



Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand). This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.
<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library [Thesis Consent Form](#)

**EFFICIENT AND “FAIR” PRICING UNDER
NEW ZEALAND’S POWER DISTRIBUTION SECTOR REFORMS:
A MODEL OF INTERTEMPORAL CROSS SUBSIDIES AND ECONOMIC DEPRECIATION**

CALUM IAN MAXWELL GUNN

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY IN ECONOMICS,
THE UNIVERSITY OF AUCKLAND, AUGUST 2002**

ABSTRACT

The New Zealand Government's objective in reforming the electricity supply industry has been *economic efficiency*. Policy actions specific to the power distribution sector have been based on the premise that electricity distribution is a natural monopoly, and a key desired outcome has been efficient and "fair" prices—those which allow electricity distributors to make a "*fair return*" on their network investments, while ensuring that consumers face prices which are "*subsidy-free*".

This thesis poses the question: *what are the characteristics of efficient and "fair" prices for power distribution network services?* Of crucial significance to this question is the *time dimension*, since debates over pricing principles posit: *static* versus *dynamic* efficiency; *short run* versus *long run* marginal cost; *backward-looking* versus *forward-looking* costs; *historic* cost versus *replacement* cost valuation; and *back-loaded* versus *front-loaded* depreciation. To address this question, a deterministic two-good/two-period model of intertemporal subsidy-free prices and economic depreciation is presented, by extending the model of intertemporal unsustainability developed by the contestability theorists, William Baumol, John Panzar and Robert Willig.

This new model indicates that intertemporally subsidy-free prices are *forward-looking*, indexed to the hypothetical amortised *opportunity costs* incurred by a coalition of current and future consumers optimally constructing a *greenfields* network to meet their own demand. Depending on the similarities between this notional asset configuration and the incumbent distributor's actual network, such prices may or may not reflect the distributor's historic or replacement costs. Where spare capacity is optimally built today, in *anticipation* of future demand, prices should cover the opportunity cost of the total capacity required to meet current and future demand. Where capacity does not require *expansion* or *replacement* until some later date, prices should initially cover the opportunity cost of the capacity required to meet current demand alone, then rise to the cost of total capacity at such time as it would become optimal for consumers to construct greenfields capacity sufficient to meet both current and anticipated demand. These results reaffirm Marcel Boiteux's position that spare capacity has its own income, as well as Ralph Turvey's view that the expectation of lower costs in future raises today's prices, providing—in some cases—justification for accelerated depreciation. However, under New Zealand's light-handed regulatory regime, electricity distributor prices and associated depreciation schedules do not appear to have exhibited these characteristics.

PREFACE AND ACKNOWLEDGEMENTS

This thesis is submitted in the Department of Economics at the University of Auckland, and concerns efficient and “fair” pricing by electricity lines businesses (i.e., power distribution companies) in New Zealand. However, this was not the original topic of inquiry, or even the same discipline. The research presented in this dissertation has its origins back in October 1992 with a meeting between myself, Nalin Pahalawaththa of the Department of Electrical and Electronic Engineering at the University of Auckland, as well as Peter Yeung, Network Planning Manager, and Graham Slack, Head of the Energy Efficiency Centre, both of what was then the Auckland Electric Power Board (AEPB). AEPB was prepared to fund two years of PhD-level research into the New Zealand prospects for demand side management (DSM)—namely, influencing consumer electricity demand in order to defer distribution network investments—with specific case studies relating to problem areas in AEPB’s network. Thanks to the support of John Boys, Head of the Department of Electrical and Electronic Engineering, and of Richard Gibbons, AEPB’s General Manager Network, the work program was approved by AEPB’s Chief Executive, Peter Cebalo, and research commenced after my PhD enrolment over a year later in May 1994, with Nalin as my supervisor.

Research began on two fronts: data collection concerning highly loaded sections of AEPB’s power distribution network, and background research on DSM programs in other countries, particularly via changes to power tariff schedules and energy efficiency initiatives. However, by this stage, New Zealand’s power distribution sector had begun down a path of reform that is still ongoing. AEPB became Mercury Energy Ltd., a company required by law to operate as a successful business, whereas previously its objectives had been more aligned to the provision of a public service.

The fruitful relationship with Mercury lasted until February 1997, and the usual disclaimer of course applies. The views presented in this thesis are my own, and should on no account be attributed to AEPB, to Mercury Energy, or to Vector Ltd.—the current name of what previously was Mercury’s network business. Since leaving Mercury, and New Zealand, I have been kept updated on New Zealand’s power sector reforms by the tireless efforts of Peter van Duivenboden of Vector, who also provided me with much of the power distribution cost data used in Chapters III, VI and IX.

The initial research program led to collaborative efforts with Robert Tromop and Graham White of New Zealand’s Energy Efficiency and Conservation Authority. Unfortunately, this work, published as Tromop *et al.* (1996), but originally presented at the 1995 Annual Conference of the Institution of Professional Engineers New Zealand (IPENZ) held at Massey University, is relevant to but a single footnote of this dissertation. For it was already becoming clear that most DSM initiatives would be at odds with the new commercial environment in New Zealand’s electricity supply industry, and that only DSM measures directed toward making power tariffs more allocatively efficient—in

other words, “getting the price right”—would be consistent with the Government’s goal of economic efficiency in the power sector.

In parallel, the research had been attempting to grapple with the question of how an electricity lines business (ELB) could best set its line charges (i.e., network conveyance and connection prices) to meet the new requirement to operate as a successful business. Peter Yeung was instrumental in encouraging me to examine this topic, and sought my appointment to an internal review team of Mercury’s line charges over the summer of 1994/95. Susan Wells, Network Commercial Manager, and the other members of the review team—namely Simon MacKenzie, Phil Henderson and William Meek—all helped me to gain an understanding of the internal and external constraints faced by an ELB on its pricing behaviour under the prevailing commercial environment, as did Bruce Turner and Neil Williams. During this time, both Robert Burley and Mei Leong at Mercury helped me to understand the Government’s new light-handed regulatory regime and, in particular, the Optimised Deprivation Valuation (ODV) methodology, discussed in Sections 7.4, 7.5 and 9.3 of this dissertation.

In light of these developments, work on DSM case studies ceased, and a revised thesis topic and scope was presented to a PhD panel comprising Nalin Pahalawaththa and Peter Yeung, as well as Des Tedford from the Department of Mechanical Engineering, and Andy Philpott from the Department of Engineering Science. The revised scope focused on pricing and investment strategies for a typical New Zealand power distribution company, both on the supply and the demand sides. Andy suggested that I model the problem using mathematical programming techniques, and this work was summarised in Gunn (1996). The qualitative analysis of regulatory and competitive constraints performed as part of this work formed the foundation for the discussion of distribution network competition in Chapter III. (This work, and parts of Chapter III, also benefited from personal experience gained in power distribution optimisation and planning techniques while employed by Worley Consultants Ltd.—now Meritec Ltd.—and Power Technologies Inc., during the period 1986 to 1990; and thanks are due to my mentors and colleagues during that period: Chi-Nai Chong, Norm Castle, Ian Grant, Kou Chang, Rob Blackburn, Miles Wyatt, Mike Breckon and Garth Harris. However, in submitting the research on the mathematical programming model for publication in the engineering literature, it was rejected on the grounds that it was too economic in content and should be submitted to a different forum.

This advice by peer reviewers, and the burgeoning recognition of the crucial importance of economic theory to the power distribution sector reforms, led to Nalin co-opting Basil Sharp of the Department of Economics as a co-supervisor in the second half of 1995. (This was good timing, as I was at this time completing my Diploma of Commerce majoring in economics, which—at least partly—prepared me for some of the challenges ahead). Basil encouraged me to present my work on the tension between energy efficiency and economic efficiency objectives to an audience of

economists, and this resulted in a paper presented at the 1995 Joint Conference of the New Zealand Association of Economists and the Law and Economics Association of New Zealand, held at Lincoln University. Basil and his colleague in Economics, Tony Endres, provided editorial input into this paper, much of which has found its way into Chapter II. This paper was eventually revised and published in *Energy Policy* as Gunn (1997). Reading this publication spurred Gordon Edge, editor of *Power Economics*, into inviting me to write an article overviewing the first decade of New Zealand's power sector reforms (i.e., Gunn, 1998), and this has also formed the basis for some of the sub-sections in Chapter II.

During this period, discussions with my co-supervisors—as well as with Tim Hazledine, Department of Economics at the University of Auckland; David Nutt at the Auckland University of Technology; Grant Read and Pat Bodger, respectively of the Departments of Management Science and of Economics at the University of Canterbury; and David Currie, Head of Network Economics at London Business School—all helped me to hone in on what should, and what should not, be the final topic and methodological approach. The first paper resulting from the new approach—published in *Energy Economics* (i.e., Gunn and Sharp, 1999)—was grounded in contestability theory, and charted the course for the eventual development of the model of intertemporal subsidy-free prices and economic depreciation presented in Chapters VIII and IX—currently under peer review in the *Journal of Regulatory Economics* (i.e., Gunn, 2002). In early 2002, Basil was instrumental in overseeing my change in registration from Engineering to Economics, since minimal engineering content remained in the research program. Furthermore, because Nalin had by this time left the University to take up a position with Transpower New Zealand Ltd., Basil assumed all supervisory responsibilities.

Thanks are due to all those mentioned above, particularly to my supervisors, Basil and Nalin, to Peter Yeung, without whom the PhD would never have got off the ground, and to Mercury Energy Ltd. which provided the funding for the initial (although mostly discarded) two years of research. At the other end of this long journey, thanks are due to Mohammad Farhandi, my friend and mentor at the World Bank in Washington DC since mid-1997, who allowed me the highly flexible work schedule which allowed me to complete this dissertation. Finally, Alicia Scott, my wife since New Year's Eve 2001, ensured that I had all the other support essential for me to ultimately bring this work to a close.

*Nel mezzo del cammin di nostra vita
mi ritrovai per una selva oscura
ché la diritta via era smarrita.*

Thanks Ali, for bringing me out of the wood.

Calum Gunn, Mt. Pleasant, Washington DC, August 2002

**EFFICIENT AND “FAIR” PRICING UNDER
NEW ZEALAND’S POWER DISTRIBUTION SECTOR REFORMS:
A MODEL OF INTERTEMPORAL CROSS SUBSIDIES AND ECONOMIC DEPRECIATION**

TABLE OF CONTENTS

ABSTRACT	iii
PREFACE AND ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS, AND LIST OF TABLES AND FIGURES	ix
CHAPTER I – INTRODUCTION: THESIS INQUIRY AND SCOPE	1
1.1 Research question, and analytical framework	1
1.1.1 <i>Efficient and “fair” prices for power distribution network services</i>	<i>1</i>
1.1.2 <i>The sectoral focus on electricity distribution</i>	<i>2</i>
1.1.3 <i>The analytical framework of contestability theory</i>	<i>3</i>
1.2 Other key sources, and the historical and normative context	4
1.2.1 <i>Historical theoretical context</i>	<i>4</i>
1.2.2 <i>Ideological and normative context</i>	<i>5</i>
1.3 Thesis outline: New Zealand’s power sector reforms, and key cost concepts	7
1.3.1 <i>New Zealand’s power sector reforms: toward allocative or dynamic efficiency?</i>	<i>7</i>
1.3.2 <i>Electricity distributors in New Zealand: natural monopolies or natural competitors?</i>	<i>7</i>
1.4 Thesis outline: key pricing and depreciation principles	8
1.4.1 <i>Efficient pricing of electricity distribution: short run or long run concept?</i>	<i>8</i>
1.4.2 <i>Subsidy-free prices in distribution networks: backward-looking or forward-looking?</i>	<i>9</i>
1.4.3 <i>Subsidy-free prices and optimal investment: toward static or intertemporal equity?</i>	<i>10</i>
1.4.4 <i>Depreciation and valuation of network assets: historic cost or replacement cost based?</i>	<i>11</i>
1.5 Thesis outline: pricing and depreciation model, and conclusions	12
1.5.1 <i>A Two-Good/Two-Period model of intertemporal subsidy-free and sustainable prices</i>	<i>12</i>
1.5.2 <i>Key conclusions</i>	<i>13</i>
CHAPTER II – NEW ZEALAND’S POWER SECTOR REFORMS: TOWARD ALLOCATIVE OR DYNAMIC EFFICIENCY?	15
2.1 Economic rationale for New Zealand’s power sector reforms	16
2.1.1 <i>Economic liberalisation and ‘light-handed’ regulation</i>	<i>16</i>
2.1.2 <i>Economic efficiency, perfect competition and marginal cost pricing</i>	<i>18</i>
2.1.3 <i>Efficiency versus competition</i>	<i>19</i>
2.1.4 <i>The long run, the short run, and dynamic efficiency</i>	<i>20</i>
2.1.5 <i>Monopoly pricing and natural monopoly</i>	<i>21</i>
2.1.6 <i>Transaction cost theory and contestability theory</i>	<i>24</i>
2.1.7 <i>Perfectly contestable markets and subsidy-free prices</i>	<i>25</i>
2.1.8 <i>Sustainable prices and intertemporal unsustainability</i>	<i>28</i>
2.2 Sectoral rationale for New Zealand’s power sector reforms	31
2.2.1 <i>Electricity supply and electricity distribution in New Zealand</i>	<i>31</i>
2.2.2 <i>Transformation of electricity supply industries world-wide</i>	<i>32</i>
2.2.3 <i>Efficiency and fairness issues with New Zealand’s pre-reform electricity supply industry</i>	<i>33</i>

2.3	Policy framework, power sector reform objectives and desired outcomes	36
2.3.1	<i>Energy policy framework, power sector policy objectives and desired outcomes (1992-1998)</i>	36
2.3.2	<i>Policy Statement on Market Power in the Electricity Sector (1998)</i>	39
2.3.3	<i>Policy Statement on the Further Development of the Electricity Industry (2000)</i>	40
2.4	Power sector reform actions (1987-2001)	41
2.4.1	<i>Initial reforms under the 1984-1990 Labour Governments</i>	41
2.4.2	<i>Introduction of contestability in distribution through light-handed regulation</i>	42
2.4.3	<i>Role of valuation benchmarks in the light-handed regulatory regime</i>	44
2.4.4	<i>Establishment of a competitive wholesale electricity market</i>	46
2.4.5	<i>Separation of electricity distribution from retailing</i>	48
2.4.6	<i>Proposals for stronger regulation of distribution natural monopolies</i>	49
2.4.7	<i>Reforms Post-1999</i>	51
2.5	Broad critiques of the power sector reforms	52
2.5.1	<i>The narrow focus critique</i>	52
2.5.2	<i>The static versus dynamic efficiency critique</i>	54
CHAPTER III – ELECTRICITY DISTRIBUTORS IN NEW ZEALAND: NATURAL MONOPOLIES OR NATURAL COMPETITORS?		57
3.1	Distribution network services: characteristics and demand	58
3.1.1	<i>Characteristics of an electricity distributor’s product: connection capacity</i>	58
3.1.2	<i>Consumer demand for network connection capacity</i>	61
3.1.3	<i>Efficiency implications of distribution network planning and design</i>	63
3.2	Perspectives of the nature of electricity distribution	66
3.2.1	<i>Electricity distributors: natural monopolies, potential competitors, or essential facilities?</i>	66
3.2.2	<i>New Zealand Government perspectives of electricity distribution</i>	70
3.2.3	<i>Contestable functions of electricity line businesses (ELBs)</i>	71
3.3	Distribution competition in New Zealand	73
3.3.1	<i>Possible modes of distribution network competition identified in New Zealand</i>	73
3.3.2	<i>Evidence of distribution bypass and subnetwork competition</i>	75
3.3.3	<i>“Unnatural” competitors: the case for unsustainability and greater regulation</i>	78
3.3.4	<i>“Unnatural” monopolies: the case for less regulation</i>	82
3.4	Natural and regulated monopoly characteristics of distribution networks	82
3.4.1	<i>Number of distribution firms</i>	82
3.4.2	<i>Distribution network economies and the significance of common or joint costs</i>	84
3.4.3	<i>Studies of electricity distribution costs in New Zealand</i>	89
3.4.4	<i>Appropriate tests of natural monopoly under contestability theory</i>	93
3.4.5	<i>Barriers to entry for distribution networks in New Zealand</i>	95
3.5	Sunk costs, asset specificity and fungibility in electricity distribution networks	97
3.5.1	<i>Barriers to entry from sunk costs, asset specificity and imperfect fungibility</i>	97
3.5.2	<i>Sunk costs and barriers to entry in contestability theory</i>	100
3.5.3	<i>Sunk costs as sources of intertemporal unsustainability</i>	102
3.5.4	<i>The spectrum of sunk cost definitions from historical to irrecoverable</i>	103
3.5.5	<i>The fungibility of distribution network assets</i>	106
3.6	Electricity distribution costs in New Zealand	107
3.6.1	<i>Distribution Network Asset Costs</i>	107
3.6.2	<i>Zone Substation Design and Costs</i>	108

CHAPTER IV – EFFICIENT PRICING OF ELECTRICITY DISTRIBUTION:	113
SHORT RUN OR LONG RUN CONCEPT?	
4.1 Efficient pricing of electricity supply	114
4.1.1 <i>Marginal cost pricing of electricity</i>	114
4.1.2 <i>Peak load pricing, time-of-use pricing and spot pricing</i>	116
4.1.3 <i>The short run versus long run marginal cost debate: the case for LRM</i>	119
4.1.4 <i>Critiques of LRM pricing and advocacy of SRM pricing</i>	121
4.1.5 <i>Long run incremental cost pricing and opportunity costs</i>	125
4.2 Revenue reconciliation	126
4.2.1 <i>The revenue reconciliation problem</i>	126
4.2.2 <i>“Second best” Ramsey pricing</i>	127
4.2.3 <i>Multiple component tariffs and cost allocation methods</i>	129
4.2.4 <i>Optimal two-part pricing</i>	130
4.2.5 <i>A partial resolution of the revenue reconciliation problem and the SRM vs. LRM debate</i>	131
4.2.6 <i>Issues of electricity distribution pricing</i>	132
4.3 “Third best” subsidy-free pricing	135
4.3.1 <i>Equity and “fairness” in pricing: price discrimination and the removal of cross subsidies</i>	135
4.3.2 <i>Subsidy-free prices: the stand alone cost (SAC) and incremental cost (IC) tests</i>	139
4.3.3 <i>Anonymously equitable and “fair” prices</i>	140
4.3.4 <i>Anonymously equitable prices in the peak load pricing problem</i>	142
4.3.5 <i>Subsidy-free prices: the consumer perspective and game-theoretic approach</i>	144
4.3.6 <i>Subsidy-free and sustainable prices in contestability theory</i>	146
4.4 Distribution line charges in New Zealand	151
4.4.1 <i>Government perspectives regarding natural monopoly pricing</i>	151
4.4.2 <i>Ministry of Commerce recommended distribution line charge methodology</i>	152
4.4.3 <i>Approaches to line charge setting by New Zealand’s ELBs</i>	153
4.4.4 <i>Cross subsidies in New Zealand’s electricity prices</i>	154
CHAPTER V – SUBSIDY-FREE PRICES IN POWER DISTRIBUTION NETWORKS:	157
BACKWARD LOOKING OR FORWARD LOOKING?	
5.1 Constrained market pricing in practice	158
5.1.1 <i>Subsidy-free pricing as a regulatory tool: willingness-to-pay and demand side issues</i>	158
5.1.2 <i>Regulation of absolute, relative and access prices for New Zealand’s electricity distributors</i>	160
5.1.3 <i>Applicability of SAC and IC bounds</i>	162
5.2 Incremental and stand alone costs in a static world	165
5.2.1 <i>Common costs and incremental costs</i>	165
5.2.2 <i>Bounds on subsidy-free revenues in a power distribution network</i>	166
5.2.3 <i>Average incremental costs in constrained market pricing</i>	168
5.2.4 <i>Fixed costs in constrained market pricing</i>	170
5.3 The time dimension in constrained market pricing	173
5.3.1 <i>Sequencing of demands in constrained market pricing</i>	173
5.3.2 <i>Fallacies of forward-looking and backward-looking costs</i>	174
5.3.3 <i>The “incremental cost fallacy” in constrained market pricing</i>	179
5.3.4 <i>Toward intertemporal subsidy-free prices: option value and opportunity cost in investment</i>	186

CHAPTER VI – SUBSIDY-FREE PRICES AND OPTIMAL INVESTMENT:	189
TOWARD STATIC OR INTERTEMPORAL EQUITY?	
6.1 Dynamic efficiency: optimal investment and intertemporal subsidy-free prices	189
6.1.1 <i>Dynamically efficient spare capacity and asset duplication in distribution networks</i>	189
6.1.2 <i>Constrained market pricing, contestability theory and dynamic efficiency</i>	192
6.1.3 <i>The zero profit constraint over the long run</i>	195
6.1.4 <i>The optimal investment rule versus “greenfields” optimality</i>	197
6.2 Optimal construction configurations in a power distribution network	199
6.2.1 <i>Optimal construction configurations: the sequencing of demands example revisited</i>	199
6.2.2 <i>Optimal construction configurations using New Zealand distribution cost data</i>	202
6.2.3 <i>Total cost equations for a single zone substation</i>	204
6.2.4 <i>Optimal construction configurations for a single zone substation</i>	206
6.2.5 <i>Optimal construction configurations for one or two zone substations</i>	209
6.3 The appropriate costs to include in stand alone costs and incremental costs	210
6.3.1 <i>Opportunity costs and replacement costs in subsidy-free bounds</i>	210
6.3.2 <i>Cost changes due to changes in technology and demand patterns</i>	214
6.3.3 <i>The case for the inclusion of historic costs</i>	215
6.3.4 <i>The case for compensation due to breaches of the regulatory contract</i>	218
6.3.5 <i>A return to Boiteux’s pricing prescription</i>	219
6.4 Regulating for intertemporal efficiency and fairness	221
6.4.1 <i>Incentive regulation versus rate of return regulation</i>	221
6.4.2 <i>Intertemporal anonymous equity: protecting consumer interests over the long term</i>	225
6.4.3 <i>Intertemporal cross subsidies, amortisation and depreciation</i>	229
CHAPTER VII – DEPRECIATION AND VALUATION OF DISTRIBUTION NETWORK ASSETS:	231
HISTORIC COST OR REPLACEMENT COST BASED?	
7.1 Capital cost recovery and economic asset valuation	232
7.1.1 <i>Capital cost recovery and depreciation</i>	232
7.1.2 <i>Valuation under regulation: the circularity problem</i>	233
7.1.3 <i>Economic asset valuation in contestability theory</i>	234
7.2 Economic depreciation	236
7.2.1 <i>Economic depreciation, marginal costs and spare capacity</i>	236
7.2.2 <i>Economic depreciation in contestability theory</i>	237
7.2.3 <i>Economic depreciation under constant payments to capital</i>	240
7.3 Asset value and intertemporal subsidy-free prices in theory and practice	242
7.3.1 <i>Deriving economic asset value from intertemporal anonymously equitable prices</i>	242
7.3.2 <i>Firm value versus asset value: the role of accumulated depreciation</i>	246
7.3.3 <i>Optimal real world depreciation: front-loaded or back-loaded?</i>	248
7.3.4 <i>Opportunity cost-pricing of specialised assets: alternative uses versus alternative users</i>	251
7.4 Optimised deprival valuation (ODV) methodology in New Zealand	253
7.4.1 <i>The roots of New Zealand’s deprival valuation methodology</i>	253
7.4.2 <i>The optimised deprival valuation (ODV) methodology</i>	255
7.4.3 <i>Economic value (EV) of network segments</i>	257
7.4.4 <i>The evolution of the ODV methodology</i>	258
7.5 Criticisms of the ODV methodology	259
7.5.1 <i>The windfall critique</i>	259
7.5.2 <i>The circularity critique</i>	261
7.5.3 <i>The optimal spare capacity critique</i>	265
7.5.4 <i>The tomorrow’s costs critique</i>	267

CHAPTER VIII – A TWO-GOOD/TWO-PERIOD MODEL OF	271
INTERTEMPORAL SUBSIDY-FREE AND SUSTAINABLE REVENUES	
8.1 A Single-Good/Two-Period model of contestable natural monopoly	271
8.1.1 <i>BPW’s model of intertemporal unsustainability under capacity expansion</i>	271
8.1.2 <i>Sustainability of revenues in the BPW model</i>	273
8.1.3 <i>Subsidy-free revenues derived for the BPW model under capacity expansion</i>	275
8.1.4 <i>Intertemporal subsidy-free revenues and the sequencing of demands example</i>	276
8.1.5 <i>Sustainable and subsidy-free revenues under anticipatory construction optimality</i>	277
8.1.6 <i>Anticipatory construction with fungible assets</i>	279
8.2 An extended Two-Good/Two-Period (TGTP) model of natural monopoly	281
8.2.1 <i>Extending BPW’s model to two goods</i>	281
8.2.2 <i>Sustainability and resale feasibility in the two-good model</i>	282
8.2.3 <i>Subsidy-free revenues</i>	283
8.2.4 <i>The difference between sustainable and subsidy-free revenues</i>	285
8.3 Relaxing the finite demand assumption: “perpetual demand”	285
8.3.1 <i>Total costs and construction optimality under “perpetual demand”</i>	285
8.3.2 <i>Impacts on stand-alone costs, incremental costs and resale feasibility</i>	287
8.3.3 <i>Sustainable and subsidy-free revenues under perpetual demand</i>	289
8.4 Intertemporal unsustainability revisited	291
8.4.1 <i>Intertemporal unsustainability and the distinguishability of goods and consumers</i>	291
8.4.2 <i>Unsustainability under different construction configurations</i>	292
8.4.3 <i>The significance of sunk costs and fixed costs to unsustainability</i>	293
8.4.4 <i>Unsustainability due to disequilibrium and symmetry</i>	294
8.4.5 <i>Relaxing the Bertrand-Nash assumption: contestability theory and strategic behaviour</i>	296
8.5 Toward intertemporal subsidy-free prices in the TGTP model	300
8.5.1 <i>Game theoretic determination of intertemporal sustainable and subsidy-free prices</i>	300
8.5.2 <i>Allocation of non-fungible fixed costs</i>	302
CHAPTER IX – INTERTEMPORAL SUBSIDY-FREE PRICES AND ECONOMIC DEPRECIATION	305
IN THE TWO-GOOD/TWO-PERIOD MODEL	
9.1 Intertemporal subsidy-free prices in the TGTP model	305
9.1.1 <i>Intertemporal subsidy-free prices and anonymous equity</i>	305
9.1.2 <i>Assumptions for deriving subsidy-free prices rather than revenues</i>	307
9.1.3 <i>Subsidy-free prices under anticipatory construction</i>	307
9.1.4 <i>Subsidy-free prices under capacity expansion: the influence of optimal investment</i>	311
9.1.5 <i>Subsidy-free prices under capacity replacement</i>	313
9.1.6 <i>The transition point for increases in the subsidy-free price</i>	316
9.2 Economic asset valuation and depreciation in the TGTP model	317
9.2.1 <i>Asset valuation and depreciation in the model under anticipatory construction</i>	317
9.2.2 <i>Asset valuation and depreciation in the model under capacity expansion</i>	320
9.2.3 <i>Asset valuation and depreciation in the model under capacity replacement</i>	322
9.2.4 <i>Justification for subsidy-free prices in the context of perfect contestability</i>	323
9.2.5 <i>A simple example of subsidy-free price paths for a zone substation</i>	325
9.3 The efficiency and “fairness” of line charges in New Zealand	330
9.3.1 <i>Price paths for subsidy-free, LRIC, and ODV pricing under anticipatory construction</i>	330
9.3.2 <i>Price paths under capacity expansion and capacity replacement</i>	333
9.3.3 <i>Evidence of monopoly rents</i>	336
9.4 Limitations of the TGTP model	339
9.4.1 <i>Implications for regulatory policy</i>	339
9.4.2 <i>Negative economic asset values</i>	341
9.4.3 <i>The implicit obligation to supply</i>	342
9.4.4 <i>Modelling only two service</i>	345
9.4.5 <i>Uncertainty and risk</i>	346

CHAPTER X – CONCLUSIONS: TOWARD EFFICIENT AND “FAIR” LINE CHARGES	349
10.1 The question of natural monopoly	349
10.2 Efficient pricing and the time dimension	352
10.3 Intertemporally “fair” pricing and economic depreciation	355
10.4 The characteristics of efficient, “fair” (and sustainable) line charges	358
10.5 The efficiency and “fairness” of line charges in New Zealand	361
 BIBLIOGRAPHY AND REFERENCES	 363
New Zealand Electricity Line Business Data and Associated Unpublished Reports	363
New Zealand Acts of Parliament, Bills and Regulations	365
New Zealand Government Publications, Reports, Media Releases, Letters and Speeches	366
Non-Government Reports, Newspaper Articles, Unpublished Papers, Monographs, Dissertations, Letters and Submissions	369
Published Works	374

LIST OF TABLES

<i>Table 3.1</i>	<i>Indicative Zone Substation Firm Capacities (MVA)</i>	<i>111</i>
<i>Table 6.1</i>	<i>Optimal Single Zone Substation Construction Configurations</i>	<i>207</i>
<i>Table 6.2</i>	<i>Optimal Construction Configurations where Initial Substation Cannot be Modified</i>	<i>210</i>

LIST OF FIGURES

<i>Figure 3.1a</i>	<i>No Competition</i>	<i>76</i>
<i>Figure 3.1b</i>	<i>Retailer Interposed Agreement</i>	<i>76</i>
<i>Figure 3.1c</i>	<i>Conveyance Agreement</i>	<i>76</i>
<i>Figure 3.1d</i>	<i>Bypass Scenario</i>	<i>76</i>
<i>Figure 3.2</i>	<i>11kV “Average Incremental Cost” Curve</i>	<i>92</i>
<i>Figure 3.3</i>	<i>Zone Substation Total Costs (by Contingency Criteria and Transfer Capacity)</i>	<i>112</i>
<i>Figure 3.4</i>	<i>Zone Substation Average Costs (by Contingency Criteria and Transfer Capacity)</i>	<i>112</i>
<i>Figure 6.1</i>	<i>Single Zone Substation Average Costs under Different Construction Configurations</i>	<i>208</i>
<i>Figure 6.2</i>	<i>Single Zone Substation Optimal Construction Configurations</i>	<i>209</i>
<i>Figure 9.1</i>	<i>First and Second Period Subsidy-Free Prices for Transformer Capacity as a Function of the Year of Demand Growth (T)</i>	<i>326</i>
<i>Figure 9.2a</i>	<i>Annual Subsidy-Free Prices under Anticipatory Construction (T = 10 years)</i>	<i>327</i>
<i>Figure 9.2b</i>	<i>Economic Asset Value under Anticipatory Construction (T = 10 years)</i>	<i>328</i>
<i>Figure 9.3a</i>	<i>Annual Subsidy-Free Prices under Capacity Expansion (T = 30 years)</i>	<i>328</i>
<i>Figure 9.3b</i>	<i>Economic Asset Value under Capacity Expansion (T = 30 years)</i>	<i>329</i>
<i>Figure 9.4a</i>	<i>Annual Subsidy-Free Prices under Capacity Replacement (T = 45 years)</i>	<i>329</i>
<i>Figure 9.4b</i>	<i>Economic Asset Value under Capacity Replacement (T = 45 years)</i>	<i>330</i>
<i>Figure 9.5</i>	<i>Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing (Anticipatory Construction Optimality)</i>	<i>333</i>
<i>Figure 9.6</i>	<i>Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing (Capacity Expansion Optimality)</i>	<i>335</i>
<i>Figure 9.7</i>	<i>Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing (Capacity Replacement Optimality)</i>	<i>336</i>
<i>Figure 9.8</i>	<i>Ratio of Actual Payments to Capital to Minimum Economic Payments to Capital for New Zealand ELBs (under Different WACCs)</i>	<i>338</i>

CHAPTER I

INTRODUCTION: THESIS INQUIRY AND SCOPE

On May 25 this year I announced that the Government would move to corporatise the electricity distribution industry. Our motivation for this is to keep prices as low as possible: New Zealand Minister of Energy, David Butcher (1990b)

So began the long-expected reform of New Zealand's power distribution sector, a process that had already got underway some years before in the power generation and transmission sectors. As the quote above implies, from the outset a key focus of this reform program was *prices*. Later however, the Government made it clear that the overall objective in reforming the electricity supply industry was to be much broader; the attainment of *economic efficiency* in all its forms—productive, allocative and dynamic. Lower prices would simply be a serendipitous by-product of improved efficiencies. But apart from the axiomatic concepts of economic efficiency encapsulated in this overarching goal, many other economic concepts have underpinned the power sector reforms, including those relating to industry structure (for instance, '*natural monopoly*'); market behaviour ('*contestability*' and '*sunk costs*'); the valuation of firms ('*opportunity costs*' and '*depreciation*'); and the costs of supply ('*marginal cost*', '*incremental cost*' and '*stand alone cost*'). Moreover, underlying all the policy actions has been the notion that the prices of all services provided throughout the power delivery system should give both producers and consumers that idiosyncratically-Kiwi sentiment of a "*fair go*". On the one hand, prices should allow firms to make a "*fair return*" on their investment, while on the other, consumers should be free from monopolistic exploitation on the part of their suppliers and should face prices which are "*subsidy-free*".

1.1 Research Question, and Analytical Framework

1.1.1 Efficient and "Fair" Prices for Power Distribution Network Services

This thesis poses the question: *what are the characteristics of efficient and "fair" prices for power distribution network services?* (Such prices relating to electricity distribution are also termed '*line charges*' by New Zealand policymakers and those in the industry). The discussion throughout this thesis highlights that assessments of efficient and fair line charges crucially depend on how the *time dimension* is treated in the concepts germane to the Government's reform program, especially in definitions which relate to the *costs* of providing distribution services. Clearly, the manner in which economic concepts are defined can have a marked influence on the implementation of sector policy.

The significance of "time" is evidenced in the polarity of the wide-ranging debates over the appropriate pricing of electricity supply, both in New Zealand and internationally. The surrounding arguments endorsing various pricing principles posit: *static* efficiency versus *dynamic* efficiency; *short run* marginal cost versus *long run* marginal cost; *backward-looking* costs versus *forward-looking* costs; *historic* cost valuation versus *replacement* cost valuation; and *back-loaded* depreciation versus

front-loaded depreciation. The analysis presented in this thesis—which is based on a very simple two-good/two-period model of intertemporal subsidy-free prices—by no means fully answers the question presented above. However, in discussing the tensions inherent in these opposing concepts—within the context of electricity distribution—the dissertation concludes that the appropriate pricing rule is not new, but bears significant similarities to the pricing prescriptions of Marcel Boiteux, a French engineer and economist. Boiteux (1956) suggested that the appropriate electricity pricing rule is to base prices on what *would* be the optimum network configuration, were capacity perfectly adjustable. Moreover, both Boiteux, and another pioneer of marginal cost pricing principles—the UK’s Ralph Turvey—implied how depreciation schedules can ensure that both firms and consumers can get their “fair go”.

1.1.2 The Sectoral Focus on Electricity Distribution

Although there are efficiency and fairness considerations for the pricing of the services provided by every sector which comprises the electricity supply industry—namely, power generation, wholesaling, transmission, distribution and retailing—this thesis focuses firmly on electricity distribution. This is primarily because of this sector’s relative neglect in the academic literature. With respect to power sector reform in general, Berger and Spiller (1996, p. 2) point out that: “much of the academic discussion on the move towards a competitive electricity sector has focused on the organization of the wholesale market and on transmission pricing”. And although much has been written about the reforms of the power distribution sector in New Zealand, such references generally take the form of commentaries or criticisms, phrased as opinions, rather than that of peer-reviewed analytical work. Not much has changed in the almost twenty years since Joskow and Schmalensee (1983, p. 59) wrote that: “Little theoretical or empirical work focuses explicitly on the economic characteristics of electric power distribution systems”. For instance, in 1995, Claggett *et al.* (1995) could still state with reasonable surety that “relatively few academic researchers have examined electrical distribution networks in isolation, although fully integrated utilities have been the subject of numerous studies that examined various aspects of economic efficiency”.

By contrast, the economic and engineering literature over the past decade or so, both internationally as well as from New Zealand, is overflowing with academic papers relating to wholesale power pools and other wholesale market mechanisms (e.g., David and Li, 1993a; Ring and Read, 1995), to forward and hedge contracts for electricity supply (e.g., Kaye *et al.*, 1990), to optimal transmission pricing (e.g., Tabors, 1994; Read, 1997), as well as to other highly technical aspects of electricity supply in a deregulated power sector. To some extent this could be because the technology involved in distributing electricity is much simpler than that involved in its generation or transmission, and the nature of the business itself, including the possible nature of financial and commercial relationships, is also less

complicated.¹ Perhaps this complexity makes these elements of power supply more attractive to researchers. On the other hand, as this thesis suggests (Chapter III), the prevailing “common sense” view that electricity distribution is a natural monopoly may be partly responsible for the lack of attention.

A similar situation exists in relation to the more general literature relating to the economics of networks, which encompasses sectors other than power, such as natural gas, water and, most significantly, telecommunications. As Economides (1996) laments, “the literature on networks is so extensive, that it is futile to attempt to cover it”. Research into telecommunications in particular has benefited from the publicity surrounding such highly visible disputes as the Telecom versus Clear case in New Zealand (§5.1.2). Yet within this vast literature on networks, very few authors have considered issues specific to electricity distribution. Nevertheless, the business of electricity distribution is enormous worldwide. In the US (for 1996), it has been estimated that the value of distribution assets for investor-owned utilities alone was US\$100 billion, and the associated revenues these companies received was US\$40 billion (Lowry and Kaufmann, 1998, p. 3). And in New Zealand for the 1999/2000 tax year, electricity distributors were valued at a total of NZ\$4.23 billion (Energy Markets Regulation Unit, 2000d, p. 9). While there is some controversy over the method used to value these assets (§7.5), it is clear that the sector is significant, and relatively small gains in efficiency could translate into substantial absolute benefits for the economy.

1.1.3 The Analytical Framework of Contestability Theory

Aside from Marcel Boiteux, if but a single figure is selected to tower over the discussion and analytical work presented in this thesis, then this is William Baumol, whose work on contestability theory clearly had a substantial influence on the original architects of New Zealand’s power sector reforms (Chapter II). A key reference work cited throughout this thesis is the second edition of William Baumol, John Panzar and Robert Willig’s (1988) consummate text: *Contestable Markets and the Theory of Industry Structure* (cited as ‘BPW’). In particular, their model of intertemporal unsustainability in a multiproduct natural monopoly—which is presented in Chapters 13 and 14 of that text—is used as the basis for the model of intertemporal subsidy-free prices and economic depreciation developed in

¹ The costs of power generation change from second to second, whereas the costs involved in network connection are relatively fixed. And although electricity transmission and distribution are both network industries, transmission networks are operated as a mesh, whereas distribution networks are generally operated (if not constructed) in a radial fashion (§3.1.2). Therefore, the analysis of optimal pricing, operation and investment for generation and for transmission networks, lend themselves to complex mathematical programming techniques (e.g., Drayton and Read, 1996; Hogan *et al.*, 1996), although these can be also—but infrequently are, at least in New Zealand—applied to problems of distribution planning (§3.1.3). In a few cases, generation spot pricing and transmission nodal pricing techniques have, at least in the theoretical literature, been extended down to distribution level (e.g., Murphy *et al.*, 1994; Farmer *et al.*, 1995; §4.2.6). However, generally this has been with the objective of more closely exposing consumers to the time-varying nature of *energy* prices at various “nodes” in the network, rather than to investigate the costs and prices specific to investment in the distribution *network* itself.

Chapters VIII and IX of this thesis. A key supplementary text written within the contestable markets paradigm is Baumol's collaborative tract, written with Gregory Sidak, on competition in local telephone markets, which—among other issues—considers the derivation of bounds on subsidy-free prices for a regulated monopoly (i.e., Baumol and Sidak, 1994a).

But apart from being the catalyst for much of the work on contestable markets, and the lead author of its definitive text, Baumol made relevant contributions to many of the economic concepts discussed throughout this thesis, including: natural monopoly theory (e.g., Baumol, 1977; §3.4.4); the role of fixed and sunk costs (Baumol and Willig, 1981; §3.5.2); stranded costs (Baumol and Sidak, 1995b; §3.5.1); marginal cost pricing and Ramsey pricing (Baumol and Bradford, 1970; §4.2.2); optimal depreciation policies (Baumol, 1971; §7.1-§7.2); cost allocation (Baumol *et al.*, 1987; §4.2.3); cross subsidies and “fairness” in pricing (Baumol, 1986; §4.3.1); constrained market pricing (Baumol, 1979; §5.1.2); price sustainability (Baumol *et al.*, 1977; §4.3.6); network access pricing (Baumol and Sidak, 1994b; §5.1.2); rate regulation (Baumol, 1968; §6.4.1); incentive regulation (Baumol and Sidak, 1994a; §6.4.1); as well as more philosophical work on the contributions of economic theory to policy (Faulhaber and Baumol, 1988; §4.3.6). Baumol's collaborator in this latter work—Gerald Faulhaber—is also a key figure, since his work on cross subsidisation in public enterprises (i.e., Faulhaber, 1975; and Faulhaber and Levinson, 1977; §4.3.2-§4.3.4) formed the basis for much of Baumol's work on the topic of subsidy-free and fair pricing.

1.2 Other Key Sources, and the Historical and Normative Context

1.2.1 Historical Theoretical Context

The fact that the key conclusion of this thesis exhibits strong parallels with work performed half a century ago is significant. Bronfenbrenner (1971), for one, surmises that economics is “parasitic”, feeding off its past, and that economic concepts are rarely “relegated to the antiquarian's dustbin”. And Faulhaber and Baumol (1988) themselves discuss such parasitism in regard to some of the key theories discussed in this dissertation. They also muse whether the discipline of economics acts simply to describe and predict market behaviour, or whether economic research itself influences market behaviour. Although the neoclassical model of the economy assumes rational optimising behaviour on the part of economic agents, Faulhaber and Baumol observe that various elements of economic theory have in fact helped those agents to *improve* their competitive behaviour. They trace the cross-fertilisation of theory and market behaviour in some of the key elements of economic theory discussed in this dissertation, namely: marginal analysis (§4.1.1); peak load pricing (§4.1.2); Ramsey pricing (§4.2.2); and the stand alone cost (SAC) test (§4.3.2).

The historical background and other context of economic propositions is therefore important, for although many theories might appear to become superseded by events and advances in technology—such as the short run marginal cost (SRMC) versus long run marginal cost (LRMC) pricing debate with

respect to power generation (§4.2.5)—these old controversies often include issues still disputed today, but in a different setting. For instance, the SRMC versus LRMC debate still crops up in regard to issues of power distribution pricing (§4.1.1, §4.2.6 and §5.3.2). And the fact that the literature on cross subsidy initially stemmed from examples applying to the railroad industry may partly explain why many of the normative prescriptions arising from the theory appear to assume that “common” costs are also “fixed” costs (§5.2.4 and §5.3.3). Consequently, the work of such pioneers of marginal cost pricing in practice as Boiteux and Turvey are plumbed, since they dealt with many of the same problems that policymakers face—and still have not entirely resolved—to this day. This circularity is particularly evident in the insightful work of Boiteux, who was one of the few early researchers to examine the specific economic characteristics of electricity distribution in any depth (§3.1, §4.1-§4.2, §6.1, §6.3.5 and §9.2.4).

As Berg and Tschirhart (1995) observe, “there is no doubt that economists’ impact is limited when analyses ignore the historical context in which issues arise”. For example, Turvey’s work (§4.1.3) is often not cited these days because it arose in a period when changing electricity prices in real time was not technologically (or institutionally) feasible. This does not alter the fact that his exposition of theory ignored these limitations, yet he was keenly aware of the complexity of moving from theory to practice. From his underlying rich theoretical discussions of marginal cost and economic depreciation, Turvey moved to addressing how marginal cost pricing and depreciation schedules could be implemented in practice, taking pragmatic constraints of his day into consideration (§4.1-§4.2, §5.3.2, §6.1.4, §7.2-§7.3, and §9.2.4). Similarly, the much later developers of the theory of real time pricing (§4.1.2) took a very pragmatic approach to implementing their theory, cognisant of contemporary restrictions.

1.2.2 Ideological and Normative Context

In addition, economic methodologist Mark Blaug (1980) has suggested that economists need to be reminded “of the fallacy of trying to appraise particular theories without invoking the wider metaphysical framework in which they are embedded”. This is particularly true in observing the subtle changes in the normative prescriptions arising from proponents of the theory of contestable markets. Of particular interest is the work of Gregory Sidak, initially a key collaborator with Baumol on the application of contestability theory to real world problems in the telecommunications and power industries (i.e., Baumol and Sidak, 1994a-b and 1995a-b). While Baumol continued to strongly defend himself and his colleagues from criticisms that they were strict adherents to *laissez faire* policies (§2.1.7)—stating that “Contestability theory supports neither extreme interventionists nor extreme noninterventionists” (Baumol, Panzar and Willig, 1988, p.476)—by contrast Sidak, in his later collaborative work with Daniel Spulber (i.e., Sidak and Spulber, 1997), can be seen: citing the neo-

Austrian school; showing scepticism of any form of regulation; endorsing positive economic profits; and viewing perfect contestability as only one of a number of possible market equilibrium models (§5.3.2).²

This thesis does not attempt to cast judgement on any of these somewhat philosophical issues. Nor does it add to the ideologically-charged debate over the appropriate governance framework for New Zealand's power distribution sector, the appropriate level of state intervention, and the right mix of regulatory or competitive solutions. Instead, as stated above, this thesis directs its attention toward one of the key *outcomes* which would indicate the success (or otherwise) of any such regime governing electricity distribution—namely, efficient and “fair” line charges. If policymakers have no benchmark to indicate what the characteristics of such prices might be, then they have no basis for declaring that their policy goals have been achieved. This is not to suggest that the way those benchmark prices are derived can entirely be an exercise in positive economics, free from normative influences. Proponents of *laissez-faire* would likely claim that the appropriate benchmark is provided by the competitive market outcome itself, and that prices observed in the marketplace are by definition efficient and fair. Others might suggest that this is a tautological argument exhibiting a major fallacy of consequence.

But rather than be caught up in such a dispute, this thesis presents its discussion within the framework which the New Zealand Government *itself* established in engaging upon its reform program. The Government's overall objective was economic efficiency, key outcomes were to be efficient, fair and subsidy-free power prices, and policy actions were firmly grounded in the concepts of natural monopoly and contestable markets which stem from contestability theory. Hence, the discussion draws conclusions on how the Government might begin to evaluate the success of its reforms, with its own ideological outlook and analytical framework taken as given. Nevertheless, in doing so, the nature of that outlook and framework are identified, as are the key criticisms of reform, since such critiques can sharpen the understanding of both.

² Controversially, Shepherd (1995) appears to view the policy prescriptions of the contestability theorists as stemming from vested interests, rather than from some ideological standpoint. Shepherd sees the fact that Baumol, as well as his colleagues, John Panzar, Robert Willig and Elizabeth Bailey, were all at one point in time employees of, or contractors to, the Bell System as being particularly significant. Shepherd states that the Baumol group were hired at least partly to formulate ideas to resist the anti-trust policies which the Bell System (subsequently AT&T) faced during the 1970s and 1980s: “Contestability theory—particularly as it was framed by the Baumol group—tended directly to advance the interests of this company”. Shepherd does not include Gerald Faulhaber and Stephen Levinson in his list of those at times on the AT&T payroll who, although not directly contributing to the standard contestability texts, provided the theoretical underpinnings to much of the Baumol group's *policy* prescriptions (§4.3.2-§4.3.4).

1.3 Thesis Outline: New Zealand's Power Sector Reforms, and Key Cost Concepts

1.3.1 New Zealand's Power Sector Reforms: Toward Allocative or Dynamic Efficiency?

The main body of this dissertation opens in Chapter II with an outline of New Zealand's power sector reform program, in the context of both the liberalisation of other sectors of the country's economy (§2.1), and the transformation of electricity supply industries worldwide (§2.2). In particular, the theoretical justification for policy actions provided by '*contestability theory*' is examined, and the benchmark of industry conduct articulated by contestability theory is contrasted with that associated with the traditional neoclassical model of '*competition*', in which the mere presence of '*monopoly*' is deemed coincident with inefficient pricing behaviour. Contestability theory maintains that the presence of '*natural monopoly*'—which occurs when a single firm can serve the market in question at least cost—does not in itself indicate the existence of monopoly power, and the need for some form of regulatory price control.

Before turning to a brief review of the actual policy actions themselves (§2.4), the various statements issued by successive Governments regarding reform objectives and desired outcomes are presented (§2.3). (At the time of the submission of this thesis the review of the power sector's regulatory framework was still ongoing. However, in order to bring the research program to some form of closure, this dissertation's review of the Government's power sector reform program ceases with the passage of the Electricity Industry Bill on 7 August 2001). Finally, the Chapter closes with a brief summary of the key critiques of the reform program to that date (§2.5), which range from those dismissing the entire package as ideologically driven, to those in agreement with the policy framework, but not with the means to achieve or to measure the desired outcomes. Most notably, a key question is whether the Government might have placed an overemphasis on gains in the '*static*' components of efficiency—namely '*productive*' and '*allocative*' efficiencies—at the expense of '*dynamic*' efficiency.

1.3.2 Electricity Distributors in New Zealand: Natural Monopolies or Natural Competitors?

The focus narrows to electricity distribution in Chapter III, and to the question whether distributors are natural monopolies, or instead whether they might be able to engage in competitive behaviour. Firstly, the definition and characteristics of the '*product*' which distributors provide are identified, as are the key issues involved in optimal distribution network investment (§3.1). Electricity distribution is described as a multiproduct activity providing '*network connection capacity*' to consumers distinguished by *location* and *time* of demand. It is then pointed out that the Government's specific policy actions in the power distribution sector have been driven by the presupposition that the sector has naturally monopolistic characteristics, an assumption also made in much of the relevant international literature (§3.2). Some overseas authors also observe that there may be no set of prices available to an incumbent natural monopolist which are '*sustainable*'; namely, prices which prevent inefficient entry and wasteful asset duplication. The appropriate response to this problem, they suggest, might be regulatory

protection of the incumbent, whereas the contestability theorists themselves suggest that such ‘*unsustainability*’ may require regulators to allow firms greater pricing freedom (§3.3.3).

Interestingly, many local commentaries affirm that distributors are *not* natural monopolies. Instead, these claim that electricity distribution comprises some contestable functions, and some evidence exists that competition for network connection has already been occurring (§3.3.1-§3.3.2). As such, limited regulation might be warranted (§3.3.4). The Government, industry commentators and industry participants have supported their positions with observations and speculations regarding the number of firms in the sector, various network economies—such as ‘*economies of scale*’ and ‘*economies of scope*’—as well as industry cost data (§3.4.1-§3.4.3). Yet no conclusions have been drawn by applying the cost tests for natural monopoly which are actually derived from contestability theory (Baumol, Panzar and Willig, 1988; §3.4.4).

Critics of contestability theory stress the role of ‘*barriers to entry*’ in allowing firms to raise prices to inefficient and exploitative levels (§3.4.5). However, to contestability theorists, the only significant barriers to entry are ‘*sunk costs*’; a concept which is indicated as having numerous shades of meaning (§3.5.1-§3.5.2, and §3.5.4). Moreover, whereas ‘*fixed costs*’ are considered to improve the sustainability of an incumbent monopolist, sunk costs are seen as contributing to unsustainability, particularly when investment over time is taken into account (§3.5.3). This is because sunk costs are associated with assets which are ‘*non-fungible*’—in other words, they have no alternative use inside or outside the industry in question—and investment decisions pertaining to those assets thus become ‘*intertemporally interdependent*’. Subsequently, a brief assessment is performed of barriers to entry in New Zealand’s power distribution sector, and of the fungibility of distribution network assets (§3.4.5 and §3.5.5). Chapter III closes by deriving a simple cost function for the assets comprising a typical New Zealand zone substation (§3.6), since zone substations are the elements in any power distribution network which exhibit the strongest economies of scope. (This cost function is utilised in later sections of the thesis: §6.2 and §9.3).

1.4 Thesis Outline: Key Pricing and Depreciation Principles

1.4.1 Efficient Pricing of Electricity Distribution: Short Run or Long Run Concept?

A broad review of traditional electricity pricing theory within the framework of the static competitive market model is presented in Chapter IV, which encompasses ‘*marginal cost pricing*’, ‘*peak load pricing*’, ‘*spot pricing*’, and ‘*long run incremental cost*’ pricing (§4.1), as well as ‘*Ramsey pricing*’, multiple component tariffs, cost allocation methods, and optimal two-part pricing (§4.2).³ A key tension

³ In particular, the peak load pricing problem has provided a useful setting for the discussion of other aspects of the characteristics of efficient and fair prices, especially in regard to: anonymously equitable prices (§4.3.4-§4.3.5); the fallacy

underlies the presentation of these various pricing approaches: the relative significance of ‘*long run*’ and ‘*short run*’ cost factors, and whether the costs of capacity should or should not be included in prices (e.g., Boiteux, 1956; and Turvey, 1969). It is then explained that the contestability theorists neatly step aside from the SRMC versus LRMC debate by rejecting the competitive market paradigm entirely. They propose the implementation of an approach which regulators have called “*constrained market pricing*”, derived from a contestable markets framework (Baumol and Sidak, 1994a; §4.3.1).

Baumol and Sidak suggest that prices should not be set to a specific measure of cost. Rather, efficient and “fair” prices—namely those which are free from ‘*cross subsidy*’—will lie in a *range* within which firms should be free to set their tariffs, based on their observations of relative consumer demand characteristics, subject to the constraint that firms make a ‘*zero economic profit*’. This range of prices is bounded by the ‘*stand alone cost*’ (SAC) and ‘*incremental cost*’ (IC) of supplying the product or group of products in question (§4.3.2), or alternatively, the particular consumer or group of consumers (§4.3.3-§4.3.4). Prices satisfying all relevant constraints are considered to be ‘*subsidy-free*’ (Faulhaber, 1975), or from a consumer perspective, ‘*anonymously equitable*’ (Faulhaber and Levinson, 1977). In these original formulations, stand alone cost is derived by estimating the costs which a *hypothetical entrant* would incur in serving a particular subset of consumers, or the costs those consumers themselves would face should they *self-produce* the service. By contrast, the incremental cost of supplying a subset of consumers should not be derived in its own right, but evaluated by subtracting the stand alone cost of supplying the ‘*complementary*’ subset of consumers, from the zero economic profit constraint associated with serving *all* consumers. The Chapter closes by pointing out that, even though New Zealand Government agencies are well aware of all the pricing techniques discussed throughout Chapter IV, both the recommended and actual approaches for setting distribution line charges are largely based on cost allocation approaches which have no basis in either the competitive or contestable market model (§4.4).

1.4.2 Subsidy-Free Prices in Distribution Networks: Backward-Looking or Forward-Looking?

In Chapter V, some definitional problems with constrained market pricing are examined. In particular, there are questions regarding the application of the approach to issues of *absolute* price levels, *relative* price levels and *access* prices, given that the range between stand alone cost and incremental cost is often considered to be large in practice. If such is the case, then firms—even if they are monopolists—may enjoy substantial pricing freedom (§5.1). However, using a simple *static* two product model—one that ignores any influence from the passage of time—it is demonstrated that the large gap between SAC and IC is partly attributable to a fallacious approach to calculating those bounds. The fallacy—which is not made by Baumol and Sidak themselves, but by other advocates of subsidy-free pricing—is that any

that physically common costs do not contribute to incremental costs (§5.2.1); intertemporal subsidy-free prices (§5.3.4); and economic depreciation (§7.2.1).

physically ‘*common costs*’ (or ‘*joint costs*’) of capacity should not be included in the incremental cost calculation. This misconception appears to arise from an implicit assumption that common costs are equivalent to ‘*fixed costs*’, or possibly that there are constant returns to scale (§5.2).

However, introducing the *time* dimension into the model—by *sequencing* the onset of demand for the two products (§5.3.1)—illustrates a whole range of time-related “fallacies of *forward-looking costs*” and of “*backward-looking*” costs relevant to the derivation of subsidy-free bounds on prices (§5.3.2). Such fallacies—which are usually cited as being “unfair” from a regulated firm’s perspective—include: “ignoring *investment-backed expectations*”, by considering that, once capital expenditures have been “sunk”, they do not need to be recovered in prices; as well as failing to incorporate into prices the full ‘*opportunity costs*’ of all the resources utilised in an investment (Schramm, 1991; and Sidak and Spulber, 1997).

Baumol and Sidak appear to hold the view that the subsidy-free bounds do *not* change when the time dimension is introduced (§5.3.3). The Chapter closes by surmising that such a position could itself be fallacious (§5.3.4). In this case the misconception might arise should it be forgotten that, where capacity is non-fungible, current and future capital outlays will be intertemporally interdependent. As such, even fixed costs—and not just the physically common costs of capacity—might contribute to *intertemporal incremental costs*, in which case such a fallacy may disadvantage consumers rather than producers, particularly if incremental costs are estimated in their own right, rather than being derived from complementary *intertemporal stand alone costs*. Given the apparent misapplication of the time dimension in constrained market pricing, the remainder of the thesis mostly focuses on the question of the *relative* efficiency and fairness of price levels in a *temporal*—rather than a locational—sense. This is similar to examining *absolute* price levels over time, although in this case, the influence of the intertemporal interdependence of costs is explicitly taken into account.

1.4.3 Subsidy-Free Prices and Optimal Investment: Toward Static or Intertemporal Equity?

Chapter VI opens by examining how ‘*spare capacity*’ and ‘*asset duplication*’ can form part of the optimal investment program for a power distribution network (§6.1.1). In particular, as Boiteux observed, under many circumstances it can be optimal to construct network capacity in *anticipation* of future demand. Yet Baumol and Sidak are quite clear; the constrained market pricing approach sets bounds on subsidy-free prices based on estimates of stand alone costs and incremental costs *ignoring* demand, whether current or future (§6.1.2). The time dimension is mainly accommodated in constrained market pricing by applying the zero economic profit constraint on a *net present value* basis (§6.1.3). This acknowledges that returns *on* and returns *of* capital are not necessarily smooth over time. However, it fails to recognise that the intertemporal interdependence of costs “sunk” in non-fungible assets could cause the optimal investment rule (as outlined by authors like Turvey) to result in an asset configuration different from that which a hypothetical entrant would construct to capture and optimally serve all or part

of the distribution services market on a ‘*greenfields*’ basis (§6.1.4)—in other words, as if the market were currently unserved by *any* supply capacity. Incorporating the zone substation cost function developed in Chapter III (§3.6.2), the simple two product model with sequenced demands (§5.3.1) is used to illustrate the difference between the optimal investment rule and greenfields optimality for three potentially optimal construction configurations (§6.2)—namely, ‘*anticipatory construction*’, ‘*capacity expansion*’, and ‘*capacity replacement*’.

It is then discussed whether subsidy-free bounds on prices should incorporate any or all of: ‘*historic costs*’; asset ‘*replacement costs*’; ‘*opportunity costs*’; and cost changes due to new demand patterns or technological advances. Under constrained market pricing, this seems to depend on the extent to which those costs would be incurred by a hypothetical entrant in constructing an optimal greenfields network to serve current and future demand (§6.3). Following a brief overview of the relative merits of rate of return regulation and incentive regulation for implementing constrained market pricing (§6.4.1), it is conjectured that consumers’ interests are best protected by viewing a hypothetical entrant’s costs to be the same as the self-production costs of a coalition of current and future consumers (§6.4.2). These costs determine the *intertemporal anonymously equitable* prices. The Chapter finishes by highlighting that the perspective of firms and consumer groups differ most significantly in terms of the applicable *timeframe* for supply or demand (§6.4.2). Whereas a single firm may serve demand indefinitely into the future, consumers in a distribution network typically enter and leave the market over time.

1.4.4 Depreciation and Valuation of Network Assets: Historic Cost or Replacement Cost Based?

The relationship between capital cost recovery over time, and the ‘*depreciation*’ of ‘*economic asset value*’—both inside and outside the framework of contestability theory—is examined in the first part of Chapter VII (§7.1-§7.2). In particular, policymakers have grappled with the problem that, unlike firms in a competitive market—where value is dictated by the future stream of market-determined prices—for a regulated monopolist, the concepts of “value” and “making a fair return on investment” exhibit a strong element of circularity (§7.1.2). Consequently, the valuation methodology utilised by the firm, and whether this is based on *historic* asset costs, or asset *replacement* costs, has a strong impact on the prices which it sets. However, it shown that, at least in the case of a simple *single* product model with *no* demand growth, this circularity can be broken; the depreciating economic value of a firm’s assets can be derived from the intertemporal anonymously equitable prices, whereas the firm *as a whole* maintains its overall economic value (§7.3.1-§7.3.2). In contrast with this result, the relative efficiency of ‘*back-loaded*’ or ‘*front-loaded*’ depreciation schedules are discussed (§7.3.3-§7.3.4). The Chapter closes by describing the New Zealand Government’s regulatory-imposed valuation methodology for electricity distributors (§7.4)—optimised deprival valuation (ODV)—and the various critiques of this approach (§7.5).

1.5 Thesis Outline: Pricing and Depreciation Model, and Conclusions

1.5.1 A Two-Good/Two-Period Model of Intertemporal Subsidy-Free and Sustainable Prices

Whether the circularity between valuation and pricing can also be broken for a *dual* product monopolist where demand is *not* constant is examined in Chapters VIII and IX. Baumol, Panzar and Willig's single-good/two-period model of an intertemporal unsustainable natural monopoly is used as the starting point for this inquiry, and it is demonstrated that when the condition for price sustainability is met—namely that fixed costs are substantial—a set of intertemporal subsidy-free revenues can be derived for the monopolist (§8.1.1-§8.1.4). While BPW's model was only applied to a construction configuration of capacity expansion, bounds on subsidy-free revenues can likewise be derived from the model should a program of anticipatory construction be optimal (§8.1.5). Moreover, in this case, unlike capacity expansion, a feasible asset resale price can exist between the two periods, even if the asset is non-fungible. Hence, a consumer group does not need to incur the full stand alone cost of any asset it utilises, if it can resell that asset to subsequent consumers (§8.1.6). The effect that this has is to *reduce* the cost of self-production from a consumer perspective to—what this thesis terms—the '*net intertemporal stand alone cost*' (NSAC) of self-production. By complementarity (i.e., §4.3.2), this results in a '*net intertemporal incremental cost*' (NIC) of self-production, *greater* than the IC which would be found by neglecting the possibility that even non-fungible assets—which have no value in alternative *uses*—may have a value in resale to alternative *users* (i.e., §7.3.4).⁴ Consequently, where resale between successive consumer groups is feasible, the bounds on subsidy-free revenues can narrow considerably, making some of the criticisms levelled against constrained market pricing (i.e., §5.1.1) less apt.

BPW's model is then extended to two goods (§8.2) where demand for each good is *sequenced* to start at the beginning of each period. In addition, their assumption that the two periods in the model are, in total, shorter than asset lifetime is relaxed, by assuming that the demand for both goods continues in perpetuity (§8.3). As a result, capacity replacement also becomes a potentially optimal construction configuration. Significantly, the extension of the model to two goods demonstrates that the unsustainability inherent in BPW's model is largely due to their specification of a *single* product model in the first place (§8.4.1-§8.4.3). Besides, there are a number of other reasons why any concerns regarding unsustainability can be set to one side, not least that contestability theory seems to provide a stronger analytical framework for examining industry structure than it does for predicting the outcome of strategic market behaviour (§8.4.4-§8.4.5; and §8.5).

⁴ The stand alone cost of providing the service is simply the cost incurred if there were no demand in the future. The terminology used here of "net intertemporal stand alone cost" does not imply that the stand alone cost itself actually changes as a result of resale options, but is a useful shorthand for recognising the relationship between "true" SAC and the effect intertemporal interdependence has on constraining subsidy-free prices.

The intertemporal subsidy-free revenues derived from this two-good/two-period (TGTP) model (§8.2-§8.3) are then converted into intertemporal subsidy-free and anonymously equitable *prices*, for each of the three potentially optimal construction configurations (§9.1). From these prices, the economic asset values and depreciation schedules are derived (§9.2). The relevance of the model to the line charges of electricity distributors in New Zealand is assessed, by substituting the zone substation cost function used in earlier sections of the thesis (i.e., §3.6 and §6.2) into the TGTP model. Price paths based on the intertemporal subsidy-free approach, on long run incremental cost (i.e., §4.1.3 and §4.1.5), and on the optimised deprival valuation methodology mandated by the New Zealand Government, are then compared (§9.3.1-§9.3.2). This analysis is used to consider whether ELB line charges in New Zealand exhibit the characteristics of efficient and “fair” prices—in other words, whether line charges are intertemporally subsidy-free (§9.3.3). Finally, the main body of this dissertation concludes by outlining some implications of the analysis for regulatory policy and some limitations of the TGTP model (§9.4).

1.5.2 Key Conclusions

Key conclusions drawn throughout the text are summarised in Chapter X, and throughout, some directions for further research are highlighted. In particular, the closing Chapter evaluates the relative significance of the pertinent cost concepts and pricing principles, in light of this dissertation’s response to the question posed at the outset (§1.1.1). From the application of the two-good/two-period model developed in this thesis, it appears that efficient and fair line charges are characterised by providing the incumbent firm with a return on and return of what *would* be the network design, were the current asset configuration optimally matched to current and future demand, from the perspective of current and future consumers. While this conclusion bears marked similarities to that of Boiteux, the approach differs by explicitly taking into account asset indivisibilities, as well as *both* producer and consumer perspectives of the costs involved in an optimal investment program. Furthermore, this pricing approach also differs from the implementation of constrained market pricing as it originates from contestability theory, because the approach explicitly acknowledges that investments which are “sunk” (i.e., non-fungible) exhibit intertemporal interdependence of costs, and still have an opportunity cost—and thus a positive price—should the assets involved still have some value to future consumers.

Where spare capacity would be optimally built today—in anticipation of future demand, and as the result of declining average incremental costs—then today’s prices should cover the amortised opportunity cost of the total capacity required to meet both current and future demand. Where capacity does not require expansion or replacement until some later date, then initially prices only need to cover the amortised opportunity cost of the capacity optimally required to meet current demand alone, but they should rise to the opportunity cost of total capacity at that time when it would become optimal for consumers (supplying themselves on a greenfields basis) to construct capacity sufficient to meet both current and anticipated demand. These results reaffirm Boiteux’s position that spare capacity has its own income (§6.1.1), as well as Turvey’s view that the expectation of lower costs in the future raises today’s

prices (§7.3.3), providing—in some circumstances—justification for accelerated depreciation. However, under New Zealand’s light-handed information disclosure regime for electricity line businesses, line charges and depreciation schedules do not appear to have exhibited the characteristics of efficient and “fair” prices embodied in the principles of intertemporal subsidy-free pricing.

CHAPTER II

NEW ZEALAND'S POWER SECTOR REFORMS: TOWARD ALLOCATIVE OR DYNAMIC EFFICIENCY?

Any restructuring of the vital electricity sector must be carefully planned and implemented. It should not lose sight of the main objective, to improve efficiency. ... There are three key aspects of efficiency forms that electric power companies will need to achieve in economists' terms: productive efficiency; dynamic efficiency; and allocative efficiency: New Zealand Minister of Energy, John Luxton (1991a)

The Government's overall objective is to ensure that electricity is delivered in an efficient, fair, reliable and environmentally sustainable manner to all classes of consumer. To meet this objective, the Government favours industry solutions where possible, but is prepared to use regulatory solutions where necessary. This Policy Statement sets out the Government's expectations for industry action. An Electricity Governance Board should be established to ensure that the provision of electricity services is contestable wherever possible: Government Policy Statement – Further Development of New Zealand's Electricity Industry (New Zealand Government, 2000)

Regulations aimed at monopoly pricing are only justifiable, in terms of efficiency, if they produce gains in allocative efficiency that outweigh losses of productive and dynamic efficiency: Executive Director of the New Zealand Business Roundtable (Kerr, 1999, p. 4)¹

Electricity distribution is only one of the complementary activities of electricity supply. Hence, any examination of the reforms to New Zealand's power distribution sector is more readily understood in the context of both the restructuring of the electricity supply industry as a whole and the policy framework for the overall energy sector. Consequently, although the focus is still on electricity distribution, this Chapter provides an overview of the motivation, objectives, and rationale behind the reform of New Zealand's power sector from 1986 until late 2001, as well as the Government's desired outcomes and specific policy actions. This Chapter also examines the key criticisms to which the reforms have been subjected.²

¹ This quote is abridged and altered for clarity, but the meaning has not been changed.

² For the chronological details of New Zealand's power sector reforms, this Chapter is indebted to the document prepared by the Energy Markets Policy Group (2001a). Earlier overviews by the present author are also drawn upon (i.e., Gunn, 1996 and 1998).

2.1 Economic Rationale for New Zealand's Power Sector Reforms

2.1.1 Economic Liberalisation and 'Light-Handed' Regulation

Over the past twenty or so years, and beginning in Chile, the electricity supply industries of numerous nations have experienced dramatic transformations in their governance frameworks: changes which have encompassed ownership, organisational structure and commercial relations. However, these power sector reforms have not been made in isolation, and need to be viewed in light of the much wider liberalisation of many sectors in developed and developing economies that has been taking place in parallel. This process of liberalisation has aimed to relieve legal or regulatory constraints on market behaviour that are seen as being more efficiently provided by the workings of the market mechanism itself. As Berg and Tschirhart (1995) express it, "the most fundamental regulatory issue is whether firms should be regulated in the first place".

In New Zealand, such liberalising reforms began in earnest in 1984, with the accession of Prime Minister David Lange's Labour Party Government to the Treasury benches, and these reforms rapidly spread throughout a wide range of sectors (e.g., Bollard, 1991).³ This process resulted in New Zealand moving from being one of the most regulated OECD economies—under the stewardship of Prime Minister Robert Muldoon's successive National Party governments—to one of the least regulated, within the span of a few years. The key benchmark against which to test the success of reform outcomes has been the concept of '*economic efficiency*' (§2.1.2), and the key question for policy makers has been the appropriate level of intervention in a particular sector required to ensure gains, rather than reductions, in efficiency. For utilities, particularly network industries such as power, telecommunications, gas and water, this has usually meant that those functions which are seen as being '*potentially-competitive*' (§2.1.2) should be governed by the market, whereas the remaining '*naturally-monopolistic*' (§2.1.5) elements of the industry in question should still be regulated by the State.

However, New Zealand's innovative approach to the governance of many sectors, particularly for power, has differed markedly from that of most other western countries, due to its highly liberal and non-interventionist style of regulation (e.g., Bollard and Pickford, 1995; OECD, 1997, pp. 75-89).⁴ The

³ Governments in New Zealand have a three year term. The Labour Government elected in 1984 lasted two terms, but in the 1990 election a National Party Government came to power. In 1993, New Zealand's electoral system changed from the Westminster "first past the post" system dominated by two parties (the National and Labour Parties), to mixed member proportional representation. The election of that year, as well as 1996, saw National form coalition governments, but power went to a Labour-led coalition in the 1999 election.

⁴ This could partly be seen as a backlash to the Muldoon years, explained by a past Deputy Prime Minister in the Muldoon Government as follows: "Over the past decade, New Zealanders have acquired what is now a highly-developed aversion to Government regulation. ... Those who prefer the 'comfortable collusions' that are inherent in a regulated and/or controlled and/or protected business environment will not necessarily enjoy the New Zealand approach; but it must be observed that, so

emphasis has been on a ‘*light-handed regulatory regime*’ for sector monopolies, based on ‘*information disclosure*’ and ‘benchmarking’. As the International Energy Agency recently expressed it: “New Zealand’s approach to energy sector regulation is unique” (IEA, 2001b). By contrast, the more common forms of regulation are predominantly ‘heavy-handed’ by comparison, and most OECD countries generally require an industry-specific regulatory body to manage the regime. These include imposing controls on costs—including the cost of capital—through ‘*rate of return regulation*’, such as has traditionally been the case in the United States, or on prices through ‘*price cap regulation*’, as in the United Kingdom (§6.4.1), as well as ‘*yardstick regulation*’ (e.g., Shleifer, 1985). These are all types of ‘*incentive regulation*’; regulatory schemes which seek to provide incentives for the regulated entity to perform well in pursuit of its self-interest (e.g., Laffont and Tirole, 1993). Incentive regulation acknowledges the problems inherent in ‘*asymmetric information*’; namely, that the regulator almost always has access to less information than the regulated. The basic objective of incentive regulation is to decouple a utility’s price structure from its own reported costs, thus preserving incentives for the firm to act efficiently in its own interest while meeting regulatory objectives (Evans, 1998).

Small (1999c) suggests that, in New Zealand, “an active rejection of traditional regulatory methods was the primary motivation for light-handed regulation”. The approach has been industry ‘self-governance’, and successive governments have made it clear that the philosophy underlying its light-handed regime for the power sector is: the electricity supply industry should be encouraged to regulate itself (e.g., Energy Policy Group, 1994e, p. 1). As Evans (1998) expresses it, in light-handed regulation, “the governance structure is provided by the operations of private sector firms and the contractual relationships associated with the market. The role of government is confined to the establishment of a framework for property rights and competition policy. This structure includes voluntary industry self-regulation and joint ventures between competing firms. The key feature of this regime is open entry and an underlying presumption that competition is desirable. Light-handed regulation precludes statutory restrictions on entry and on-going specific regulation based on the state of the market”. Consequently, New Zealand’s approach has been cited as a “natural experiment” for testing “the appropriateness or not of leaving operational details to be negotiated by the parties rather than mandated by the regulators” (Bergara and Spiller, 1996, p. 5).⁵

far, no need for regulation (or, in the case of natural monopolies, greater regulation) has been demonstrated. More difficult to answer is the question: What would happen, and how would the Government respond, if such a need was demonstrated in the future?”: (McLay, 1993).

⁵ Australia also has a so-called “light-handed” regulatory regime governing the access of potential competitors to existing telecommunications networks (§5.1.2). Legislation requires that the party which owns the existing facility negotiate with an access seeker, subject to either party having the right to refer the matter to arbitration at any time. King and Maddock (1998) describe this approach as an “intermediate” light-handed regime lying between New Zealand’s regime—which does not

2.1.2 *Economic Efficiency, Perfect Competition and Marginal Cost Pricing*

Economic efficiency has been the primary and consistent objective for the reform of New Zealand's power sector (§2.3.1), although the primacy of this goal has been watered down slightly in recent years (§2.3.3). The high water mark for discussions concerning economic efficiency is the concept, taken from neo-classical economic theory, of '*Pareto efficiency*' (or '*Pareto optimality*'). A Pareto-efficient allocation of society's resources is one where no individual can be made better off without making at least one other individual worse off (e.g., Hay and Morris, 1993, p. 566).⁶ The two fundamental theorems of neo-classical welfare economics are that: every perfectly competitive market equilibrium is Pareto-efficient; and conversely that, every Pareto-efficient allocation of resources can be achieved from allowing the perfectly competitive market mechanism to work, with respect to some initial distribution of the given resources. Herein lies the attraction of competitive markets, although real-world markets rarely come close to satisfying the pre-conditions for the existence of a *perfectly* competitive market.

'*Perfect competition*' requires that: (a) markets are not incomplete—in other words, there is no unfulfilled demand; (b) neither producers nor consumers individually have any influence over market prices (i.e., no participant has 'market power'); (c) any negative (or positive) externalities are internalised into marginal social costs (e.g., costs of environmental damage associated with the production of a particular good); and (d) all participants in the market have access to 'perfect information', and act according to the economic model of 'rational behaviour'. As any standard first year microeconomic textbook explains (e.g., de Serpa, 1988, pp. 259-260), in a perfectly competitive market at equilibrium, the price of any good will equal the '*marginal cost*' of supplying that good (§4.1.1).⁷ Such an outcome provides a "fair return" for firms—one that provides a return on, and return of, all capital employed—as long as they are productively efficient. But fortuitously, this outcome is not only beneficial from a producer's point of view, but also from society's as a whole, since overall welfare—the sum of 'producer surplus' and 'consumer surplus'—is maximised, an outcome which ensures economic efficiency.⁸ Hence, economists sometimes equate "economic efficiency" with "*the public interest*" (e.g., Baumol and

include arbitration, only appeals under generic pro-competitive legislation (§2.1.3)—and more heavy-handed regimes, where a regulator directly sets the access price.

⁶ However, the Pareto principle requires that no interpersonal comparisons of utility be made, in other words, it is assumed that one person's utility is not affected by the utility enjoyed by another (e.g., Hay and Morris, 1993, p. 566).

⁷ In the standard textbook comparative statics diagram of supply and demand, this is the point at which the negatively-inclined consumer demand curve intersects the positively-inclined producer supply curve. The supply curve is itself the marginal cost curve.

⁸ Producer surplus is easier to measure than consumer surplus, since it is usually considered to be equivalent to a firm's net revenue (e.g., Crew *et al.*, 1995).

Sidak, 1994a, p. 26); an equivalence whose roots are often attributed to Adam Smith's celebrated "invisible hand".⁹

In practice, economic efficiency is often defined as comprising three main components: (i) '*allocative efficiency*' (or pricing efficiency), relating to the allocation of scarce resources between competing uses—it is often associated with the prescription that prices equal the marginal costs of producing those goods and/or services throughout the entire economy; (ii) '*productive efficiency*' (or technical efficiency), relating to the minimisation of the cost of production processes; and (iii) '*dynamic efficiency*', relating to the optimality of production and investment decisions *over time* (e.g., in a New Zealand context: Gallagher and Lewis, 1988, pp. 73-75; Ministry of Commerce and The Treasury, 1995, para. 39; Pickford, 1996; Williamson and Mumssen, 2000, p. 2; Commerce Commission, 2001, pp. 22-23).¹⁰ Under conditions of perfect competition, these three efficiency forms coincide, with each other, and with Pareto efficiency. Even for those who do not necessarily view *perfect* competition as the appropriate benchmark to which actual market behaviour should be compared, these three elements of efficiency, and their link with competition in its more general sense, is still basically the same.

Competitive markets are the preferable economic mechanism for achieving allocative, productive, and dynamic efficiency. Allocative efficiency is present when goods and services are allocated to the uses in which they have the highest value. Productive efficiency is present when producers use goods and services in such a manner to minimize costs, subject to technological constraints. Dynamic efficiency refers to decisions made over time and includes efficiencies in investment and technological innovation (Sidak and Spulber, 1997, p. 522).

2.1.3 Efficiency versus Competition

Acknowledging this key link between competition and efficiency, key generic legislation was enacted in New Zealand to prohibit certain business and market conduct which restricts competition—the *Commerce Act 1986*. The Act also contains provision for Government price control of markets where competition is limited. Any decision to impose price control is ultimately made by the Minister of

⁹ The roots of the invisible hand concept, however, go back further. Routh (1975, p. 75) explains that: "Quesnay repeats the idea first found in North, and also repeated by Boisguillebert, that each man, by pursuing his own interests, promotes the good of all. 'The whole magic of a well-ordered society is that each man works for others, while believing that he is working for himself'". Adam Smith's later statement was that: an "invisible hand" guides every individual so that "by pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it" (quoted in Sidak and Spulber, 1997, p. 523).

¹⁰ Sometimes a distinction is made between "investment efficiency" and "dynamic efficiency", where the former simply relates to efficiency of investments in durable assets, while the latter relates to the speed with which more efficient production techniques are adopted. Firms may do this through innovation or by adopting the ideas and/or new technologies of other firms in the industry (e.g., Murray *et al.*, 2001). In this thesis, the term "dynamic efficiency" is used to refer to both the former and latter concepts of efficiency over time, although the emphasis is predominantly on the efficiency of *investment*.

Commerce (from 2000, this post was renamed the Minister of Economic Development, and the Ministry renamed accordingly), or in the case of the energy sector, the Minister of Energy. However, the investigation leading to price control, and its subsequent implementation, is the responsibility of a quasi-autonomous generic regulatory body, the New Zealand Commerce Commission.¹¹

Some consider that the path liberalisation has taken in New Zealand has not really focused on increasing economic efficiency itself, but rather on the *means* to that end (e.g., Noble, 1991). As a result, unfettered (or ‘monopolistic’) market behaviour has been permitted in the guise of promoting ‘competition’. On the other hand, others consider that the defence of apparently anti-competitive practices on efficiency grounds under the *Commerce Act 1986* has become quite well established (e.g., Hazledine, 1994, p. 1; Patterson, 1995). As a result, market power in some cases may have remained unchecked because concerns about the possible lessening of competition have been outweighed by the assumption that the market, left to its own devices, will form its own efficient organisational structures. Although both opinions share the view that the economy is too ‘*laissez faire*’, and that abuse of market power is common, nevertheless, the latter position seems more credible. As Bollard and Pickford (1995) have expressed it: “Key features of New Zealand’s competition law are the narrow focus on a single well-defined objective—the promotion of competition as a means of increasing efficiency, with efficiency over-riding competition where the two conflict”. Consequently, the *Commerce Act 1986* “does not favour the rights of any particular group of people, such as consumers disadvantaged by monopoly pricing”.¹²

2.1.4 The Long Run, the Short Run, and Dynamic Efficiency

Apart from this tension between efficiency and competition, there is an additional tension between the components of allocative and productive efficiency on the one hand, and dynamic efficiency on the other. The first two differ from the latter component in that they are ‘*static*’ concepts. In neo-classical economic theory, the time dimension is generally addressed through the standard abstractions of the ‘*short run*’ and the ‘*long run*’. Both of these concepts are, however, inherently static in nature. The ‘*short run*’ refers to changes in supply and demand where no change in productive

¹¹ The light-handed regime has been described as: “charting a ‘middle course’ between *laissez-faire* and direct regulation. *Laissez faire* risks significant resource misallocation and inefficiency from monopoly pricing and other behaviour of firms left to their own devices. On the other hand, inefficiencies may arise from direct regulation due to the cost of the regulatory body, compliance costs, asymmetric information, as well as losses in dynamic efficiency from regulatory control of industry structure and competition, which may inhibit entry, competition and innovation” (Bollard and Pickford, 1995). The ‘asymmetric information’ problem in regulation is similar to the ‘principal agent’ problem of owning and managing a firm (§2.1.6), whereby the regulator can be considered the “principal”, and the regulated firm the “agent” (e.g., Hay and Morris, 1993, p. 629).

¹² This is because the Act does not prohibit a dominant firm from actions which are not anti-competitive. Since a monopolist by definition faces no competition, monopoly pricing *per se* cannot be considered anti-competitive.

capacity is required, whereas in the ‘*long run*’, productive capacity can be adjusted. When capacity can be adjusted, marginal and average cost relate to long run marginal and average costs. However, time is not explicitly taken into account in these concepts. The ‘*static long run*’ is no “longer” than the ‘*static short run*’; both cases represent a single point in time, or along a steady-state (e.g., Teplitz-Sembitzky, 1992, p. 25). As such, Hay and Morris (1993, pp. 568 and 660) observe that “there is every reason to doubt the dynamic efficiency of competition”, and therefore “policy needs to discover the reality, in a capitalist system, of the trade-off between short term and long term”.

Not surprisingly then, there are those who consider that, in policies directed toward real-world markets, too much emphasis is placed on static concepts, and not enough attention is given to the trade-off between static and dynamic efficiency forms. As is described shortly (§2.5.2), this concern has also been extended to the reforms in New Zealand’s power sector.

In markets where politicians and officials have concerns about efficiency, social-welfare maximising public policy will focus on dynamic efficiency and any impediments to it. This is because allocative and productive inefficiency will not persist in a dynamically efficient market, but policies focused on allocative and productive efficiency may have the unintended effect of reducing dynamic efficiency. ... [T]here is a substantial risk that regulations aimed at perceived static efficiency problems will in the long run have the net effect of reducing efficiency in the market as a whole. The need to explicitly consider the impact of proposed policy on the dynamic efficiency of a market is enhanced by the existence in New Zealand of a relatively short electoral cycle. Three year parliamentary terms may focus political interest on policies that address perceived short-term allocative and productive efficiency concerns even though these policies reduce economic growth and welfare in the long run (Evans *et al.*, 2000).

2.1.5 Monopoly Pricing and Natural Monopoly

The standard neo-classical model of the economy demonstrates that monopoly pricing is undesirable, and this is why monopolies are considered to require some regulatory controls on their behaviour. This undesirability emerges from a comparison of the market outcomes when a monopoly is free to set its own prices, with the outcomes resulting from the benchmark of perfect competition.

Monopoly pricing throws a spanner in the works of the ideal perfectly competitive world. By definition, the existence of monopoly pricing in the economy prevents a Pareto-efficient competitive equilibrium from occurring, and as such has traditionally been seen as a source of ‘*market failure*’ potentially warranting some sort of market intervention by government.¹³ If left to its own devices,

¹³ The general equilibrium of the economy cannot be Pareto optimal because any monopoly pricing behaviour will distort the market such that, marginal rates of substitution for consumer consumption of goods, will not equal the marginal rates of transformation for producer production of goods (e.g., de Serpa, p. 459).

neo-classical theory indicates that a profit-maximising monopolist will maximise producer surplus alone, and not the welfare-maximising combination of both producer and consumer surplus. Assuming all consumers face the same price, the price which will maximise producer surplus does not occur at the intersection of the supply curve (i.e., the marginal cost curve) with the demand curve, but at the output level where the supply curve cuts the marginal revenue curve instead. The monopolist would therefore obtain a level of profit higher (i.e., monopoly ‘rents’) than is necessary for “fair” compensation of capital employed (i.e., a ‘normal’ or ‘*zero economic profit*’; §6.1.3). Not only would this result in a higher price to consumers than would be the case under perfect competition, but it also causes supply to be curtailed with some demand remaining unfulfilled. (This is because neo-classical theory assumes downwardly-sloping demand curves; consequently, *ceteris paribus*, an increase in price results in a reduction of demand).

Furthermore, a wealth transfer ensues from consumers to the monopolist, and an overall welfare (or ‘deadweight’) loss is experienced by the economy as a whole. The monopolist can make this wealth transfer from consumers even greater if it is free to ‘*price discriminate*’—in other words, to charge different prices to different consumers for the same good, even though the same cost conditions relate to each consumer (§4.2.2). However, Williamson and Mumssen (2000, p. 5) caution that: “While the problem of monopoly is essentially static (prices above costs), the problems of regulation are dynamic, relating to long-term investment”. In other words, regulatory intervention to address problems of monopoly pricing needs to also consider impacts on investment (§6.1).

Historically, monopolies have often emerged intentionally as a result of protective regulation, such as through the legal right to the exclusive franchise for the supply of a particular good in a particular market. For instance, prior to the onset of power sector reform in New Zealand, no agency could generate electricity without the direct approval of the Minister of Energy, and the community and municipally-owned electricity distributors were protected by legally-defined geographical franchise boundaries. But monopolies can also arise ‘naturally’, namely where a single firm can produce the total market demand at a lower cost (i.e., most productively and dynamically efficient) than any combination of two or more firms (e.g., Sharkey, 1982a, p. 73).

Although this definition of ‘*natural monopoly*’ has not changed with time, the cost characteristics considered to signal its existence have (§3.4.4). Traditionally, natural monopolies have been associated with industries exhibiting ‘*economies of scale*’ (§3.4.2), and these have been identified by examining the industry cost curve. If an industry was subject to declining average costs in the long run, this was seen as evidence of naturally monopoly. Where a natural monopoly is present, competition in the traditional sense would require that each competitor duplicate the facilities of the monopolist, which would be dynamically inefficient. However, although an industry as a whole might exhibit economies of scale, it

became recognised by economists during the late 1970s and early 1980s that, in a complex utility industry involving one or more joint products, it might be possible that a subset of those products did not.

In cases of natural monopoly, the concern has traditionally been the opposite of protected monopoly. For monopolies where long run average costs are increasing, long run marginal costs will also be increasing and will be greater than average cost. Moreover, the profit-maximising price will be greater than the marginal cost under a perfectly competitive equilibrium (§4.1). But for a natural monopolist, the presence of declining long run average costs means that marginal costs will be lower than average costs. To recover its costs the natural monopolist has to price above marginal cost (i.e., at or above average cost) or a private sector firm will go out of business. Market demand is insufficient to support an efficient-sized plant, and the only way to escape welfare losses is to require the natural monopolist to price at marginal cost, and for the Government to subsidise the monopolist to a level equivalent to pricing at average cost. This was why goods and services produced by natural monopolies were historically perceived as requiring public, rather than private, provision (e.g., Blaug, 1990). Moreover, the neo-classical model of general equilibrium that provides the basis for Pareto optimal resource allocations assumes that technology exhibits non-increasing returns to scale, and also that commodities are perfectly divisible (e.g., Spulber, 1989).

The situation is more complex with a *multiproduct* firm, since to talk about the ‘average’ costs of producing two or more goods which are dissimilar and cannot be aggregated is not meaningful. (The analysis of the multiproduct natural monopoly is elegantly handled by contestability theory, discussed in the following subsections). In any case, more rigorous analyses of the cost conditions for a natural monopolist have show that, even for a simple *single* product firm, declining average costs might be a sufficient test for the existence of natural monopoly, but it is not a necessary one. In a single product firm it is quite possible for long run average costs to be increasing at the point of current industry output—in other words, marginal costs are greater than average costs—yet the sole firm’s cost structure still be consistent with that of a natural monopoly (Sharkey, 1982a, p. 85). This situation is, at least in theory, problematic, because if this occurs, a competitor can potentially undercut the monopolist’s price by serving less than the entire market demand, while still making a profit. A monopoly facing this type of industry cost structure is termed an ‘*unsustainable* natural monopoly’ (§2.1.8). Perversely, for an unsustainable natural monopolist, monopoly and competition are not mutually exclusive. In such a case, the prescription for Government intervention is not just to subsidise the natural monopolist, but also to raise market entry barriers to protect it from competition which would be inefficient.

In New Zealand, monopoly pricing of utilities in the energy sector (i.e., power and gas) has been controlled in two ways. Firstly, previously protected monopolies that were potentially competitive have been exposed to competition. Where potentially competitive and naturally monopolistic business functions have been combined in the same vertically-integrated organisation, they have been

‘unbundled’, either through complete ownership separation, or through transparent information disclosure under the light-handed regulatory regime (§2.1.1). Secondly, the light-handed regulatory regime has itself been used to expose naturally monopolistic organisations to public scrutiny (Energy Policy Group, 1995). The regime has been designed with the intention of highlighting overt abuses of monopoly power, and to “shame” monopolists into not taking advantage of their dominant position.

2.1.6 *Transaction Cost Theory and Contestability Theory*

Market failure, such as that due to monopoly pricing, has traditionally been used as a justification for government intervention in particular markets. However, developments in economic theory and practice during the 1980s had a profound effect on influencing the attitude of policymakers world-wide, and in New Zealand in particular, away from intervention and toward liberalisation. The burden of proof now became for Government to justify any interventions in the freely functioning market mechanism, rather than for firms to demonstrate to Government that their activities did not constitute an abuse of market power. These developments in economics criticise the perfect competition model as not fully representing or explaining reality, and contrast any potential for market failure with the likelihood of *non-market* or bureaucratic failure—in other words, intervention by government making the economy less efficient (e.g., Bollard, 1991). As Gale (1989, p. 39) expresses it, the “onus is now on the bureaucracy to establish that policy improves outcomes by more than the cost of the bureaucracy”, even where market failure appears to be the case.

One of these developments in economics, transaction cost theory, or ‘*transactions governance*’—expounded by Oliver Williamson (1986a-b)—focuses on the costs of running the institutional system that allocates resources. The perfectly competitive model of the economy assumes that the transaction costs between firms, between firms and individuals, and within firms, are zero. Transaction costs—a concept going back to Coase (1937)—comprise: the search and information costs in setting up a market exchange; the bargaining and decisionmaking costs in making an agreement; and, the monitoring and policing costs of making sure a contract is honoured. (This includes the costs inherent in the relationship between ‘principal’ and ‘agent’; e.g., Vickers and Yarrow, 1988, pp. 92-101). Transaction cost theory suggests that, given the prior establishment of property rights, the market produces governance structures that minimise not only production costs, but transaction costs as well. As such, it provides the basis behind arguments in favour of privatising state-owned corporations. Exposing companies to the sharemarket is cited as resulting in organisational structures which minimise transaction costs, as well as in better investment decisions, for it promotes greater accountability of management; the threat of takeover or bankruptcy can subject firms to powerful disciplines (e.g., Deane, 1989).

However, for many of those in favour of using the market mechanism to achieve economic efficiency, the ownership of firms is a secondary issue.¹⁴ The behaviour of firms—especially in regard to price setting—is the key, and the important consideration is the ability of the regulatory environment to promote competition. Under the perfect competition model, traditional pro-competitive regulation—epitomised in the early anti-trust legislation of the United States—concentrated on the number of firms in an industry. By contrast, the theory of ‘*contestable markets*’—which emerged in the 1970s, and grew in influence during the 1980s—suggests that it is difficult for firms to adjust prices quickly, and therefore the threat of new entrants can even force monopolists to set prices at, or close to, a competitive level (i.e., at marginal cost). Williamson links transactions governance and contestability theory together through the concept of ‘*asset specificity*’, which occurs when capital is employed in durable and immobile investments that have specific uses (§3.5.1). Contestable markets are those *without* this characteristic, thus: “contestability theory and transaction cost economics are looking at the same phenomenon—the condition of asset specificity—through the opposite ends of the telescope” (Williamson, 1986a, p. 56).

Contestability theory—in conjunction with transaction cost theory—has had a major influence on both the theory and practice of regulation (e.g., Berg and Tschirhart, 1988; Spulber, 1989; Baumol and Sidak, 1994a; Baumol and Sidak, 1995a; Sidak and Spulber, 1997), particularly with respect to network industries. And in New Zealand, the Government's economy-wide policy choices have been strongly guided by the concept of contestable markets (e.g., The Treasury, 1984; Lloyd, 1986), not least in the power sector, as has been outlined by many commentators on reform (e.g., Culy and Gale, 1987; Gallagher and Lewis, 1988, pp. 49-51; Kask, 1988a, p. 13; and Gale, 1989, pp. 1-2). Deference to the concept can be clearly seen in the language used by industry participants and the Government during the various stages of reform (§2.3-§2.4).

2.1.7 Perfectly Contestable Markets and Subsidy-Free Prices

Proponents of the paradigm of ‘*perfectly contestable markets*’—developed by a large group of economists, but definitively articulated by William Baumol with his colleagues John Panzar and Robert Willig in the tract “*Contestable Markets and the Theory of Industry Structure*” (Baumol *et al.*, 1988, henceforth ‘BPW’)¹⁵—suggest that contestability theory can provide guidance in addressing the issues of efficient and sustainable pricing, particularly in industries involving economies of scale and/or ‘*economies of scope*’ (§3.4.2). Rather than just providing new insight into what constitutes a natural

¹⁴ For example: “Access is more crucial than privatisation, since ownership of utilities has little to do with the efficiency of their operations. ... Open access of private or public systems, coupled with efficient utility pricing, will probably succeed as much as any scheme in securing the benefits of a competitive market for utilities and their customers” (IEA, 1991, p. 32).

¹⁵ Baumol *et al.* (1988) is the second edition of a work originally published in 1982, and it is this second edition which is henceforth cited as ‘BPW’.

monopoly, BPW presented a comprehensive and integrated theory of efficient industry structure. Even some of the theory's strongest critics concede that the theory provides an impressive and comprehensive presentation of the theoretical cost characteristics of multiproduct firms (e.g., Shepherd, 1984), particularly with respect to the technological conditions for natural monopoly. Spence (1983) suggests that the main value of the contestable markets hypothesis is as an analytic technique for exploring multiproduct cost functions, while Shepherd insists that this aspect of the theory can and should be separated from the normative implications arising from the application of the concept of a perfectly contestable market, which relies on "implausible", "impossibly restrictive", and "mutually inconsistent" assumptions.

BPW state that when analysing industries which exhibit increasing returns to scale—a situation considered prevalent in many network utilities—the standard of perfect competition is totally inadequate, and a rule requiring that price be set to marginal cost would be a prescription for financial disaster (BPW, pp. 503-504). In BPW's view, perfect contestability should replace perfect competition as the appropriate benchmark for guiding regulatory intervention, with respect to both price and entry. A perfectly competitive market is viewed as just one special case of a perfectly contestable market, and perfect contestability applies with equal force to circumstances where perfect competition is impossible due to economies of scale. Besides, the model of perfect competition intrinsically pre-specifies industry structure, and therefore cannot serve as a useful benchmark for the study of the determinants of industrial structure. In contrast, "contestability theory encompasses an endogenous determination mechanism from which any industry structure may emerge" BPW (p. 487). BPW cite the classic work by Alfred Kahn (1970) on regulation as stating that the proper goal of regulation is the achievement of "as-if competitive behaviour". However, from BPW's perspective: "such a goal is a *non sequitur* because generally, as-if competitive behaviour is infeasible under the technological conditions of natural monopoly that justify regulation in the first place. In contrast, as-if contestable behaviour seems always to be apt as a standard of comparison" (BPW, p. 14).¹⁶

¹⁶ Referring to research by Laffont and Tirole (1993), as well as Spulber (1989), Baumol and Sidak (1994a, p. 45) point out that: "A different criticism sometimes directed at the competitive-market model is that it fails to incorporate the results in recent theoretical literature on regulation. This body of research makes impressive use of game theory and the theory of principal-agent relationships to construct models of the behavior of firms (and regulators) in settings in which regulators are constrained by limited information. The regulator's assumed objective need not be the scrupulous pursuit of 'the public interest, convenience, and necessity'". However, Baumol and Sidak conclude that "the recent theoretical literature on optimal regulation is still highly mathematical and abstract".

Similarly, Blaug (1992, p. 240) observes that: "The field of industrial organization ... has been transformed in recent years by the introduction of noncooperative game theory. Nevertheless, one may search high and low in such leading textbooks as Jean Tirole's *Theory of Industrial Organization* (1988), exemplifying the game-theoretic revolution in industrial organization, without encountering so much as a definite empirical prediction about market behavior".

A perfectly contestable market is one in which: (i) entry is completely free and exit is costless; (ii) entrants and incumbents compete on completely symmetric terms; and (iii) entry is not impeded by the fear of retaliatory price changes. Potential entrants might not fear retaliation because incumbents are restrained by law or other impediments from making retaliatory moves, or because entrants simply do not expect retaliation and as such maintain so-called '*Bertrand-Nash expectations*' (BPW, pp. 349-350). This latter characteristic of a perfectly contestable market, the Bertrand-Nash assumption, is one of the key areas of focus for criticisms of contestability theory (e.g., Shepherd, 1995), although some critics dismiss the theory entirely, on the basis that it is "ideology masked as science" (Sassower, 1988).

While many of the early critics of regulation, irrespective of philosophic bent, recognized that inherent structural conditions preclude the emergence of competition, contemporary critics urge that regulation be eliminated and replaced with competition or a competitive surrogate (Miller, 1995).

Answering criticism that they display "ideological bias", and that contestability theory offers "carte blanche to mindless deregulation", Baumol and his colleagues profess that "the arena in which the viewpoint of contestability theory may make its main contribution" is "as a guide for regulation, rather than as an argument for its elimination" (BPW, p.503).

Where markets are contestable, the presumption should be that intervention is unwarranted. Where markets are not contestable, intervention should be guided by the performance perfect contestability can be expected to produce (BPW, p. 479).

Contestability theory is presented by BPW as being able to offer consumers in markets with unavoidable entry barriers the same sort of protection from excessive or monopoly pricing that they would have derived were there perfect freedom of entry. They claim that, prior to the exposition of contestability theory, probably the most "noteworthy" and "curious" gap in the standard theory of policy

Nevertheless, Blaug does acknowledge that "the introduction of a game-theoretic approach to economics has brought with it a new 'understanding' of what is meant by rationality and interdependence and equilibrium".

The Ministry of Commerce and The Treasury (1995, para. D21) also state that: "There has been much research into the question of the optimal regulatory mechanism in the presence of [asymmetric information]. For example, Laffont and Tirole (1993) derive a mechanism in a context in which the regulator can observe costs (but not effort) perfectly ex post. In their model the regulated firm is subject to a contractual arrangement that provides incentives for the regulated firm not just to expand its own output but also the output of competitors. Unfortunately, these and other more sophisticated incentive-compatible regulatory models appear to be rather impractical".

Consequently, this thesis sets the parallel strand of industrial organisation most clearly articulated by Tirole and Laffont to one side. Nevertheless, Baumol and Sidak's work on constrained market pricing (Chapter V) was derived from a game-theoretic model developed by Faulhaber (1975), described later (§4.3.2). Similarly, elements of game theoretic techniques are used later in this thesis (§8.5.1) to "force" an equilibrium so that subsidy-free prices can be derived without running up against the intractable problem of unsustainability (§2.1.8).

was the lack of a defensible criterion for regulatory *ceilings* on prices, and that this gap is filled by the theory (BPW, pp. 487-8). In the presence of economies of scale and scope, contestability theory—through its articulation and expansion of earlier work on cross subsidisation and pricing in public enterprises (§4.3)—provides a benchmark test for efficient prices; prices must lie between the incremental cost (IC) and standalone cost (SAC) of supply, derived under perfectly contestable market conditions (BPW, pp. 508-9). Prices satisfying this requirement will be free from ‘*cross subsidy*’ and be free from the exercise of any market power, or—where the market is most efficiently served by a single firm—from ‘*monopoly power*’. Consequently, regulatory intervention is not warranted unless industry prices lie outside these bounds, even in industries where there is a monopoly supplier. BPW (p. 504) emphasise that the practical application of the SAC test by regulators emerged from contestability theory, and could not have been deduced from the model of perfect competition.

Perfect contestability is acknowledged by its proponents as being just as much a “fictional” construct as perfect competition (Baumol, 1986, p. 129), but the object of using the concept is to give regulators a model for the design of rules for markets which are distinctly *not* contestable (Baumol and Sidak, 1994a, p. 43). As Blaug (1992, pp. 90-97) explains, the falsehood of assumptions should not deter economists from proposing new models of the market, a viewpoint that traces its origins to the eloquent argument presented by Milton Friedman (1953). Consequently, the fact that perfect competition is an unrealistic model is not the reason why it is inappropriate for use by regulators of utility industries. The reason the model is seen to be inappropriate is because it counsels the regulator to actions that are either infeasible or undesirable, such as: (i) setting prices to marginal cost (§4.1.1), resulting in financially non-viable firms; or (ii) attempting to populate an industry with a multiplicity of firms, resulting in higher costs due to unrealised economies of scale (Baumol and Sidak, 1994a, p. 32). By contrast, contestability theory is seen as providing more appropriate guidance to policy makers in terms of both (a) the criteria used to distinguish cases where government intervention is desirable rather than undesirable, and (b) the regulatory tools that will increase the public welfare benefits of any intervention (BPW, p. 498), particularly in regard to determining limits on subsidy-free prices.

2.1.8 Sustainable Prices and Intertemporal Unsustainability

Apart from providing some guidance on efficient subsidy-free price ranges, another facet of contestability theory is its analysis of the ‘*sustainability*’ of prices. The issue of sustainable prices is not one that has received much attention by policymakers in New Zealand with respect to any sector of the economy, including the power industry.¹⁷ Nevertheless, sustainability is an integral part of contestability theory, and is intrinsically linked to the concept of subsidy-free prices. Moreover, some consider that the

¹⁷ In discussing the sustainability of network industries, Heald (1997) for one considers that unsustainability is more likely to relate to a network in its infancy, rather than to a mature network covering a large geographical area and reaching a very high proportion of potential customers.

possibility of unsustainability could have a significant bearing on the regulation of electricity distribution (e.g., Teplitz-Sembitzky, 1990). Hence, the opportunity is taken at this point to outline the key features of this issue, one that appears throughout this dissertation (particularly §3.3.3, §4.3.6 and §8.4).

Sustainable prices are ones that effectively deter entry by any competitors into an incumbent's current share of a market. Under conditions of perfect contestability, BPW state that prices cannot be sustainable if they involve any cross subsidy (BPW, p. 351). Further, an enterprise that monopolises an industry can only be potentially sustainable if it is a *natural monopoly*—one in which a single firm can produce the total market demand at a lower cost than any combination of two or more firms (e.g., Sharkey, 1982a, p. 73). Unfortunately, there is no guarantee that sustainable prices exist, even for a natural monopoly producing socially valuable products (BPW, p. 357), a problem first examined with any rigour by Faulhaber (1975), as well as Panzar and Willig (1977).

Within a static framework, '*sunk costs*' associated with the construction of non-fungible capacity are considered to act as barriers to entry (§3.5), and hence make an incumbent monopolist's prices more sustainable than they would be had the free entry/exit assumption of perfect contestability held true (e.g., Templitz-Sembitzky, 1990, p. 46). A market is only perfectly contestable if entry and exit are perfectly easy and costless, which requires that a competitor can enter without incurring any costs to which incumbents are not subject. Entry must not require the new firm to make any sunk investments; that is to make outlays that cannot be quickly and costlessly reversed and fully retrieved (Baumol and Sidak, 1994a, pp. 42-43). Where sunk costs exist, the potential arises for even a natural monopolist to abuse its monopoly power.

However, in an intertemporal context, BPW indicate that, where efficiency requires construction to be spread over time and there are scale economies in sunk construction costs—a case in which the market is likely to be supplied most efficiently by a natural monopolist—there might be *no* sustainable pricing solutions. Consequently, the market mechanism may well produce an intertemporal allocation of resources that is inefficient, causing unnecessary duplication of assets (BPW, p. 406).¹⁸ BPW appear somewhat disturbed by this outcome, declaring that “we cannot be comfortable with a standard equilibrium analysis of industry structure, or feel that the invisible hand has matters here firmly under control” (BPW, p. 429). Nevertheless, to some extent BPW attempt to explain away the lack of such

¹⁸ BPW are very strong on this point stating (p. 406): “we prove that with a vector of prices the firm that at one period of time was the sole supplier of the market will find itself *vulnerable* to being driven from the market precisely at the date when efficiency calls for its expansion. ... [G]iven the incumbent's prices, an entrant will *always* have an opportunity to take at least part of the market away from the established firm if the scale economies are sufficiently strong, and to do so before the incumbent can recoup its investment. But then, that entrant's market in turn becomes vulnerable to takeover in the same fashion, before it can recover its outlays” [emphasis by BPW themselves]. The model which BPW used to draw this conclusion is discussed in detail in Chapter VIII.

observable market instability in the real world by pointing out the implications of the Bertrand-Nash assumption and the associated assumption that demands and prices are perfectly foreseeable.¹⁹

This outcome might suggest a prescription for regulatory policy; that in cases of static or *intertemporal unsustainability*, intervention is warranted to provide protection to the incumbent natural monopolist through restrictions on entry, as such, avoiding the wasteful duplication of facilities (e.g., Templitz-Sembitzky, 1990, p. 5). It was partly concerns regarding unsustainability, or ‘destructive competition’ (e.g., Sharkey, 1982a, pp. 24-28 and Ch. 6), which—at least in the United States—was one of the justifications for the historical franchising of the supply areas of vertically-integrated power utilities. Although electricity supply as a whole was generally viewed as a naturally monopolistic activity, the traditional concern was that, without protection from competition, the incumbent natural monopoly would either be bankrupted by new entrants, or at the very least some wasteful duplication of facilities would occur (e.g., Primeaux, 1975; §3.2.1).²⁰

In contrast, although couching their discussion with a number of strong caveats, BPW draw a very different conclusion with regard to regulatory policy, making recommendations that relate to price rather than to entry. They suggest that the real lesson arising from their analysis of intertemporal unsustainability is not that incumbents are inevitably doomed to failure, but that monopolists may only be able to protect themselves, and the interests of the economy, by the very sort of strategic pricing behavior that economists have tended to deplore (BPW, pp. 406-7). Consequently, they conclude that the threat of responsive pricing and its deterrent effect upon potential entrants, far from frustrating the invisible hand, may be the only effective pricing implement available to it, because increasing regulatory restrictions on price or entry runs the risk of unintended consequences and inefficient market outcomes. On the other hand, they acknowledge that relaxing price restrictions may leave consumers at the mercy of unfettered monopolistic pricing behaviour, while lowering industry entry barriers may erode the natural

¹⁹ “In intertemporal analyses, neither the Bertrand-Nash assumption nor its denial is completely convincing in general. ... [V]irtually any scenario is possible. Potential entrants may all be cowed from entry forever by fear that they in turn will suffer the same fate that their entry imposes on the incumbent. ... Such a scenario would depict a world of equilibrium with industry structure and ownership held unchanged by the forces of fear. At the other extreme, potential entrants may grasp recklessly at every entry opportunity and so produce constant upheaval and perpetual unprofitability for every firm. There is an intermediate possibility, which seems much more plausible. When entry opportunities arise, they are not recognised and used at once. Rather, imperfect information, some degree of timidity, and other frictions ... introduce substantial lags into the process. ... [T]ypically the lag will be sufficient to permit many incumbents to not only recoup their investments, but to earn handsome profits over their lifetimes. Casual empiricism seems to confirm the reality of this last scenario” (BPW, pp. 428-429).

²⁰ This wasteful duplication justification for regulating natural monopoly is typically attributed to John Stuart Mill’s *Principles of Political Economy*, dating from 1848 (e.g., Schwartz, 1971, pp. 128-142; Sidak and Spulber, 1997, p. 21).

monopolist's financial viability by making the efficient incumbent firm vulnerable to opportunistic hit-and-run entry. As such, Baumol and his colleagues lament that:

The policy designer sails between Scylla and Charybdis, apparently with no safe middle course (BPW, p. 368).

2.2 Sectoral Rationale for New Zealand's Power Sector Reforms

2.2.1 *Electricity Supply and Electricity Distribution in New Zealand*

Electricity is most efficiently (in a technical sense) transmitted at high voltage due to the problem of electrical energy losses (§3.1.1), although for safety and other reasons, it can only be used at the home or small businesses at low voltage. The meshed bulk 'transmission network' effectively interconnects all the power generating equipment and in New Zealand operates at the extra high voltage (EHV) of 110 kilovolts (kV) or 220kV.²¹ The very largest industrial consumers are sometimes supplied directly at 110kV. The voltage is "stepped down" by transformers at 'points of supply'—currently owned and operated by Transpower, New Zealand's State owned transmission company—typically to 66kV, 33kV, and sometimes 22kV. These lines are generally owned and maintained by a number of electricity distribution companies—currently termed "electricity line businesses", or "ELBs"—and each ELB may be connected to one or more Transpower points of supply. This high voltage (HV) part of the electricity supply system is generally termed the 'subtransmission network'. Some large consumers, typically ones which require high voltage supply for industrial process requirements, are supplied directly from this part of the system. The voltage is further stepped down at many ELB 'zone substations' to 11kV and/or 6.6kV, and this system forms the 'medium voltage (MV) distribution network', typically comprising the greater part of an ELB's physical network assets. 'Consumer substations' mounted on the ground along the street, or up power poles, finally step the voltage down to 400V and/or 230V, and this 'low voltage (LV) distribution system' directly supplies homes and small businesses. However, major office blocks and many other consumers may own their consumer substation, and be connected to the ELB's network at medium voltage. The entire high voltage subtransmission system, as well as the medium and low voltage distribution systems, are often collectively termed the 'distribution network'.²²

Traditionally, the term 'electricity distribution' related to three functions: (i) the '*conveyance*' or '*delivery*' of electrical energy from the transmission network, over the subtransmission, medium and/or low voltage distribution networks, to end-use consumers; (ii) the '*connection*' of new as well as existing consumers to the distribution network at high, medium and low voltage; and (iii) the '*sale*' of that energy (measured in kWh and kW; §3.1.1) to consumers. However, the current industry paradigm is to

²¹ The North and the South Island transmission networks—which utilise AC (alternating current)—are interconnected by an HVDC (high voltage direct current) link by provided by submarine cables across Cook Strait.

²² This subsection draws on earlier material from the author (Gunn, 1996).

distinguish the latter function as a distinctly separate activity, ‘energy retailing’. As such, more properly, the term electricity distribution currently only relates to the first two of these functions.

2.2.2 Transformation of Electricity Supply Industries World-wide

The electricity supply industry as a whole traditionally comprised three distinct activities: power generation, transmission and distribution. These activities were seen to be complementary and subject to significant scale economies. As such, all these activities, or at least generation and transmission, were vertically-integrated within a single organisation, and protected by regulation from outside competition. And, with the exception of most notably the United States, where historically the majority of power utilities have been privately-owned, these supply activities were primarily the responsibility of the State, through a Government department or State-owned utility. Sometimes the State’s involvement in electricity supply was supplemented by consumer co-operatives or local government, particularly at the distribution level.

Developments such as contestability theory and the theory of transactions governance (§2.1.6) began to call into question the traditional “common sense” view—or “fundamental theorem”—concerning the naturally monopolistic nature of electricity supply, and the optimal way the industry should be owned, structured and operated.²³ At the same time, changes in technology were beginning to change the traditional cost characteristics of the industry. For many years there had been clearly evident economies of scale in the production and construction of generation equipment, leading to a “bigger is better” ethic. But technological advances during the 1970s led to the development of cheaper small scale power turbines. If power generation was no longer a decreasing cost industry, perhaps it could be opened to competition (e.g., Joskow and Schmalensee, 1983). Furthermore, advances in information and communications technology, as well as the development of spot pricing theory—which emerged from the engineering literature (§4.1.2)—were beginning to make the possibility of “real time” markets for trading electrical energy a reality.

²³ There are those who, although perhaps conceding that the new paradigm of privatisation, deregulation and contestability might be applicable to the electricity supply industries of developed countries, would challenge its applicability to developing countries, with their less mature and smaller supply industries (e.g., Teplitz-Sembitzky, 1990, p. iv; Bhattacharya, 1995). For instance, Teplitz-Sembitzky (1990, p. 2) states that: “there is no a-priori reason to believe that the solutions to the difficulties developing countries have with their power sector will necessarily lie with the range of policy measures offered by the deregulation and privatization movement. Nor is there sufficient evidence that the ‘fundamental theorem’ of power sector economics, namely that power generation, transmission and distribution exhibit essential features of a multiproduct natural monopoly, should be reconsidered if not abandoned”. Although the developing country aspect of this argument does not apply to New Zealand, it is possible that, given the small population, that the issue of small markets might. Nevertheless, the International Energy Agency recently concluded that: “New Zealand has demonstrated that electricity market liberalisation can be successful in a small country” (IEA, 2001c, p. 10).

The international “best practice” power industry structure and market operation is now very different, requiring the vertical and horizontal unbundling of monolithic old power agencies and utilities, as well as the deregulation of any constraints to free market operation to the greatest extent possible (e.g., Regulatory Assistance Project, 2000; IEA, 2001a). To the traditional functions of electricity supply have been added ‘wholesaling’ and ‘retailing’, and new commercial relationships more associated with the financial services sector, such as hedge contracts for energy supply (and even transmission capacity) have emerged. Numerous generating companies actively compete in a deregulated power market, where bulk power flows and the dispatch of generators is managed by a separate transmission company, but the wholesaling of electricity—between generators and distribution companies as well as the largest consumers—is handled by an independent market operator in a ‘power pool’.

At distribution level, any old franchise boundaries protecting distributors against new entrants have been removed, and open access regimes are in place to allow ‘retail competition’ for sales of electrical energy. Like wholesale competition, such competition does not affect the actual flows of power from generators to consumers, only the nature of the sector entities involved and the commercial relations between them, especially the financial transactions associated with electricity supply (§3.3.1). Moreover, the corporatisation and subsequent privatisation of national or local government agencies involved at any stage of the electricity supply chain has substantially increased the role of the private sector in the supply of power. Consequently, the State’s involvement is limited to that of policymaker and regulator, although best practice requires that these two functions also be at arm’s length from each other. Any regulation applies solely to the remaining part of the industry considered to have naturally monopolistic characteristics due to economies of scale, scope and density (§3.4.2); namely, transmission and distribution.²⁴

2.2.3 Efficiency and Fairness Issues with New Zealand’s Pre-Reform Electricity Supply Industry

The general paradigm shift toward liberalisation, combined with the transformations beginning to occur in power sectors internationally, would probably have been sufficient drivers for restructuring New Zealand’s electricity supply industry, even had the sector been considered problem-free. However, calls for changes to New Zealand’s existing industry model went back to the early 1980s (e.g., Low and Read, 1982). Although these earlier recommendations foreshadowed the later reform program, the proposed mechanism for implementing any reforms was firmly grounded in the predominant ‘command-and-

²⁴ Individual OECD countries have of course differed in the specifics of the manner and the extent to which “best practice” reform has been implemented. A recent review is provided by Steiner (2000).

control’ paradigm of the time, maintained by successive Muldoon governments (§2.1.1). As such, major improvements to central planning and control were proposed, rather than deregulation.²⁵

Prior to 1986, the basic thrust of government policy in the power sector had been to “get power to the people as cheaply as possible” (Culy *et al.*, 1997). To achieve this, the Government’s control over the electricity supply industry was primarily exercised through statutory monopolies. The New Zealand Electricity Department (NZED), part of the Ministry of Energy, was responsible for power generation and transmission, as well as for the direct supply of the very largest consumers (such as the Comalco aluminium smelter). Sixty tax-exempt electricity supply authorities (ESAs) conveyed and sold electrical energy between mainland New Zealand’s bulk transmission network and the majority of consumers, within set geographical franchise areas. Depending on the vagaries of history, these ESAs were generally controlled and managed by either consumer-elected Electric Power Boards (EPBs), or by the Municipal Electricity Departments (MEDs) of community-elected local governments.²⁶ Many commentators would probably agree that the heavy involvement of central and local government in the initial evolution of New Zealand’s electricity supply industry was justified, given the importance of secure energy supply for economic growth and regional development, the sparsely settled nature of the country, and the abundance of a cheap hydropower resource mostly located far from major load centres (e.g., Culy *et al.*, 1997).²⁷

But in the new environment of liberalisation, the incoming 1984 Labour Government perceived that the now-mature industry’s economically “inefficient” agencies were guilty of “featherbedding”, and had produced excessive over-investment in power generation capacity, “gold-plating” of transmission and distribution networks, and “arbitrary tariff structures distorted by cross-subsidies favouring residential consumers” (Farley, 1991).²⁸ In generation, bulk electricity was sold at below cost (Galvin, 1985), and unnecessary expenditure occurred toward the end of each financial year because there was a

²⁵ Low and Read (1982, pp. iv-vi) recommended that, in the electricity supply industry: (i) the public corporation model was the most appropriate for public energy companies; (ii) energy sector companies should be required to make a profit to improve efficiencies; (iii) because load management by ESAs was handicapped by average-cost pricing, the marginal cost of generation should be more accurately reflected in seasonal and time-of-day price variations (§4.1.2); and (iv) cross subsidies for the purpose of redistributing income should be removed (§4.3.2).

²⁶ These 60 ESAs comprised 38 “special purpose local authorities” (i.e., Power Boards) operating under the *Electric Power Board Act 1925*, 21 municipal electricity departments of territorial local authorities, and one authority directly owned by the Government (Energy Markets Policy Group, 2001a). Outside mainland New Zealand, an additional Government-owned local authority provided power supply on the Chatham Islands. This thesis only deals with the power sector on mainland New Zealand.

²⁷ The subsidisation of nascent network industries has also been cited internationally as playing a significant role in nation building and integrating young nations into a coherent whole (e.g., Heald, 1997).

²⁸ “Gold-plating” means that consumers were provided with a level of security and quality of supply (§3.1.2) higher than that which they would be willing-to-pay for, had they faced the true costs of supply, and is usually considered to arise because firms are given a guaranteed return on investments, efficient or otherwise (e.g., Williamson and Mumssen, 2000, p. 14).

widespread belief that annual budgets needed to be fully spent, or they would be reduced in future years (Culy *et al.*, 1997). Moreover, for a country which at the time had a population of less than 4 million, the large number of distributors was deemed unnecessary. ESAs were insulated from commercial pressures and were thus able to retain large workforces to perform functions that could be more efficiently outsourced—some ESAs even made their own office furniture (Mercury Energy Lines Business, 1999, p. 7). Moreover, because many entities were only required to break even, the opportunity cost of capital (§6.3.1) was not being adequately reflected in prices (Evans *et al.*, 2000).

‘*Cross subsidies*’ in power tariffs were of particular concern, and have generally been viewed by successive Governments as being implicitly “unfair”. Yet when considered in the context of economic efficiency, “*fairness*” is meant in the limited sense of ‘*horizontal equity*’, which requires that individuals subject to the same conditions (typically costs), should be treated the same (§4.3.1). This is effectively a restatement of allocative efficiency, and has been popularised in New Zealand by the term ‘user pays’ (e.g., Bollard and Pickford, 1995). One implication is that firms should not be allowed to engage in ‘*price discrimination*’; charging similar consumers different prices when they face the same costs (e.g., Hay and Morris, 1993, 161-168; §4.2.2). Prior to reform, the New Zealand Ministry of Commerce considered that cross-subsidisation was able to be sustained by the nature of Power Board and local government representation.²⁹ The favouring of residential consumers by cross subsidies was seen as being highly unusual among developed economies, with a survey at the time indicating that, of the 21 countries who were members of the International Energy Agency, New Zealand was the only country where average household tariffs were below industrial tariffs. On average, domestic tariffs in IEA countries were around 75 percent above industrial tariffs (Kask and Saha, 1986, p. 6).³⁰

Business interests, often given voice in the reports and commentaries of the New Zealand Business Roundtable (NZBR), considered (and still consider) many of the inefficiencies inherent in the

²⁹ The Ministry’s critique traces its origins to George Stigler’s (1971) economic ‘capture’ theory of regulation, which perceives politicians—in this case local government politicians—and regulators as being influenced to transfer wealth to their supporting groups. However, Kask (1988b, p. 37) suggested that the primary source for favoritism was not due to the elected nature of the Power Boards, but to the lack of clarity in the definition of the purpose of the ESAs, which was simply to “protect the consumers’ interest”. As a result, Kask considered that the Boards translated the consumer interest purpose to “fairness in pricing”, for which they each had their own criteria and own responsiveness to consumer needs. This, in her view, was what had led to the industry’s pricing problems.

³⁰ This view failed to point out that, although non-residential customers in New Zealand paid significantly higher tariffs than residential customers *on average*, the lowest retail prices typically went to the largest bulk supply industrial customers (Jackson, 1990, pp. 18-21). For instance, concerns were raised that New Zealand’s largest electricity consumer, the Comalco aluminium smelter was being cross subsidised by the ESAs and their customers (e.g., Bertram, 1991a).

industry primarily to be problems relating to ownership.³¹ The introduction of competition into electricity supply was also seen as presenting a solution to the perceived inefficiencies. A typical argument was that a competitive environment places pressure on firms to provide better customer service, and to maximise internal efficiency through cost reduction, subsequently resulting in lower average prices (Saha, 1991). Defenders of the previous industry structure were few, although Kelsey (1993, p. 55) is one who claims that the pre-reform power system did provide an “efficient and cost effective service”.

To sum up, the power sector’s perceived problems fell neatly into the three components of economic efficiency described above (§2.1.2). Issues with productive efficiency stemmed from a lack of commercial pressures and incentives on all actors in the industry, as well as from a loss of scale economies due to the excessive number of distribution franchise areas. Allocative inefficiencies arose from politically-motivated and “unfair” cross-subsidies in power prices and from the lack of competition. Finally, dynamic inefficiencies were evidenced by significant over-investment in generation, transmission and distribution capacity, as well as in levels of security and quality. This was again considered to be caused by a lack of commercial pressures and competition.

2.3 Policy Framework, Power Sector Reform Objectives and Desired Outcomes

2.3.1 Energy Policy Framework, Power Sector Policy Objectives and Desired Outcomes (1992-1998)

New Zealand’s power sector reforms are best viewed in the context of successive policy frameworks issued by the Government relating to both the power sector and to the energy sector as a whole. However, the initial reforms to the electricity supply industry, which began in earnest under a Labour Government in 1987 (§2.4.1), were undertaken without the guidance of any sector-specific policy framework. Since the State’s role in the energy sector was to be diminished, and the sector to be eventually placed on a more commercial footing, the Government abolished the Ministry of Energy in late 1989, and its policy and regulatory functions were subsumed into the Ministry of Commerce (although a Minister of Energy was still retained). Some industry commentators thought that the abolition of the old Ministry left something of a policy vacuum, and it took some time after the initial stage of reform before a comprehensive policy framework for the energy sector was announced.³²

³¹ “New Zealand has had decades of experience with government ownership of distribution companies and the case for change is rightly built in dissatisfaction with outcomes under those structures” (NZBR, 1999, p. 5). This viewpoint from NZBR is still held strongly, particularly given that power generation still is predominantly performed by State-owned enterprises, albeit “competing” ones.

³² Broader discussions of specific policy actions involved in New Zealand’s wider energy sector reforms can be found in Culy and Gale (1987) and Cocklin (1993).

When an Energy Policy Framework was finally issued by a National Government in June 1992, it perhaps arrived a bit quietly, with the Minister of Energy finding the need to write to the General Manager of the largest ESA, as late as December of that year, in order to affirm the Policy Framework's existence and to send him a copy.³³ The overall energy policy objective announced as part of this Framework was as follows.

To ensure the *continuing availability* of energy services, at the *lowest cost to the economy as a whole*, consistent with *sustainable development* (Luxton, 1992a) [emphasis added].

The three highlighted phrases of this objective provide three sub-objectives which neatly group together the outcomes that the Government stated it sought to achieve within the Energy Policy Framework: (a) "*continuing availability*", which relates to outcomes intended to ensure the short and long run *security* of energy supply; (b) "*lowest cost to the economy as a whole*", which relates to outcomes intended to achieve *economic efficiency*; and (c) "*sustainable development*", which relates to outcomes intended to improve *energy efficiency* and to mitigate any adverse *environmental impacts* associated with energy production and use. In addition, the term "*energy services*" acknowledges that consumers of energy (and electricity) are not interested in the consumption of energy *per se*, but in the end-uses that energy use provides (such as heating, lighting and watching television).³⁴ Specific outcomes that the Government cited as seeking to achieve from its energy policy, consistent with the economic efficiency aspect of the overall objective, included: "the efficient and effective provision of energy services through well-functioning commercial systems with competitive incentives, operating within an effective and stable regulatory environment"; and "reduced statutory and structural barriers to enterprise and innovation in the supply and use of energy services".

Although policy initiatives and desired outcomes specifically relating to the electricity supply industry were outlined as part of the Government's Energy Policy Framework, there was no policy *objective* in the Framework associated with the power sector *per se*. Nevertheless, the same Minister of Energy who issued the 1992 Framework, had earlier declared that his vision for the reformed power sector was one of a "deregulated, efficient, and competitive service industry, with minimal Government involvement". The main objective of the power sector reform was as presented in the first quote at the beginning of this Chapter, and is repeated here.

³³ "I believe that many are not fully aware of the Government's policies on energy. Certainly, a number of commentators have continued, quite erroneously, to suggest that the Government has no energy policy. The Government has a comprehensive and coherent policy framework" (Luxton, 1992b).

³⁴ This taxonomy for the phrases in the energy policy objective and for the criticisms of the power sector reforms presented at the end of this Chapter (§2.5) have been based on earlier work by the present author (i.e., Gunn, 1995a; Gunn, 1997).

Any restructuring of the vital electricity sector must be carefully planned and implemented. It should not lose sight of *the main objective, to improve efficiency*. ... There are three key aspects of efficiency forms that electric power companies will need to achieve in economists' terms: *productive efficiency; dynamic efficiency; and allocative efficiency* (Luxton, 1991a) [emphasis added].

This objective for the power sector thus appears to place the emphasis firmly on the “lowest cost to the economy as a whole” strand of the overall energy sector policy objective. Key policy initiatives outlined in the Framework which were directed at the power sector included: the removal of statutory barriers to competition in electricity retailing; the separation of natural monopolies from competitive activities to address issues of market power; and the requirement that wholesale electricity prices be determined by negotiation between electricity generators, wholesale buyers and major consumers. These actions were broadly in line with emerging international power industry best practice (§2.2.2), namely, the deregulation of the industry’s potentially-competitive subsectors (i.e., generation, wholesaling and retailing), and the imposition of pro-competitive regulation on those subsectors considered to have natural monopoly characteristics (i.e., transmission and distribution).

In addition, the Energy Policy Framework announced the Government’s intention to corporatise publicly-owned electricity distribution and retail businesses in order to improve efficiency incentives. Overall, the preferred regulatory regime for the energy sector was indicated as being a light-handed one based on comprehensive information disclosure and the monitoring of performance and outcomes. However, the Government specifically provided for the possible future control of electricity charges to residential consumers, should they be deemed necessary, and also raised the issue of “fairness” in the context of general business conduct.³⁵

Apart from the reform outcomes enshrined in the Government’s Energy Policy Framework, two other outcomes were less formally envisaged with respect to the power sector reform process; namely: “lower prices”—as can be seen from the quote opening Chapter I—and “consumer choice”. Successive Labour and National governments have claimed that a key aim of any reform was “lower” or “fairer” electricity prices (e.g., Butcher, 1990a; Peters *et al.*, 1998a; Hodgson, 2000a), as this was thought to lead to lower costs for businesses, further leading to enhanced international and local competitiveness, resulting in jobs and economic growth. In 1990, the Minister of Commerce and Energy went as far as claiming that: “Savings in electricity bills represent an immediate improvement in costs and living

³⁵ The Government’s wider objectives for the economy and for markets dominated by natural monopolies were expressed in a statement of the Government’s Strategic Result Areas for the Public Sector (1994-1997): “[The] establishment, implementation and monitoring of legislative frameworks for the fair and efficient conduct of business and the operation of markets, which rewards innovation, promotes efficiency and enhances investor confidence” (quoted in Ministry of Commerce and The Treasury, 1995, para. 2).

standards for everyone and will help to achieve the Government's target of full employment by 1995" (Butcher, 1990b). By "lower prices", what was generally meant was lower *average* prices (e.g., Saha, 1991), as the desire to remove cross subsidies to improve allocative efficiency clearly implied that residential and rural area retail prices would have to *increase*, irrespective of any improvements in productive and/or dynamic efficiency.³⁶

The reforms have also been consistently touted as augmenting consumer choice (e.g., Peters *et al.*, 1998a), and were seen as likely to result in electricity distributors, and subsequently retailers, restructuring themselves under the auspices of the "marketing concept" (e.g., McLay, 1993). This led the Minister of Energy who presented the 1992 Policy Framework to predict that: "The future is likely to see reduced distribution costs for electricity and real choice for consumers, to the extent of looking up the Yellow Pages for competitive electricity quotes" (Luxton, 1991b).

2.3.2 Policy Statement on Market Power in the Electricity Sector (1998)

The Energy Policy Framework remained the Government's key policy statement pertaining to the power sector until late 1999. But in December 1998, as part of its Electricity Reform Package (§2.4.5), the Government reaffirmed the 1992 energy policy objective, and also issued a formal Statement of Economic Policy on Market Power in the Electricity Sector.³⁷

This Statement specified goals specific to the power sector; ones required to enable the achievement of the wider energy policy objective. These goals were that: (a) "electricity prices should signal the full cost of providing each extra unit of electricity"—a statement implying the importance of marginal cost pricing (§4.1.1); and (b) "electricity costs and prices should be subject to strong and sustained downward pressure"—the "lower prices" objective, not necessarily consistent with the previous goal. The Government considered that these goals would be best achieved by: (i) "strong competition in all sectors of the electricity industry where competition is possible (electricity generation and retailing)"; and (ii) "robust regulation which replicates competitive pressures in the natural monopoly sectors of the electricity industry (electricity distribution and transmission)"; (New Zealand Government, 1998). This

³⁶ For instance, the Ministry of Commerce (Rural Supply Working Party, 1990, p. 3) estimated that the removal of the cross subsidies for rural supply, and the requirement to make a reasonable rate of return, would on average imply an increase of 78% in the electricity retail prices to rural consumers. Removal of cross subsidies, or "tariff rebalancing" as it was termed, was also a key theme of an initial Government discussion paper on the appropriate philosophy for setting distribution line charges (Ministry of Commerce, 1991, pp. 79-80).

³⁷ The *Commerce Act 1986* (s26) allowed for the Government to issue formal statements of economic policy for the purpose of guiding the Commerce Commission in its execution of the Act. Statements specific to electricity transmission and to the development of a wholesale electricity market had earlier been issued, in December 1994 (superseding an earlier Statement of October 1993) and in December 1995, respectively.

latter part of the statement was an explicit articulation of the rationale that had already been driving the Government's policy actions over the previous few years.

2.3.3 Policy Statement on the Further Development of the Electricity Industry (2000)

Not long after returning to the Treasury benches as a result of the 1999 election, the new Labour Government presented its own package of power sector reforms: the "Power Package". The core of this Power Package was a new overall Energy Policy Framework relating to energy efficiency and renewables, climate change, and to the electricity, natural gas and transport sectors. This was soon followed by a formal Policy Statement on the Further Development of New Zealand's Electricity Industry (New Zealand Government, 2000), and this revoked all the economic policy statements that past National governments had issued relating to the sector. The new Energy Policy Framework, and the more specific Policy Statement on the electricity industry, shared parallel objectives as follows.

To ensure the delivery of energy services, and electricity, to all classes of consumer in an efficient, fair, reliable, and environmentally sustainable manner (Ministry of Economic Development, 2000b).

To meet this objective in the electricity industry, the Government stated that it favoured "industry solutions where possible" but was "prepared to use regulatory solutions where necessary". Any industry solutions, or arrangements, were to "promote the satisfaction of consumers' electricity requirements in a manner which is least-cost to the economy as a whole and is consistent with sustainable development". Consequently, the same three sub-objectives of security, economic efficiency and environmental concerns, cited in the earlier 1992 energy policy objective, were effectively still in place. The evolution of more industry self-regulation now became a "guiding principle" (New Zealand Government, 2000). The only new concept explicitly added to the mix of objectives and principles was the notion of *fairness*, as is evidenced in the second quote opening this Chapter. The Minister of Energy even subtitled his new package of policy actions: "*A Fair Deal for Electricity Consumers*" (Hodgson, 2000a). While the concept of "fairness" has not been defined by the Government, the sentiment of giving people a "*fair go*" is firmly embedded in the New Zealand psyche, having formed part of the country's egalitarian heritage over the course of many years (e.g., Jefferies, 1997)—"Fair Go" itself being the title of a long-running New Zealand television programme championing the rights of consumers.

The Policy Statement outlines the key outcomes sought in the power sector, and requires that: (a) "the full costs of producing and transporting each additional unit of electricity are signalled so that investors and consumers can make decisions consistent with obtaining the most value from electricity"—

which could be interpreted as a statement of *marginal cost pricing*,³⁸ (b) “delivered electricity costs and prices are subject to sustained downward pressure”—the *lower prices* objective; (c) “energy and other resources are used efficiently”—which could be read as a restatement of the *economic efficiency* objective; and (d) “the quality of electricity services, and in particular trade-offs between quality and price, should as far as possible reflect customers’ preferences”—the *consumer choice* objective. (The first two objectives are derived from the goals identified by the previous National Government in its own Policy Statements). In electricity distribution, these desired outcomes were described as being best achieved by “costs minimisation, so that transmission and distribution companies seek to minimise costs while providing the level of services demanded by customers” (New Zealand Government, 2000). This wording was very similar to the specific goal for the distribution sector that, although it did not find its way into the Statement of Economic Policy, had been part of the National Government’s Electricity Reform Package (Energy Markets Policy Group, 1998b). Hence, the change in Government did not mark a significant change in objectives or desired outcomes, although it placed renewed emphasis on the concept of “fairness”.

2.4 Power Sector Reform Actions (1987-2001)

2.4.1 Initial Reforms under the 1984-1990 Labour Governments

Notwithstanding the scope of perceived problems in the electricity supply industry (§2.2.3), the initial stage of reform was fairly limited in scope, focusing mainly on generation and transmission. Moreover, initial policy actions can be seen as part of the Government’s broader program of corporatising state trading activities under the *State Owned Enterprises Act 1986*, rather than as specific power sector reforms undertaken within an integrated policy framework for the energy sector. The *State Owned Enterprises Act* was part of measures taken by the Government to improve the performance and accountability of the public sector, since the Act required State Owned Enterprises (SOEs) to operate with commercial structures and incentives, subject to a principal objective of being a successful business.

In April 1987, the NZED was incorporated under the Act as a tax-paying SOE, and the ESAs were also required to pay income tax. The new SOE, the Electricity Corporation of New Zealand Limited (ECNZ or “Electricorp”), was required to operate as a profitable and successful business, and to permit the connection of private generators to its transmission network. However, a number of commentators, including its first CEO (Deane, 1989), considered that privatisation would impose more

³⁸ The absence of the word “marginal” is similar to the nod given to marginal cost pricing in the 1978 Public Utility Regulatory Policy Act (PURPA) of the United States. Berg and Tschirhart (1995) cite PURPA as promoting six pricing principles, including the “cost-or-service standard” that: “Methods should take into account the extent to which total costs are likely to change if *additional* kWhs or electric energy are delivered to electric customers” [emphasis added]. They note that this principle was generally acknowledged at the time (e.g., Joskow, 1979) as having the intention of encouraging marginal cost pricing.

stringent market pressures on electricity sector participants than could corporatisation, and thus privatisation was regarded by many in the business community as the logical next step for all parts of the supply industry (with the possible exception of transmission).

Further reforms in generation and transmission, and the extension of the reform program to the distribution sector, were planned during the Labour Government's second successive term. An Electricity Task Force was established in 1988, and was required to review the structure and regulatory environment of the industry, subject to an overriding objective of economic efficiency. Its key recommendations (e.g., Cabinet Policy Committee, 1989, p. 11) concerning each part of the supply industry (i.e., generation, wholesaling, transmission, distribution, and retailing) were that: (i) barriers to entry in the generation sector should be removed (to promote wholesale competition); (ii) control of the transmission grid should be separated from generation (to allow open access to competitors); (iii) line and energy charges should be separated, with line charges being only subject to light-handed regulation (to minimise cross-subsidisation); (iv) retail franchise areas for supply should be removed, and there should be no regulation of energy prices (to promote retail competition); and (v) like NZED, ESAs should be corporatised (to expose them to commercial pressures).

However, the Labour Government only had the opportunity to partially implement one of these proposals: namely, the separation out of ECNZ's transmission assets and operations into a wholly-owned subsidiary, Transpower New Zealand Ltd.—'Transpower'—in April 1988. The Government planned next to tackle the corporatisation of the ESAs, but the only reform made in the distribution sector was through the *Electric Power Boards Amendment Act 1990*, under which commercial directors were appointed by the Government to EPB boards. Any previous board members not appointed as directors became trustees, and were to hold the shares upon corporatisation of the EPBs (Energy Markets Policy Group, 2001a, p. 4). Shortly after implementing this measure, the Labour Party lost the 1990 election, and the baton of power sector reform was passed to a National Party Government.

2.4.2 Introduction of Contestability in Distribution through Light-Handed Regulation

The Energy Policy Framework of 1992 laid the groundwork for some substantial reforms relating to sector ownership and structure, particularly in what had been the traditional activity of electricity distribution. The National Party Government took up where the Labour Party had left off, enacting the *Energy Companies Act 1992* in July of that year, which provided for the corporatisation of the ESAs. The ESAs now became Energy Companies (ECs), and the Act required that the "principal objective of an energy company" be "to operate as a successful business". ECs were given a set timeframe to prepare establishment plans detailing how their shares were to be allocated. A diverse range of ownership patterns resulted, although the consumer-elected trust ownership structure, resulting in annual rebates or dividends to consumers, was initially the most favoured (Energy Markets Policy Group, 2001a, p. 5). Other ECs gave away their shares directly to consumers, which were rapidly snapped up by foreign (as

well as local) energy companies, keen to secure a stake in the soon-to-be deregulated energy market. Nevertheless, in some of the old MEDs, consumers voted for shares to remain in local government hands.

These changes in the ownership of the entities responsible for electricity distribution and retailing were soon followed by structural reforms, as well as the transformation of commercial relations between ECs distributors and consumers. In April 1993, the *Electricity Act 1992* came into effect. This Act partially removed EC franchise boundaries and thus the exclusive right to supply that had been inherited from the old ESAs. It also removed the prior obligation-to-supply with respect to any new consumer-requested connections within the old franchise area, although existing lines were required to be maintained until 2013, unless all consumers associated with a particular line requested disconnection. In response to the Act, and recognising the rationale behind the reforms, the ESAs had already developed a code of practice for “contestable energy trading” (ESANZ, 1993).

Initially, franchise restrictions were removed for only the smallest consumers, those consuming under 500 MWh per annum (typically residential consumers and small businesses). The stated reason was to avoid the possibility that the smaller consumers might later end up cross subsidising the larger consumers, since the level of competition for various consumers was expected to increase in proportion to their level of consumption (Energy Markets Policy Group, 2001a, p. 7). The partial removal of the franchise meant that retail competition was now possible, at least in theory. However, much more prevalent than any signs of burgeoning competition, were rapid changes in ownership for those ECs with some private shareholding. The passage of the new Act resulted in a flurry of mergers, acquisitions and highly complex cross-ownership arrangements (e.g., Small, 1995). Although trusts were able to purchase shares in other companies, they were themselves protected from takeover. As had been earlier advocated (§2.2.3), this led to a process of industry rationalisation, and the number of ECs began to reduce appreciably.³⁹

The franchise boundaries were removed entirely one year later. All consumers were now, in the view of the Government, “contestable” (Energy Markets Policy Group, 2001a, p. 8). The mechanism of this contestability was to be achieved through a new light handed regulatory regime based on information disclosure. The promulgation of related regulations was explicitly allowed for in the *Electricity Act 1992*, and these were issued as the *Electricity (Information Disclosure) Regulations 1994*, focusing primarily on the natural monopoly aspects of an EC’s business. These Regulations viewed an EC as two distinct and “transparent” businesses: an ‘energy retailer’ (or ‘energy trader’), the potentially-competitive business unit responsible for reselling to consumers the energy which it purchased from ECNZ; and a ‘line business’, the owner and operator of the naturally-monopolistic distribution network assets,

³⁹ For instance, toward the end of 1995, 41 of the original 60 ECs remained, 19 of which were run by trusts, 8 by local government, 7 by private interests, and 7 by a mixed or other form of ownership structure.

responsible for the connection of consumers, via their own system, back to Transpower's bulk transmission network.

Retail competition became possible because all line businesses were required to set and disclose 'line charges' for connection to, and for the conveyance of power (or 'wheeling') across, their distribution networks. The same charges were to apply whether it was the traditional *incumbent* energy retailer selling energy to the existing consumers, or an *external* energy retailer owned by a competing EC.⁴⁰ The external energy retailer would enter into a 'use-of-system agreement' with the incumbent line business, to compete with the incumbent retailer on energy sales and service. Regardless of whether the energy retailer was internal or external to the EC, they were required to face the same line charge structure, but in setting the price that their end-use consumers would pay, they were able to 'rebundle' the line charges, the wholesale electricity purchasing costs from ECNZ, the transmission charges from Transpower, along with the desired margin, in any manner. Typically, most consumers were presented with two or three price components: variable charges, dependent on consumption (in kWh) and/or on peak demand (in kW); as well as a fixed charge, usually related to the electrical connection capacity (i.e., kVA) of a consumer's site (§3.1.1), and combined with the consumer's allocation of ELB overhead costs. Since distributors were now no longer statutory monopolies, not only could they potentially compete through their energy retailing business to supply energy services, but line businesses were also free to compete for network connection. ECs could now construct new subnetworks embedded within the old franchise boundaries of a neighbouring or any other EC (§3.3.2).

2.4.3 Role of Valuation Benchmarks in the Light-Handed Regulatory Regime

Apart from providing a transparent mechanism for allowing retail competition, the light-handed regulatory regime was intended to provide benchmarked comparisons of line businesses, in an attempt to expose any strongly monopolistic behaviour. The regime, as outlined in the *Electricity (Information Disclosure) Regulations 1994*, required that certain information relating solely to any line business (including Transpower, since it is also a network operator) must be publicly disclosed. Disclosure information, subject to independent audit, included a range of financial and technical performance indicators, line charge schedules, price setting methodologies, as well as the terms and conditions of line business contractual arrangements. By contrast, the regime required little in the way of information disclosure from energy retailers or generators (and all such requirements were removed with the new information disclosure regulations promulgated in 1999; §2.4.5). The general provisions against anti-

⁴⁰ An incumbent energy retailer was one that had effectively been the business unit of an old ESA, associated with a specific geographical franchise, that had been responsible for all electricity sales in that area. An external energy retailer was one that had been previously associated with an ESA covering a different geographical franchise (whether this be a neighbouring area or some region in a different part of the country), or a *new* company set up for the *sole* purpose of energy retailing.

competitive behaviour enshrined in the *Commerce Act 1986* (§2.1.3) were seen as providing sufficient safeguards against any abuse of retailer market power.

The key performance statistic that the Ministry of Commerce was to keep a watch over was the accounting rate of profit (ARP)—later termed the return on investment (ROI)—for each line business. Where the ARP was significantly in excess of the post-tax weighted average cost of capital (WACC) faced by a line business, then that business could be considered to be abusing its monopoly power, and potentially be subject to heavy-handed regulation in the form of price control provided for under the *Electricity Act 1992* (§2.1.3). However, unlike most international rate of return (ROR) based regulation (§6.4.1), there was reasonable latitude in the behaviour of any line business, since successive New Zealand governments were perceived by the industry as loathe to engage in overtly interventionist actions.⁴¹

The ARP is effectively the ROR on the valuation of the distribution network owned and operated by the line business. To try and get comparable and benchmarked ARP valuations, the Ministry required all ELBs to use the same methodology, the Optimised Deprival Valuation (ODV) method. Historical book valuation methods were considered unreliable as there had been little common ground between the methodologies and assumptions that were applied by the various ESAs prior to distribution sector reform. Thus the main purpose of the New Zealand Government’s requirement for all line businesses to disclose the value of their assets using a common methodology was to provide an implicit restriction on monopoly pricing behaviour consistent with a contestable market outcome. The requirement has primarily served the *regulatory* objective of determining a valuation base that can subsequently be used to derive benchmarked comparisons of the rates of return for all ELBs.

The Government has not required that the ODV be used as the basis for determining accounting ‘book’ value, and consequently for tax purposes. Neither has it been required that line business prices be set indexed to their ODV. However, the Ministry of Commerce (Energy Policy Group, 1994e, pp. 8-9) noted that it was acceptable for the ODV to be used in audited financial statements, and that certified ODVs satisfy the accounting criterion of reliability and relevance more satisfactorily than historic cost, as well as giving a more “true and fair view” of value. Proponents of the ODV method assert that the use of ODV for pricing leads to prices that are a good proxy for those that would eventuate in a competitive market where supply and demand are balanced (TPEB, 1991, p. 28). Tariffs reflecting ODV are seen as sending the “correct” signals to consumers, thereby ensuring they value their consumption of electricity

⁴¹ Or so the line companies themselves have at times believed. In a 1997 survey of the CEOs of New Zealand line businesses, 29% of respondents believed the Ministry of Commerce would allow up to an additional 4% return above the WACC, and 71% believed that the Ministry would allow a 0 to 3% leeway. However, 44% of respondents did not even expect the Ministry to act on *any* result until after the year 2000.

distribution services appropriately (Saha, 1993). As a result, ODVs frequently became the *de facto* basis for setting line charges. The significance of this outcome cannot be underrated—as Bertram and Terry (2000, p. 4) express it: “asset valuation is *the* issue if one is examining lines company pricing in New Zealand”. Given its significance, the ODV methodology is examined in greater depth later (§7.3-§7.4).

2.4.4 Establishment of a Competitive Wholesale Electricity Market

By early 1995, although retail competition was now possible, the Government conceded that: “clearly, it is going to be some time before competition is a reality rather than a promise for other than large electricity consumers” (Kidd, 1995). Even supporters of the reform program had earlier warned that, without a competitive *wholesale* market for electricity, *retail* competition would be rendered ineffective (e.g., Cole, 1993; Heffernan, 1993; Lough, 1994; Wilson, 1994). This is because, while generation still accounted for around 50% of the retail electricity price, retailing costs only contributed to about 3% of the average consumer's bill. This left little room for energy retailers to compete on the basis of price.

Wholesale competition was not viable because ECNZ was still in control of around 95% of all electricity generation. (The remainder was contributed by small old power plants owned and operated by the ECs themselves). Not only was transmission not truly independent of generation but one study of the corporatisation of ECNZ, performed with the co-operation of the Corporation, indicated that ECNZ had actively attempted to deter entry into the market through wholesale price reductions (Spicer *et al.*, 1991, pp. 28 and 81). One potential entrant to the generation market implied that ECNZ had also engaged in directly anti-competitive behaviour, using the planning approval and resource consent processes available under the *Resource Management Act 1991* (McLachlan, 1992). ECNZ's main justification for its behaviour was that it was always acting to maximise its net worth to its shareholder, the Government, with a view to maximising the eventual sale price upon privatisation (Spicer *et al.*, 1991, p. 24). Moreover, ECNZ was clearly unhappy with the Government's increasing desire to break the corporation up for the purpose of promoting competition in generation, on the basis that such a policy would also reduce the company's value (FERNYHOUGH, 1990).

Indeed, the failure of new generators to enter the generation market was of clear concern to both the National and earlier Labour Governments. Consequently, a number of studies had been commissioned to examine possible structures for a wholesale market, and possible mechanisms for breaking up ECNZ. As a first step to implementing the wholesale market, the recommendation of the Electricity Task Force to separate transmission from generation (§2.4.1) was finally implemented in July 1994. Transpower, initially just a wholly-owned subsidiary of ECNZ, now became incorporated as a separate SOE (Transpower New Zealand Ltd.). This separation was designed to ensure effective open access to the transmission network for new generators, as well as a distinct separation of the generation (potentially-competitive) and transmission (naturally-monopolistic) components of the wholesale

electricity price. In June 1995, the decision was made to split ECNZ into two competing SOEs as of February 1996, and to establish a competitive wholesale market, in October of that year. Contact Energy Ltd. was the new SOE spun off from ECNZ, comprising just under 30% of ECNZ's generating assets.

An industry market framework for wholesale electricity trading was already in place. The Electricity Supply Association of New Zealand (ESANZ), which represented all ECs, in tandem with ESANZ and Transpower, had set up the Electricity Market Company Ltd (EMCO) in 1993.⁴² A key aspect of EMCO's work was to administer the Metering and Reconciliation Information Agreement (MARIA), which detailed the procedures for recording and reconciling power flows to meet the needs of parties contracting in the wholesale and retail markets. Under MARIA, Transpower reconciled metered power flows against contracts, and passed information back to market participants for billing purposes. Once the wholesale market commenced operations in October 1996, EMCO took on the role of market administrator, clearing manager and pricing manager, and Transpower retained the role of generation dispatch. Wholesale electricity prices became based on bids and offers from market participants (i.e., generators, energy retailers and major consumers), although most electricity became bought and sold through long term hedge contracts rather than through this spot market (Energy Markets Policy Group, 2001a, pp. 8 and 11). In expectation of these reforms, construction of the first major private power plant had already begun, and this generator commenced operation in December 1996.

In September 1997, the Government revised Transpower's objectives to more strongly emphasise the need for it to continually improve the efficiency of transmission services, by "making the services contestable wherever possible" and "producing customer driven services at least cost" (Energy Markets Policy Group, 2001a, p. 11). The first privatisation of State electricity assets took place in March 1999, with the sale of a 40% cornerstone shareholding in Contact Energy to a US-based utility, and the remaining shares were floated two months later. Responding to criticisms that the Government had set up a dominant duopoly in place of a dominant monopoly generator (e.g., Electricity Week, 1995; O'Sullivan, 1995), the remaining assets of ECNZ were again split in April 1999, this time into three new SOEs: Genesis Power Ltd., Meridian Energy Ltd., and Mighty River Power Ltd.

⁴² This, along with the 1993 Code of Practice for contestable energy trading (§2.4.3), were indications that the electricity supply industry was prepared to take steps to set up its own governance framework, rather than to have to wait for Government-mandated solutions. Earlier, in 1992, private sector interests had performed their own Wholesale Market Electricity Study (WEMS, 1992), which can also be seen as the precursor to a number of industry-led reforms. It was primarily because the Government felt this privately-sponsored study warranted some carefully deliberated response that it began to accelerate moves to establish the wholesale market based on the recommendations of its own Wholesale Electricity Market Development Group (WEMDG), which reported back in 1994.

2.4.5 *Separation of Electricity Distribution from Retailing*

While preparing for the operation of the wholesale market, the Government had been monitoring the behaviour of the line businesses of ECs through the information disclosure statistics (e.g., Energy Markets Regulation Unit, 1998), as well as from consumer and external energy retailer complaints to the Commerce Commission (summarised in Bollard and Pickford, 1995). The possibility that line businesses could subsidise their incumbent energy retailers, or make the terms of access contracts difficult, thus raising entry barriers to external retailers, was of particular concern.⁴³ In early 1995, the Minister of Energy warned ECs that: “Cross-subsidising competitive consumers by captive customers, cross-subsidising potentially competitive energy supply from monopoly line businesses, unnecessarily restrictive provisions in line service agreements, and ensuring generation profits off the back of captive consumers, are all examples of behaviour which once identified will not be tolerated” (Kidd, 1995). The Government acted on this threat three years later, once it had become clear that competition was only a reality for either the very largest consumers, or for consumers with nationwide accounts that were able to benefit from consolidated billing (such as supermarkets and other nationwide businesses or retail chains).

In releasing a new Electricity Reform Package in April 1998, entitled “A Better Deal for Electricity Consumers”, to address this problem, the Government announced that: “The current level of competition in retailing is very disappointing, and has been shrinking rather than growing. At present only three per cent of all electricity generated is traded across regional monopoly boundaries. ... As a result of these reforms, average electricity prices are expected to fall over the next few years” (Peters *et al.*, 1998a). The new Package focused on promoting greater competition in the interests of lower prices, through the split of ECNZ into three companies (§2.4.4), as well as through the complete ownership separation of energy retailers from line businesses. To some (e.g., Small, 1998) this increase in competition was seen as being at the expense of efficiency, should economies of scope (§3.4.2) exist between the functions of energy retailing and network services.

The separation was likened to the separation of Transpower from ECNZ, and was seen as removing incentives for ECs to frustrate retail competition (Peters *et al.*, 1998b). An EC’s naturally-monopolistic line business would be separated from its “contestable” activities, (i.e., retailing and any generation), allowing the network business to be exposed to closer scrutiny, particularly with respect to the identification of any monopoly pricing behaviour. Further, the separation was not only designed to recognise the underlying differences in cost structure between lines businesses and energy retailers, but also the differences in business risks and business priorities. The focus for the line business was seen as

⁴³ Some limited research was undertaken that suggested such cross-subsidisation could be occurring (e.g., Gilbert, 1997).

being efficient asset management, security of supply and cost minimisation, whereas the focus for retailers was regarded as managing entrepreneurial risk (Energy Markets Policy Group, 1998a).⁴⁴

To implement the separation, the *Electricity Reform Act 1998* was passed in July of that year. Subsequently, the information disclosure regime was strengthened, made more transparent, as well as revised to be consistent with the structural changes implemented under the Act, through the issuance of the *Electricity (Information Disclosure) Regulations 1999*. The Act gave ECs eight months to achieve corporate separation of their lines and energy businesses, and full ownership separation was mandated by the end of 2003. However, ECs chose to move quickly, and full separation was completed by the date for corporate separation. As a result, by April 1999, the 32 remaining Electricity Line Businesses (ELBs) were left to concentrate solely on the business of electricity conveyance and distribution network connection. The spun-off retailers amalgamated into just seven companies, many owned by generating companies, including the three post-ECNZ State-owned generators.⁴⁵ To enable consumers to switch retailers easily and at low cost, as was required under the Act, the industry established The Marketplace Company Ltd (M-co) to administer the required electricity profiling system (e.g., NZIER, 1997). The Government had indicated that it was prepared to regulate for the introduction of such a system should the industry not find its own solution (Energy Markets Policy Group, 2001a, pp. 12-13).

2.4.6 *Proposals for Stronger Regulation of Distribution Natural Monopolies*

The Electricity Reform Package also signalled the intention of the Government to enhance the credibility of the threat of actual price control on the post-split ELBs. Since ELBs were still deemed natural monopolies (Peters *et al.*, 1998b), the risk of excessive prices still remained, and the information disclosure regime—in combination with the *Commerce Act 1986*—were now regarded by the Government as insufficient to deal with access, cross subsidy and monopoly rent problems (Energy Markets Policy Group, 1998a). As such, an effective price control threat was considered necessary to supplement the existing regulatory regime. The Government proposed empowering the Commerce Commission to impose price control should ELBs breach Government-set performance criteria termed “thresholds”. In preparation for this move, the Government issued a formal policy statement on market

⁴⁴ The perceived need for the focus on asset management and security of supply arose as a result of the major blackouts caused by multiple cable failure that had been experienced in Auckland, New Zealand’s largest city, in early 1998. The subsequent Ministerial Inquiry into the Auckland power failure recommended that the information disclosure regime be extended by including a requirement for line businesses to disclose asset management plans and security standards (Ministry of Commerce, 1998). These recommendations were implemented through the *Electricity (Information Disclosure) Regulations 1999*.

⁴⁵ The Government expressed some reservations about this re-aggregation of generators and retailers into vertically-integrated structures: “while likely providing benefits in terms of efficient management of risk and minimising transaction costs, [this] may affect the quality of price signals in the marketplace” (New Zealand Government, 1998).

power in the electricity sector (§2.3.2), as well as a discussion document outlining the proposed threshold criteria, in late 1998 (Energy Markets Policy Group, 1998c).

The new proposal acknowledged that, in most western countries, some form of price control (rate of return, CPI-X, or sliding scale; §6.4.1) was used to address the monopoly pricing problem for utilities. It was recognised that New Zealand's regime was implicitly a form of rate of return regulation, although implemented in a uniquely light handed manner (§7.3). The "specific thresholds" proposal for price control, a regime of yardstick regulation (or 'yardstick competition'; §2.1.1), was a substantial departure from the previous approach in two respects. Firstly, the key indicator of an abuse of monopoly power was no longer solely the ARP based on an ELB's optimised deprivation valuation (ODV; §2.4.3), but a combination of seven indicators relating to service quality, averaged costs and averaged revenues (i.e., prices). Only one of these indicators was based on the ODV. Secondly, ELBs would be ranked according to the indicators, and would be assigned a weighted average of demerit points associated with their rankings every six months. Only ELBs receiving a consistently high score (above a certain threshold) would become subject to direct price control by the Commerce Commission. Because demerit points were based on an ELB's ranking, rather than on the underlying values of the performance indicators, the scheme was seen as capable of providing continual pressures on ELBs to improve performance (Energy Markets Policy Group, 1998c). However, the proposal was viewed by many industry commentators as being too heavy handed (e.g., Evans, 1998), as well as focusing on too many potentially-conflicting objectives (e.g., Small, 1999a). In particular, submissions and independent advice on the scheme suggested that, based on the currently available data, it was not possible to adequately take into account factors outside the control of the ELBs. In response, the Government shelved the proposal (Energy Markets Policy Group (1999a)).⁴⁶

Instead, in May 1999, the Government decided to delegate the problem to the Commerce Commission to solve, and as such tabled the Commerce (Controlled Goods and Services) Amendment Bill before Parliament (in order to amend the *Commerce Act 1986*). Although the Government declared that its industry reforms were already delivering benefits to most consumers in terms of falls in both wholesale and retail electricity prices, it expressed concern that many ELBs had: "used the reforms as an excuse to increase their profit, while others are simply increasing their [line charges] because they think

⁴⁶ Such a justification for dropping the scheme was consistent with research that had been performed on the possibility of introducing a similar type of scheme for the Regional Electricity Companies responsible for power distribution in the UK. Weyman-Jones (1995) suggested that: "yardstick comparisons need measures which are independent of, or control for, the influence of exogenous variables that may have a major role in determining a utility's measured performance on a variety of indicators". He concluded that: "making the textbook model of yardstick competition operational requires the specification of relatively sophisticated models and estimation methods. This, however, requires a degree of regulatory oversight and intervention which is at odds with the original intention to have 'regulation with a light hand'".

they can get away with it". Appearing to consider that this situation was inherently "unfair", the Minister of Energy stated that, if these problems were not addressed, "consumers would continue to be fleeced" (Bradford, 1999). The Bill, had it been enacted, would have required the Commerce Commission to develop and administer incentive and/or revenue cap based price controls for all ELBs. The National-led coalition Government was unable to obtain the majority needed to enact the Bill, and in the election held soon afterward, a Labour-led coalition assumed the responsibility for power sector reform.

2.4.7 Reforms Post-1999

One of the first steps of the incoming Labour Government was to make good on its pre-election promise of commissioning an inquiry into the electricity supply industry as a whole. The Inquiry reported back in June 2000, and recommended further evolution of the self-regulatory arrangements in the electricity industry, as well as that the threat of price control for ELBs be made credible (Ministry of Economic Development, 2000a), possibly through introducing some form of price cap regulation (§6.4.1). In response to the Inquiry's findings, the Government presented its Power Package in October 2000, the core of which was a new overall Energy Policy Framework (§2.3.3). The Minister of Energy stated that the Package was designed to "sort out the mess left by the previous Government and give consumers the deal they deserve" (Hodgson, 2000a).

For the distribution sector the Power Package included new rules for the valuation of ELBs, in the form of a new and more stringent edition of the ODV Handbook (Energy Markets Regulation Unit, 2000c), and an indication that the Commerce Commission's role in the distribution sector was to be expanded. The Commission was to be charged with the responsibility of overseeing the information disclosure regime, with reviewing whether the ODV was the best valuation methodology after all, and mandating any changes (Hodgson, 2000b). The Package also set up a specific timetables for the industry to develop its own Electricity Governance Board. In keeping with the rationale underlying past policy actions, one of the requirements was that the Electricity Governance Board should "ensure that the provision of [electricity] services is contestable wherever possible" and develop industry rules to "promote enhanced competition wherever possible". The Board would also be required to develop "model" use-of-system agreements and approaches to distribution pricing (New Zealand Government, 2000).

An Electricity Industry Bill was tabled in Parliament in November 2000, to provide the Government with the power to introduce regulations in areas where the Board fails to deliver. Should the Governance Board itself entirely fail to deliver industry rules that meet the Government's expectations, then the Bill also enables the Government to replace the industry's Board with a Crown entity. This possible return to more intervention in the power sector led one member of the opposition, himself an ex-Minister of Energy, to declare that the Bill was "a monster, a shocker which should be thrown out. ... A rogue elephant is being set loose" (Wilson, 2001). Nevertheless, the Government made it clear that it

“favours industry solutions ahead of regulation” (New Zealand Government, 2000), and successfully passed the Bill on 7 August 2001. (This by no means heralded the end of changes to the sector’s governance framework, but this dissertation does not consider the impact wrought by the passage of the Bill on the sector, or any further reform actions taken subsequently).

2.5 Broad Critiques of the Power Sector Reforms

2.5.1 *The Narrow Focus Critique*

New Zealand’s power sector reform program has not been without its share of detractors. While criticisms of the Government’s 1992 overall energy policy objective (§2.3.1) are hard to find, critics have been quick to pull holes in the individual merits of the various policy actions that have been taken in the power sector. Notwithstanding the Government’s position that “the reforms being introduced are part of a world-wide realisation that electricity is a tradeable commodity, rather than a community service” (Luxton, 1991b), there are still those who consider that electricity service provision has rather more unique characteristics. Although such a “uniqueness” view is held by some of the strongest critics of reform,⁴⁷ it is interesting that this view has also been shared by some supporters of reform, including the Chairman of New Zealand’s Wholesale Electricity Market Development Group, former Deputy Prime Minister, Jim McLay.

Nowhere in the world has an electricity system been allowed to develop on a totally (or even largely) *laissez faire* basis. Electricity is not just a commodity that can be traded as such. It is an essential service; with unique political, social, technical and commercial features (McLay, 1993).

Other industry commentators consider that the singular focus on economic efficiency as the primary or sole objective guiding the power sector reforms is inappropriate, and that the other goals enshrined in the overall energy policy objective warrant just as much attention. For instance, some see security of supply as too important a goal to be left to markets alone (e.g., Bertram, 1992; Leyland, 2000), whereas others consider that energy markets might be rife with market failures, particularly in relation to energy efficiency and environmental goals (e.g., Terry, 1991, pp. 27-30; Peterson *et al.*, 1992, p. 32; Harris *et al.*, 1993, pp. 11-25).⁴⁸ Consequently, these commentators prescribe a more substantial role for the Government in energy markets. Moreover, successive governments—from both Labour and

⁴⁷ For instance: “The debate and difference of views as to what role electricity should play in the development of New Zealand’s society and economy is not new. ... What is different in the current debate is the frequency and vehemence with which the concept of electricity as nothing more than a commodity to be bought and sold is expressed. Such an approach constitutes the narrowest of opinions expressed since the late-nineteenth century” (Jackson, 1990, p. 4).

⁴⁸ The author was also involved in work which suggested that the focus on “getting the price right”, and allowing only supply side investments in distributors’ valuation bases, would act as disincentives to energy efficiency and demand side management (DSM) initiatives (i.e., Tromop *et al.*, 1996). Similar views were presented in a study of energy efficiency in New Zealand performed by the International Energy Agency (IEA, 1999).

National parties—have even stood accused of embarking on a reform process based on (free market) ideology alone (e.g., Jackson, 1990; Bertram, 1991b; Noble, 1991; Russell, 1991a; Peet, 1992; Kelsey, 1993).⁴⁹ As Berg and Tschirhart (1995) observe, whether policy conclusions reached by particular individuals tend to be based more on analysis or ideology is an open question. Even doubts have been expressed from *within* the Government caucus, as the following statement concerning the power sector reforms from a past Minister of Conservation indicates.

There is a feeling in some quarters that the Government is building a fascinating new tin budgie but nobody is really sure whether the damn thing will fly or whether we are placing too much reliance on the sloping shoulders of the free market (Dennis Marshall, quoted in Kelsey, 1993, p. 55).

The typical response from the successive Ministers of Energy in charge of the reform program has been that its critics are themselves blinded by their own ideologies (e.g., Luxton, 1991a; Kidd, 1994).⁵⁰ Yet other industry commentators elicit elements of ideology on both sides of the sector reform debate (e.g., Cocklin, 1993; McLay, 1993).⁵¹

⁴⁹ Jackson (1990, p. 2): “Development at the bulk generation and supply level depend as much upon *political ideology* ... as upon issues related specifically to the electricity industry itself”.

Bertram (1991b): “The deregulation polices have been guided by a *social philosophy* which gives rather little credence to any conception of the ‘public interest’, and which is hostile to the proposition that unregulated markets fail to provide adequate incentives towards sustainability and energy efficiency”.

Noble (1991): “[There has been] inappropriate application of mock-competitive market rules to the transmission and distribution sectors of the industry. The remedy is to let competition take control where that is suitable, and let reality rather than *ideology* take control in the rest of the industry”.

Russell (1991a): “It seems to [the] Consumers’ Institute that the [Electricity] Task Force and some politicians have an *ideological preference* towards privatisation and this has led to a determination to find a way to achieve it”.

Peet (1992): “For a government ministry to promote free markets in energy under the guise of achieving economic efficiency, while at the same time ignoring the fact that, even in theory, free markets are incapable of achieving what is intended, represents economic ignorance. I suspect it reflects commitment to a *simplistic ideological belief in free markets* rather than broad-based economic rationality”.

Kelsey (1993, p. 55): “[Electricity supply authorities] were being replaced by an untried scheme driven by *ideological dogma*. The National government seemed determined to bring the rigours of the market-place to electricity” [*emphasis added in all quotes*].

⁵⁰ The initial reform requirement that ECs make a profit was the basis for criticism from those who viewed electricity provision as a public service. One Minister of Energy had this response to such criticism: “Profit should not be regarded as a dirty word in the electricity industry! ... Profit is the worldwide, accepted measure of business performance. It provides the return required to compensate for risk and uncertainty. It is the force behind the planning, investment and management decisions which are necessary for sustained high-quality low-cost service provision. Business profit in a service industry is not some New Right theory but common sense. McDonalds, Pizza Hut, the hotel industry, or the rest of the economy outside central or local government are driven by it. Profits are not only consistent with the best interests of consumers, they are essential if costs are to be minimised and efficiency maximised” Luxton (1991a). A later Minister of Energy, similarly responded to the

As with all political, economic and commercial issues, there are those at both ends of the spectrum who will advocate ideologically-based electricity reforms that have little regard for political, commercial and economic reality. In dealing with electricity sector reform, we must never allow ourselves to be blinded by ideology. With something as essential as electricity, there is no place for ‘interesting experiments’. Instead, we should seek simple and workable solutions to real and present problems (McLay, 1993).

2.5.2 *The Static versus Dynamic Efficiency Critique*

Critics also include those who agree that economic efficiency is the appropriate objective in restructuring the electricity supply industry, but disagree on the appropriate means required to achieve it, particularly with respect to the appropriate level of State intervention in the sector. The New Zealand Business Roundtable prescribes lesser State involvement and greater private sector participation in the industry, and echoes the past Minister of Energy John Luxton by stating that “the focus should be on the overall objective of economic efficiency, as measured by the sum of producer and consumer surplus” (NZBR, 1998). A common concern is that equal weight has not been given to the three components of efficiency (§2.1.4), as is seen in the third quote at the opening of this Chapter. This concern is shared in a general sense by some commentators on power sector reform world-wide.

In economics, the concept of efficiency is ambiguous and has different meanings under different circumstances. It can mean productive or allocative efficiency, technical efficiency, or even process efficiency, while in each case the presumption equates the analytical results with a social optimum (Bhattacharya, 1995).

The Government’s emphasis on productive efficiency gains has been subject to criticism, although such a focus is understandable given that such gains are more easily measurable than those relating to the other two efficiency components. For instance, while the managing director of the company which is New Zealand’s largest consumer of electricity agreed that corporatising ECNZ no doubt contributed to productive efficiency gains, he suggested that the increased profitability of this SOE should not necessarily be seen as guaranteeing beneficial effects for the economy as a whole (McDonald, 1991). Applying Government-recognised measures, productive efficiency had increased markedly by establishing ECNZ under the SOE model, and the reduction in wholesale prices of 16% between 1988 and 1993 was also presented as evidence for the early success of reform.⁵² However, one study of the

ideology criticism: “You [power companies] are now confounding your slogan wielding critics. Most of them know nothing of the power for good of competitive market forces because of the *political ideology* to which they subscribe. You are doing a better job than any government can do” (Kidd, 1994) [*emphasis added*].

⁵¹ Cocklin (1993): “There is no doubt that arguments in support of market led economies are couched more in *ideology* and theory than they are demonstrable in fact, but so too are the counterarguments”.

⁵² Increases in sales per electricity sector employee is another key indicator used as evidence of efficiency gains. For example, from 1988 to 1993, gigawatt-hours (GWh) per employee at ECNZ rose from 5.56 to 9.93. However, this does not account

corporatisation of ECNZ, performed with the co-operation of the corporation, indicated that the initial part of this price reduction at least can be explained as part of a strategy designed to deter competitors (Spicer *et al.*, 1991, pp. 28, 81). Moreover, as Lewis and Evans (1998) have pointed out, because excess generation capacity had existed prior to reform, falling real electricity prices do not provide a measure of social welfare gains in the sector over time.

Even more concerns have been expressed about a lack of attention to dynamic efficiency, in the sense of efficient investment, than with respect to issues of energy security.⁵³ Some industry commentators are sceptical about the ability of the market mechanism to send efficient and effective signals for investment in long-lived and capital intensive assets (e.g., Bertram, 1992). While likely to place much more faith in the invisible hand, the New Zealand Business Roundtable has nevertheless warned the Government that: “Dynamic efficiency is particularly important in capital intensive industries and must be explicitly considered. Measures that might primarily transfer wealth from producers or taxpayers to consumers in the short term, at the risk of adverse long-term effects, should be avoided” (NZBR, 1999, p. 5). The lack of attention given to dynamic efficiency is often considered as arising from placing too much emphasis on allocative efficiency, and especially to short term pricing impacts, hence the Government should focus on “innovation-favouring policies, rather than price-reducing policies” (Irwin, 2000, p. 7).⁵⁴ The subsequent Chapters of this thesis present the discussion of efficient and fair distribution line charges mostly within the context of this allocative versus dynamic efficiency critique. One of the architects of the valuation methodology for ELBs articulates the significance of the interest in this efficiency conflict as follows.

It is easier to focus on allocative rather than dynamic efficiency as a policy objective and consequently, greater weight is often placed on allocative efficiency and the ‘first round’ effects of policy. Yet it is important to remember that it is dynamic efficiency which determines our future welfare (Wilson, 2000a).

for the effects of “contracting out”. Other measures considered by the Government to be indicative of productive efficiency gains have included: reductions in unit costs (\$/kWh) and increases in sales volume (kWh); (McLay, 1993). Given that ECNZ was at the time practically the only supplier, this last indicator seems somewhat irrelevant.

⁵³ Even the International Energy Agency, in commenting on the early phases of New Zealand’s reforms, expressed concerns regarding the incentives for ensuring dynamic efficiency (IEA, 1989, p. 401). On the other hand, the IEA’s most recent review of New Zealand’s energy policy, focuses more on incentives for energy security (IEA, 2001b).

⁵⁴ Irwin (2000) compares allocative efficiency losses (due to inefficient electricity prices) versus productive efficiency losses (due to inefficient costs) over time, and suggests that the latter far exceed the former. This is based on the well-known static “triangles” versus “rectangles” surplus loss argument (e.g., Teplitz-Sembitzky, 1992, pp. 20-26).

CHAPTER III

ELECTRICITY DISTRIBUTORS IN NEW ZEALAND: NATURAL MONOPOLIES OR NATURAL COMPETITORS?

While much has been written on the electricity industry in general, little attention has been focused on electricity distribution. This is probably due to widespread acceptance of the belief that it is a natural monopoly: Report on Economies of Scale in the New Zealand Electricity Distribution System (Wyatt et al., 1989, p. 6)

The politics of the industry appear to have reached the stage where the assertion of natural monopoly suffices to justify additional regulation. ... The proposition that local lines do not present intractable natural monopoly problems should be taken seriously. Were it not for a century of state monopoly, markets would surely have eroded many natural monopoly attributes in electricity lines companies long ago: Executive Director of the New Zealand Business Roundtable (Kerr, 1999, p. 5)

The assumption underlying the proposal that lines businesses are natural monopolies has not been tested by officials, but has important implications for the design of policy to induce downward pressure on prices: Submission to the New Zealand Ministry of Commerce (Mercury Energy Lines Business, 1999, p. 3)

Upon the initial removal of franchise boundaries in April 1993 (§2.4.2), the energy companies responsible for electricity distribution and retailing in New Zealand were in fact observed to *compete* with each other, not just for retail services (i.e., ‘energy sales’), which was intended, but for ‘network services’ as well (i.e., electricity conveyance and network connection). But if network services are not “contestable” activities (§2.1.7), then how can competition be possible? As Broadman and Kalt (1989) ask: “how can competition exist in a market presumed to be a natural monopoly?” In light of this question, this Chapter starts by defining the characteristics of, and demand for, electricity distribution network services. It then outlines various international and local perspectives of the nature of electricity distribution, particularly the presupposition that distribution is a natural monopoly, and contrasts this view with the evidence for the existence of network competition in New Zealand. Studies of distribution cost data from New Zealand electricity line businesses (ELBs) are examined in the context of appropriate tests for determining the presence of natural monopoly. Finally, the possibility that ELBs might be monopolies for historical reasons—rather than as a result of their cost characteristics—is considered, by assessing the type and impact of barriers to entry and sunk costs in New Zealand’s electricity distribution market.

3.1 Distribution Network Services: Characteristics and Demand

3.1.1 Characteristics of an Electricity Distributor's Product: Connection Capacity

In assessing the nature of the business of electricity distribution, clearly it is important to define the product correctly. What exactly are distribution network services, and how are they measured? Traditional perspectives of electricity distribution have sometimes taken the simplest approach, viewing distributors as selling a *single* bundled product, namely energy sales (kWh). Sometimes, two products have been considered, energy sales (kWh) and power (kW), or different consumers have been grouped together as purchasing distinct products, such as residential, commercial and/or industrial energy supply. This view, however, fails to appreciate the service which an electricity line business actually provides. As discussed earlier (§2.2.1), the primary *function* of the unbundled ELBs in New Zealand is to convey electrical energy to consumers. However, this conveyance is not itself the ELB's *product*. In determining the characteristics of an ELB's product, firstly an examination is required of the characteristics of electric energy.

'Instantaneous electrical energy', or 'real power', is measured in Watts, or in thousands of Watts (kW). The standard unit of energy itself is the Joule (J), and one Watt is a Joule per second. However, in the electricity industry, useful energy is expressed in Watthours (Wh) rather than Joules, and usually in multiples of one thousand or one million (i.e., kWh and MWh respectively). Provision of real power over time, and thus useful electrical energy, is measured in kWh, which is found from the integral of real power over time.

However, ELBs do not produce power, they *convey* power. The amount of power that can be conveyed at any point of the network is limited by the 'capacity' of each of the 'upstream' distribution assets in that network. But not all power conveyed is useful or real power. The total or 'apparent power' conveyed comprises both real power (kW) and 'reactive power' measured in various levels of 'vars' (e.g., kVAR). Apparent power is generally measured in thousands (or millions) of VoltAmps (VA), and equals the square root of the sum of real power (kW) squared and reactive power squared. The ratio of real power to apparent power is termed the 'power factor'. Although appliances like electric heaters and incandescent light bulbs may have power factors close to unity, motors (e.g., those in washing machines) typically have power factors closer to 0.8, and some industrial processes have even lower power factors. The reason that apparent power is important is because the capacity of the equipment conveying electrical energy to consumers must be sized to satisfy apparent, rather than real, power requirements. Hence, a power line connected to an appliance which draws 1 kW of power at a power factor of 0.8, must have a capacity of at least $1/0.8$ kVA (i.e., 1.25 kVA).

Although unbundled distributors convey power, this thesis considers that the product which they actually sell is 'connection capacity' at various locations on their network. Connection capacity is sized in kVA, and by paying for connection, the consumer is effectively purchasing the right to consume an

amount of apparent power, up to the level of connection capacity, *at any time*. Weisman (1991) describes any consumer of power as “purchasing a capacity option”. The cost of the connection capacity is the cost of the physical and temporal share of the distributor’s upstream network required to convey power from the transmission grid to the consumer. The consumer pays for the actual energy conveyed separately (or if real time metering and pricing were implemented, then the consumer could pay for power on a moment-to-moment basis). Any consumer’s connection capacity needs to be sized greater than their ‘peak power requirement’ (measured in kW), adjusted for the power factor. There is a distinction to be made, therefore, between “demand” and “use”. Simply because a durable good is not being “used” does not mean that the “demand” for that good is zero. Consequently, there is demand for capacity of the durable assets involved in electricity conveyance even if electrical energy is not actually being conveyed.

In New Zealand, the electricity retailer is usually the purchaser of network connection on the consumer’s behalf, and it bundles this cost together with its wholesale energy purchases, to sell ‘delivered energy’ to the consumer (§3.3.1). The retailer’s selling price comprises both the cost of electricity production and its conveyance. Although, depending on the components of the retailer’s price, this may blur the signal which consumers specifically see relating to the cost of network services, this bundling does not alter the fact that the distributor’s product is connection capacity. Consequently, even though cost complementarities exist between electricity production and conveyance (§3.4.2), New Zealand’s electricity supply industry structure truly allows connection capacity to be examined as a distinct product in its own right. Nevertheless, connection capacity and electrical energy are ‘complementary goods’, where electrical energy is ‘downstream’ of capacity, in the sense that, while electrical energy cannot be consumed without some corresponding demand for capacity, a demand for capacity can exist in the absence of demand for energy. For instance, industrial processes which generate their own electricity, either directly or as a by-product (e.g., ‘co-generation’), may not need to draw much energy from the local distribution network. However, they may require the guarantee of a backup supply in the event of an outage of their own generation plant. Consequently, their demand for connection capacity might be high, even if that capacity is rarely used to draw electrical energy.

Because the consumption of electrical power changes over time, and the consumption patterns of residential and commercial consumers, as well as of various industrial processes, are very different, a distributor’s product has often been distinguished by the ‘class’ of consumer, particular in distribution tariff schedules (e.g., Scherer, 1977; and with respect to New Zealand, Kask, 1988b, p. 9). Different consumption patterns are typically identified by the ‘load factor’ associated with a particular consumer class. This is the ratio of the actual energy consumption in kWh over a particular period (usually one year), to the maximum possible consumption of electrical energy over that period, which (for one year) is

8,760 hours multiplied by the peak power requirement for that consumer or group of consumers.¹ Although the demand for capacity, power and for energy are clearly related through the load factor expression, the costs of the distribution network are much more strongly correlated to the product it actually sells—connection capacity—than it is to the amount of energy conveyed (e.g., Mercury Energy Lines Business, 1999, p. 5). And as noted in the previous paragraph, in some cases a consumer’s demand for capacity may even be wholly independent of its demand for energy.

However, the time dimension does have a significant impact when it comes to considering the cost of energy conveyance. Because not every consumer’s peak power or ‘maximum demand’ requirement necessarily occurs at the same time, the peak power requirement of a *group* of consumers will only be the sum of all the individual maximum demand requirements if all those consumer have homogeneous energy consumption patterns. The more heterogeneous the consumption patterns are, the lower will be the maximum demand requirement for the group. This characteristic of electricity supply is termed ‘diversity’, and it is measured by the ‘diversity factor’ (or ‘co-incidence factor’): the ratio of the maximum demand exhibited by a group of consumers, to the sum of the maximum demands of the individual members of that group. (A value of unity for the diversity factor does *not* indicate homogeneous energy consumption patterns, but it does indicate that the maximum demands all occur at the same instant). Since the capacity of distribution equipment upstream from consumers is sized to be sufficient to meet the maximum demand experienced at the equipment’s location, diversity has a significant impact on the required installed capacity of distribution network assets, and hence on the overall cost of the network. At each transformation to a higher voltage level there is an opportunity for there to be increased diversification of demand.

To describe the effect of diversity on network costs, Boiteux and Stasi (1952) provide a taxonomy of distribution assets on the basis of the “responsibility” a consumer has for the cost of assets upstream from its location in the network. Boiteux and Stasi distinguish three key zones upstream of any consumer: (i) the ‘collective network’, namely the remotest part of the distribution network from most consumers, the ‘installed capacity’ (in kVA) of which depends primarily on the average demand of consumers at the time of the collective network’s maximum demand; (ii) the ‘individual connection’, the capacity of which is solely dependent on the demand of a single consumer (often termed the consumer’s ‘dedicated assets’); and (iii) the ‘semi-individual network’, the capacity of which depends on the uncertainties of consumption of each downstream consumer. If connection capacity at different locations

¹ Such approaches to electricity tariff setting have been criticised as being inconsistent with marginal cost pricing (§4.1.1 and §4.2.3), and it has also been concluded that it is often difficult to find any strong correlation between load factors, consumption levels and consumer class (e.g., Turvey and Anderson, 1977, Ch. 16; Munasinghe and Warford, 1982, Ch. 6).

is a distributor's product, then distribution assets comprising the collective network, the semi-individual network and the consumer's dedicated assets, are the key 'inputs', or 'factors of production'.

One further characteristic of distribution networks that is of importance, is that the conveyance of electricity causes electrical energy 'losses'. As electric energy is conveyed through distribution network (as well as generation and transmission) equipment, energy is lost in the form of heat, due to the resistance of electrical conductor (i.e., both lines and cables), and as a result of so-called copper (and iron) losses caused in transforming voltages (e.g., Cichetti, 1977, p. 20). The majority of the amount of electrical energy lost is directly related to the electric resistance of the equipment, as well as to the square of the electric current (measured in Amps) conveyed through that equipment. However, larger capacity equipment with the same function, tends to have a relatively lower resistance. Consequently, for the same amount of power conveyed, losses will be much lower in a lightly-loaded item of equipment, than they would be in a smaller item of equipment operating close to its operational capacity. Thus, the cost of losses are a function of distance, voltage level and voltage changes, and equipment capacity.

Losses also cause voltage to drop with distance from the generation or point of supply source. This is important because end-use equipment is designed to operate within a particular voltage range. Should delivered power drop below the lower limit of some item of electrical equipment's voltage range, then it will either cease to operate, or it will operate at reduced quality of service (e.g., severely dimmed lightbulbs).

3.1.2 Consumer Demand for Network Connection Capacity

Now that an electricity distributor's product has been defined, it is important to consider the characteristics of consumer demand for that connection capacity. Demand drives the investment in distribution network capacity (e.g., Vogelsang, 1994; Lesser and Feinstein, 1999), and it is the cost characteristics of those network assets which identify whether electricity distribution is a natural monopoly. Joskow and Schmalensee (1983, p. 59) suggest that electricity distribution is a multi-product activity because of variations in customer location and in the times which customers purchase electricity. Similarly, Boiteux and Stasi (1952) concluded that, if the costs of distribution network expansion are rigorously calculated, no two customers could ever be offered the same tariff, since their situation on the system is never identical. Nevertheless, the demand for capacity associated with a particular consumer, given that this demand is satisfied by connecting the consumer to long-lived, lumpy assets, does not necessarily change much with time.

The important distinction between the demand for power and the demand for capacity needs to be reiterated at this point. Because capacity is sized to meet the consumer's demand for peak power, it is sometimes assumed that in off-peak periods consumers should not have to contribute to the cost of capacity. This is the traditional peak load pricing problem (§4.1.2). However, if an electricity line

business is considered to sell the *right* (or option) to a level of connection capacity, then, as long as the consumer's peak power requirement does not change, its demand for capacity remains constant, even while its moment-to-moment demand for power fluctuates. In fact, as long as the consumer is connected to the network, and has the potential to demand power up to the level of connection capacity, then the demand for capacity at the consumer level remains unchanged.²

Demand for capacity predominantly increases in a distribution network due to the connection of *new* consumers, rather than due to a greater demand for capacity coming from existing consumers. In either case, consumers will enter and leave the market—or move around within the market—from time to time, (re)locating at points in the network where there is an existing connection satisfactory for their needs, or at sites where an existing connection needs to be resized or built from scratch. New consumers may request connection at a new '*greenfields*' site (i.e., where there is no pre-existing supply capacity), or they may move into premises vacated by a previous consumer. In the latter case, if such a consumer has a similar, or a lower, demand for energy to the previous consumer at that location, then it is unlikely to be cost effective to change the dedicated connection capacity at that location, and there will be no change in the costs of connection. This is because, for residential and small commercial consumers that take supply at LV level, standard sizes of local connection equipment usually make it optimal to size most connections for only two (or possibly three) different levels of capacity, so that a consumer's purchase of a new appliance, or changes in the ownership of a particular household, do not warrant the costly replacement of the dedicated connection assets already installed at that site.³

Similarly, an industrial consumer's demand for capacity is only likely to change if it expands its plant or changes its process. This would suggest that, as long as the willingness-to-pay of a consumer exceeds its total electricity bill (both for connection and energy consumption), then that consumer's demand curve for capacity alone could well be flat (i.e., horizontal). In other words, demand is 'perfectly inelastic' up to the price level at which demand ceases completely.⁴ Furthermore, the demand curve will not continue indefinitely, but will cease at the point where a standard size of dedicated connection

² Nevertheless, even though the consumer's peak power requirement may not change, if it changes its power consumption pattern, then this may affect the upstream diversity factor, and thus the required capacity of upstream network assets. Consequently, even though the consumer's individual demand for dedicated network capacity has not changed, the result of a change in the demand for power may significantly alter the total cost of conveyance. The question is whether the distributor optimally planned for this possibility, and sized the upstream network capacity accordingly (§6.1.1).

³ For example, New Zealand's current ODV Handbook (Energy Markets Regulation Unit, 2000c, Table B.1; §2.4.3 and §7.4) only explicitly provides for *two* different sizes (i.e., capacities) of customer service connection for LV connections: a standard single phase service connection, appropriate for most houses; and the three phase service connection, required for small businesses, or houses with special equipment that draw an atypical amount of power, such as a spa pool.

⁴ Miller (1995), for one, notes the high inelasticity of demand in network utilities generally. The validity of a flat demand curve for network connection is discussed later (§5.1.3).

capacity is sufficient to meet that consumer's peak power requirements. In other words, demand is subject to 'indivisibilities' inherent in distribution network equipment. Since each consumer is associated with a different product, the demand curve for each product has the same characteristics.

Therefore, demand for capacity increases in two main ways: (a) greenfields developments, such as new residential subdivisions, commercial office parks, and industrial developments; and (b) system augmentation, existing service areas experiencing a greater concentration of development, such as through infill housing, the construction of multi-storied apartment or commercial buildings in place of smaller complexes, as well as new 'spot' loads (such as a new industrial development in an established area). Older residential areas are therefore unlikely to experience significant growth in the demand for connection capacity, whereas industrial zoned areas at the fringes of an urban area may experience rapid growth in the demand for capacity. As Lesser and Feinstein (1999) explain: "Growth trends are driven by real events such as changes in zoning laws and shifts in the local economic conditions".

A key feature of the demand for connection capacity is that, to ensure demand is met, the distributor needs to build some level of redundancy into the network to allow for the possibility of equipment failure. There is of course a trade-off between the cost of improvements in the security of conveyance, and the consumer's willingness-to-pay for greater 'reliability'.⁵ In dense urban areas, distribution networks are typically constructed as a mesh, similar to a transmission network, but are operated radially. In the event of a fault, the mesh design allows the fault to be isolated, and electricity conveyed to consumers along an alternate path (although the new operational configuration is still radial) with no interruption of service. Another way of improving security is to install network equipment in pairs, particularly transformers at zone substations. Large industrial consumers in particular, for whom interruptions in supply may be very costly (e.g., Sharp *et al.*, 1985), may be prepared to pay for dual lines and paralleled distribution transformers. Generally, equipment installed in pairs (or sometimes in threes) are the same capacity as each other. This reduces fault levels and the possibility of equipment damage during outage conditions.

3.1.3 Efficiency Implications of Distribution Network Planning and Design

Willis *et al.* (1995) list three categories of distribution planning: (i) new expansion planning, such as the development of a greenfields site at the periphery of an existing network; (ii) system augmentation planning, the upgrade of an existing subnetwork to accommodate new demand (or 'load') growth; and (iii) operational planning, which primarily relates to the switching patterns of an already-built subnetwork. All of these planning activities, and thus optimal distribution network design, require making a trade-off between the capacity and operational costs associated with the network, and the cost

⁵ In practice, the level of security is measured through various reliability indicators. Hence, an improvement in security is usually identified through increased reliability.

of energy losses (§3.1.1). The cost of losses is equivalent to the ‘opportunity cost’ of the electrical energy which is unable to be sold as a result as those losses (§4.1.5). The energy losses incurred in transmitting that energy would not need to have been incurred had that energy been sold at its source (i.e., the power generation plant), hence the opportunity cost of losses is the forgone revenue from energy sales lost due to power transmission and distribution. Since the forgone revenue is dependent on the current price for electrical energy, the opportunity cost of losses will depend on the market price for electricity, and will vary as that price fluctuates.

The three key interrelated aspects of distribution network configuration that need to be considered in making such a trade-off between network capacity and operating costs are: (a) the asset configuration of the network (i.e., its layout and routing); (b) the most appropriate asset size for each part in the distribution system (particularly conductor and transformer capacities); and (c) the location of normally open switches in the network, which make a meshed design operate in a radial manner. This trade-off will also be made within the context of the network’s ‘contingency criteria’—namely, the target levels of system reliability that specifically impact the configuration of the distribution network, and any redundancy which might be built into the system. The first two aspects of distribution network configuration impact dynamic efficiency. As is discussed below (§3.5.1), any decisions made with respect to network design (i.e., asset configuration and/or size) will affect and constrain later decisions of a similar nature. Such decisions are therefore ‘*intertemporally interdependent*’. On the other hand, the switching configuration relates to a much shorter time frame, and decisions made now relating to the currently-optimal configuration, do not preclude other possible configurations at some later period in time.

New expansion planning is complicated by the sheer number of possible design options, while system augmentation is complicated by the number of constraints, such as practical limitations on equipment sites and conductor routes. Given the non-linear nature of electrical losses, finding an optimal solution to any of these highly complex planning problems—even if the firm were to have access to perfect information—requires the utilisation of non-linear optimisation methods. Because of the complexity, even the computer programs used by large US utilities to aid in distribution system planning typically simplify the formulation of the problems, and use linear programming techniques instead.⁶

The three main cost drivers associated with a distribution network are demand for capacity, distance factors, and security standards. In particular, security (or contingency) levels may require the

⁶ Willis *et al.* (1995) suggest that the use of software utilising just *linear* programming techniques can reduce costs by 5-10% on average compared to manual design approaches, although they note that IEEE and EPRI publications have reported economic savings of up to 19%. They suggest that linear programming techniques can achieve up to 85% of the cost savings which would be achieved through using more accurate non-linear programming techniques.

duplication or oversizing of some assets. However, the additional capacity costs due to added security will be at least partially offset by a reduction in the cost of losses, and by reduced voltage drop. The cost of lines and cables are clearly impacted by capacity, demand and the distance over which power must be distributed. While transformer costs are primarily related to capacity and reliability, these are also indirectly impacted by distance. Because of problems of voltage drop (§3.1.1), there will be a cost trade-off between the number of zone substations and the geographical area served, and the greater the number of zone substations, the greater the number of costly zone substation transformers.

In New Zealand, such distribution planning problems were only just beginning to be addressed through the application of computational techniques prior to distribution sector reform. Optimisation techniques were generally limited to operational planning through analyses of optimal switching configurations. Given the cost of the software involved, which at that stage had to run on a mainframe computer, ESAs typically did not have their own in-house ability, but engaged external consultants.⁷ Post-reform, there is little incentive for ELBs to minimise losses by realising the ‘cost complementarities’ between network design and energy conveyance (§3.4.2). Average loss factors are calculated by each ELB for various parts of their distribution networks. Electricity retailers are billed by the electricity wholesaler for the cost of losses, based on these loss factors, but the retailer can simply pass this cost directly through to the consumer. There is therefore no direct incentive for the ELB to optimise its system design with respect to losses. Although the ODV methodology (§2.4.3 and §7.4) is intended to value an ELB’s network based on an optimal network configuration, there is no requirement for ELBs to use such readily available optimisation techniques in system design.⁸

In particular, the *time dimension* aspect of distribution planning is crucial, given the long lifetime of the assets involved (typically 15 to 60 years), and Lesser and Feinstein (1999) maintain that an optimal solution requires the application of *dynamic* programming techniques (e.g., Hillier and Lieberman, 1986) using some probabilistic model of load growth (e.g., Feinstein *et al.*, 1997). Lesser and Feinstein state that “the objective of the distribution utility capacity planning problem is to meet customers’ capacity and energy needs over the indefinite future at the lowest expected present value of all future cost,” and they sum up the characteristics of the “distribution investment problem” as being “long-term, dynamic, and uncertain”.

⁷ The present author was employed by such a consultancy in New Zealand from 1986-1990, and was heavily involved in such optimisation and distribution planning studies. Hence, the discussion in these subsections is tempered by direct personal experience.

⁸ This is not meant to imply that ELBs are *not* using such techniques, but simply that the governance structure provides them with no direct incentive to do so.

3.2 Perspectives of the Nature of Electricity Distribution

3.2.1 *Electricity Distributors: Natural Monopolies, Potential Competitors, or Essential Facilities?*

Currently, the prevailing view is that distribution networks are naturally monopolies (e.g., Steiner, 2000, p. 9). Yet only twenty years or so ago, the entire electricity supply chain was considered to have strong natural monopoly characteristics due to substantial economies of scale—if not as a whole, at least on a regional basis. Accordingly, electricity supply was typically a vertically-integrated activity of generation, transmission and distribution (§2.2.2). As it became recognised that competition in generation was possible, and the wave of power industry deregulation began, electricity distribution as a standalone business was itself still perceived as a vertically-integrated activity of bundled network services and energy sales.

For instance, during the 1980s, comments similar to the initial quote presented at the beginning of this Chapter by New Zealand analysts, regarding bundled distributors, were also to be found in the international literature. Joskow and Schmalensee (1983, p. 59) observed that: “Little theoretical or empirical work focuses explicitly on the economic characteristics of electric power distribution systems. It is generally thought that they have important natural monopoly characteristics within limited geographic areas, although these areas could conceivably be smaller than the boundaries of a large city”. And as recently as 1989, the New Zealand industry analysts quoted at the beginning of this Chapter, were still referring to *bundled* distribution services as a natural monopoly.

Since 1998, New Zealand’s electricity distributors have been completely separated into lines and energy businesses (§2.4.5), although as early as 1993, distributors had to transparently separate the accounts of the two business functions (§2.4.2). Electricity line businesses (ELBs) now manage the network services functions of electricity conveyance and network connection, whereas energy retailers handle the competitive business of energy sales, since it is now considered that electricity retailing is not a naturally monopolistic activity after all, and it should be separated out from the business of network connection. Nevertheless, the assumption that the remaining unbundled line businesses are natural monopolies is still prevalent in the international literature. Given that the classification of natural monopoly has been lifted from other aspects of the electricity supply chain in recent times, it is also worth questioning whether this label is still appropriate for the function of electricity distribution in its own right.

Although proposals for promoting competition in parts of the electricity supply industry had already been around for some time (e.g., Phillips, 1975), Joskow and Schmalensee were one of the first to write a text which comprehensively analysed the prospects for deregulating all components of the power

sector and for introducing market forces.⁹ However, they made the assumption that “whatever structural changes are applied to the electric power industry, the distribution function will be performed by a franchised monopoly enterprise. This assumption is made by almost all other analysts, and distribution seems to have such pervasive natural monopoly characteristics that franchised distribution monopolies (whether regulated or publicly owned) are almost certainly the most efficient type of organization” (Joskow and Schmalensee, 1983, p. 60).

One of the few dissenting voices to this view had already been provided by Primeaux (1975), although he was still talking about vertically-integrated power utilities that offered bundled distribution services. Primeaux noted that: “Monopoly in the local electric utility industry is so taken for granted that it is almost forgotten that competition ever existed”, and analysed cases in the US where duopolistic competition had occurred.¹⁰ Duplication of facilities had sometimes existed; lines from two different distributors ran down the same street, and as such, consumers could switch power suppliers at will. He concluded that there appeared to be certain circumstances in which competitive electric utility firms produce at lower average costs than monopoly utilities, at least “within certain size limits”. Primeaux also foreshadowed the later application of contestability theory in relation to electric supply, noting that: “Merely the threat of competition should tend to make the monopolist perform well”. Nevertheless, for various reasons, mainly relating to perceived flaws in Primeaux’s econometric methodology, Joskow and Schmalensee (1983, p. 62) considered that “too much attention has been paid to [Primeaux’s] results, and we conclude that they do not cast appreciable doubt on the proposition that distribution is a natural monopoly”. And in commenting on New Zealand’s reforms in light of Joskow and Schmalensee’s conclusion, Michaels (1989) stated that there was little international evidence that structural competition, or duplicative duopoly, could be of any value in electricity distribution.

Hobbs and Schuler (1986) are one of the few to discuss possible modes of power distribution competition, although they are also discussing bundled distribution services, again in the US industry. Their modes include: (i) “borderline competition” (sometimes termed boundary competition or fringe competition), relating to competition between neighbouring distributors to serve new housing projects or commercial and industrial developments near the boundaries of existing service areas; (ii) “duplicative competition”, referred to by other authors as “bypass”, the form of competition examined by Primeaux;

⁹ This was, however, in the context of the US power utility industry, which was predominantly owned by private sector interests—subject to strong regulatory control—rather than to Government-owned entities.

¹⁰ Primeaux (1975) states that, in 1966, direct distribution-level competition between at least two electric utility firms existed in 49 US cities with populations of 2,500 or larger. Neufeld (1987) characterises the incipient US power industry earlier still—in the first decade of the 20th century—as being highly competitive, although this competition primarily arose not between utilities themselves, but in terms of the prices that utilities offered to large industrial consumers to convince them not to self-generate.

(iii) “industrial location”, relating to distributors attempting to attract new large consumers into their service area; (iv) “franchise competition”, based on the work of Demsetz (1968), as well as being Joskow and Schmalensee’s prescription for distribution sector competition; (v) “institutional competition”, through the threat of municipal or public takeover of private distributors; and (vi) “co-generation”, in other words, self-generation of power made viable by utilising the waste heat from industrial processes.¹¹

Of these various modes of competition, bypass in network industries has been given some (mainly theoretical) attention in the international literature, both with respect to the distribution of power (e.g., Black, 1994; King and Maddock, 1996) and also of natural gas (e.g., Broadman and Kalt, 1989; MacAvoy *et al.*, 1989).¹² MacAvoy and his colleagues were concerned about the potential for wasteful duplication in gas distribution should competitive entry be allowed in gas markets. However, their concern related to partial deregulation, in other words, removing entry barriers while still retaining some forms of price control, particularly those which resulted in inefficient prices. Broadman and Kalt were also concerned that bypass in gas transmission and distribution might be inefficient, but felt that incentives for bypass generally resulted from distorted pricing practices, including cross-subsidisation between consumer classes. Bypass was not seen as being desirable *per se*, but the argument in favour of

¹¹ Apart from their taxonomy of distribution competition modes, Hobbs and Schuler (1986) provide a theoretical spatial model of oligopolistic boundary competition. By investigating the equilibrium prices of oligopolistic distribution versus monopolistic distribution (under uniform pricing), they assessed how much productive efficiencies would have to increase as a result of deregulation, to offset losses in allocative efficiency, because prices were expected to rise due to competition. This is because two firms would be operating at a higher point on the average cost curve. One interesting aspect of their analysis, which has a bearing on appropriate valuation for ELBs (§7.3.3), is that, under competition, prices will rise to the replacement costs of distribution facilities. Since replacement costs are higher than historic equipment costs, it is more attractive for a firm to charge a high price to its remaining (captive) consumers, and to let other firms take away its existing consumers at the border, rather than to meet the competitor’s price.

¹² The telecommunications sector is the network industry which has received the most attention in the academic literature in recent years. Discussions regarding bypass appear throughout (e.g., Sharkey, 1982a, pp. 181-213; Curien, 1991; Baumol and Sidak, 1994a; Sidak and Spulber, 1997). Although this dissertation draws on some of the literature relating to the telecommunications sector, there are greater differences between electricity distribution networks and telecommunications systems, than there are between power and gas distribution networks. Many of the differences between telecommunications and power networks are outlined by Evans and Quigley (1998). For example, power and gas networks are generally radially operated, involve one-way traffic, and do not experience as significant “positive network externalities” when compared to telecommunications networks. Also, the types of regulatory issues are somewhat different, with access pricing being a key concern in telecommunications, whereas for power distribution, the absolute price level and cross subsidies between consumer classes are of more interest to policy makers in New Zealand (§5.1.2). Moreover, access pricing often involves questions of “bilateral monopoly” (e.g., Economides, 1996), a situation not applicable to distribution networks. Finally, because the technology and the nature of the network is different, possible mechanisms of competition are also different in telecommunications networks (e.g., Baumol and Sidak, 1994a, Ch. 2) from those in electricity distribution networks. Also see the comments by King and Maddock (1996), fn. 15.

not restricting bypass was that it might enhance competition, or the threat of entry might place pressures on prices to become more economically efficient.¹³

Nevertheless, Broadman and Kalt concluded that the gas transmission industry cannot be characterised as contestable market, given its high level of sunk costs, and these sunk costs already provide a barrier to entry to potential competitors, enhancing the sustainability of incumbent firms (§3.5.1). Consequently, concern that any natural monopoly that might exist in gas transmission might be *unsustainable* was likely to be unfounded (§2.1.8), and therefore no regulation to restrict entry should be required. One prudent observation made by Broadman and Kalt, as well as MacAvoy *et al.* (1989), was that the avoided costs and net benefits associated with investments with and without the bypass proposal would need to be assessed carefully. Broadman and Kalt went as far as saying that some degree of physical duplication of assets was, in itself, not necessarily inefficient, and could even be optimal. Consequently, they concluded that the burden of proof should be towards rather than away from bypass, and the appropriate question is not “when is competition excessive?”, but rather “when is regulation necessary?”

Given the paucity of literature on the subject of bypass in electricity supply, Black (1994) resorted to referencing the papers on the natural gas sector by Broadman and Kalt, as well as MacAvoy and his colleagues, when discussing the issue. Black suggested that their conclusions were of less relevance to the power industry, and asserted that if any bypass actually occurs in power markets, then it is unlikely to be uneconomic. However, Black viewed bypass as primarily a means for consumers to access cheaper wholesale electricity, rather than to obtain a less expensive distribution network connection. Consequently, in Black’s view, any potentially inefficient duplication of distribution assets would be offset by substantially reduced generation costs. Such an analysis is not applicable in a New Zealand context, since the wholesale market is separate from any “market” which might exist for network services. A change in network provider will not necessarily alter the wholesale electricity component of a consumer’s bill.

More recently, in an Australasian context, King and Maddock (1996) have labelled (residential) electricity distribution networks as a special type of natural monopoly—namely, an “*essential facility*”.¹⁴ King and Maddock present a taxonomy of monopolies involving two tests: firstly, whether the facility in

¹³ Broadman and Kalt (1989) describe that bypass in gas transmission comprises three different forms of entry: (i) ‘complete bypass’, equivalent to duplicative competition or borderline competition; (ii) ‘partial bypass’, which relates to firm versus interruptible service, and has no direct equivalent in electricity distribution; and (iii) ‘contract carriage’, which is equivalent to the ‘direct access’ phase of introducing retail competition in electricity distribution. The various modes of complete bypass were their key concern, given that this would involve some duplication of assets.

¹⁴ King and Maddock (1996) also label local telecommunications networks as essential facilities, an association for which the term is often applied (e.g., Baumol and Sidak, 1994a, p. 7; Sidak and Spulber, 1997, pp. 48-50).

question involves a natural monopoly technology; and secondly, whether the facility provides an input which is essential to the provision of another good or service. Facilities which meet both tests are classified as essential facilities, whereas those which fail both are “competitive facilities”. In between lie “almost essential facilities”, the first type being a facility which is not a natural monopoly but does provide an essential input (i.e., a “regulatory monopoly”), and the other involving a natural monopoly which does not provide an essential input (i.e., a “convenient facility”). An input is only deemed essential to the provision of a specific good or service if there does not exist either: (a) an alternative input that can enable a competitor to provide an equivalent final product at a comparable cost (e.g., a different upstream bypass network in the case of electricity distribution); or (b) an alternative final product that is able to be supplied at a competitive price without that input (e.g., self-generation of electricity).¹⁵

The problem with essential facilities is that, even if competition is introduced into the market for a final product involving an upstream essential facility, efficient pricing is unlikely to occur. Since the facility is essential, the upstream natural monopolist can abuse its monopoly power by manipulating the price of the final product. Allowing bypass of the essential facility is not the solution however, since King and Maddock state that any duplication of an upstream facility would be socially wasteful by definition. Rather, the solution is to establish a regulatory regime which requires the monopolist to allow competitors “open access” to its essential facility based on an efficient access price, one likely to be specified under the regulatory regime. Although the New Zealand Government did not explicitly refer to the essential facilities concept in its reform of the electricity distribution sector, retail competition has been introduced in electricity supply in accordance with this open access principle.

3.2.2 New Zealand Government Perspectives of Electricity Distribution

New Zealand’s power sector reforms clearly indicate that successive Governments recognised the paradigm shift that had taken place in regard to electricity distribution—what had been termed “distribution” involved not only monopolistic line functions, but also potentially competitive energy sales functions. By 1998, the Government had stated this in black and white: “Electricity retailing is not a natural monopoly” (Energy Markets Policy Group, 1998b).¹⁶ In parallel, successive administrations, as well as the Commerce Commission (e.g., Pickford, 1996), also described the remaining lines part of the

¹⁵ This definition of an essential input is used by King and Maddock (1996) to distinguish between power and gas distribution networks. Although electricity distribution is viewed as an essential facility, gas networks are considered to be only a convenient facility. They argue that natural gas is not an essential input for end-uses, such as heating, since close substitutes in the form of other fuels exist. By contrast many uses of electricity, particularly for residential purposes such as lighting and operating small appliances, have no substitutes.

¹⁶ Not long beforehand, the Ministry of Commerce and the Treasury (1995, p. 4) had still been referring to bundled electricity distribution as a “vertically-integrated natural monopoly”.

distribution business as a natural monopoly: throughout press releases, papers, regulations, legislation and policy statements (§2.3). But in more recent times, this description of electricity distribution has been qualified somewhat, particularly by Government Ministries, and other terms have been used, including: “natural regional monopoly” (Energy Markets Policy Group, 1998b); “strong natural monopoly characteristics” (Energy Markets Policy Group, 1998c); and most recently, “effective monopolies” (Ministry of Economic Development, 2000b). Similarly, the recent Inquiry into the Electricity Industry (Ministry of Economic Development, 2000a) stated that, “in spite of the abolition of franchise areas, distribution companies essentially remain monopolies”, even if these monopolies are not strictly natural. Earlier on in the reform program, the language had been more definite. For instance, the first edition of the Optimised Deprival Valuation (ODV) Handbook (Energy Policy Group, 1994b, p. 5; §2.4.3) lists one of the “fundamental starting point assumptions” underlying the ODV valuation methodology is that an ELB is a natural monopoly.

Notwithstanding the terminology, the justification and desired outcomes of reforms in the distribution sector were clear. The removal of the franchise boundaries, and the unbundling and subsequent ownership separation of lines and energy businesses, were designed primarily to promote competition in *retailing*, resulting in greater *allocative* efficiencies, whereas in *distribution* it was to allow ESAs to rationalise through mergers, hopefully resulting in gains to *productive* efficiencies. The general approach throughout the reform process, made explicit in the 1998 Policy Statement, has been the introduction of competition into the potentially-competitive activities of electricity wholesaling, retailing and generation, and the imposition of pro-competitive regulation on the naturally-monopolistic activities of transmission and distribution (§2.3.2). Nevertheless, the possibility of some competition for network services was noted. Although stating that “the distribution of electricity is a natural monopoly service”, and that “consumers are not likely to have alternatives to using distribution networks in the foreseeable future”, the Ministry of Commerce did acknowledge that an exception was the “limited cases such as where bypass is viable” (Energy Markets Policy Group, 1998b). Consequently, in the ODV valuation methodology, the economic value of any network segment is capped via the possibility of bypass (§7.4.3)—in other words, direct supply by a neighbouring ELB, or by a competing ELB connected back to the nearest Transpower point of supply (§2.2.1).

3.2.3 Contestable Functions of Electricity Line Businesses (ELBs)

One ELB states that bypass and other forms of network competition are already occurring, and claims that ELBs are not natural monopolies at all. Instead, ELBs are labelled regional monopolies, able to raise ‘barriers to entry’, such as through the terms of contractual agreements for interconnection. (Conversely, it might be argued that ELBs can be viewed as regional monopolies because of barriers to entry, rather than vice versa; §3.4.5). The current monopolistic industry structure is explained as being the outcome of historical regulatory interventions (Mercury Energy Lines Business, 1999, p. 9). Under the essential facilities taxonomy discussed above (§3.2.1), King and Maddock (1996) state that facilities

(like ELBs) which are “usually monopolies for historical reasons”, are typically “regulatory monopolies”. King and Maddock’s prescription for improving efficiencies in regulatory monopolies is to encourage competition, and to require open access as one way of reducing barriers to entry during the transition period from monopoly to full competition.¹⁷

In presenting its argument that ELBs are not natural monopolies, Mercury Energy Lines Business (1999, p. 10)—henceforth ‘Mercury’¹⁸—provided a useful breakdown of the functions and assets relating to a line business. An ELB is stated as comprising the functions of: (i) network planning; (ii) customer management; (iii) contracting for maintenance and construction; (iv) network control or co-ordination, in order to operate the network to achieve certain reliability and quality targets, as well as to agree on the terms of interconnection with other networks (including Transpower); (v) owning or seeking access to corridors for laying cable, siting transformers and other equipment; and (vi) ownership of reticulation assets (i.e., the network itself). Mercury does not even mention the traditional activity of network construction, since as Berger and Spiller (1996, p. 8) point out, a national market in network contracting services has already evolved in New Zealand, with construction and installation of most network assets contracted out—function (iii) above.

Mercury suggests that functions (i)-(iii) clearly do not involve any characteristics of a natural monopoly. Although conceding that incumbent providers of these services may retain some initial advantage, no serious barriers to entry exist to obtaining planning, customer management or contracting capabilities. These are seen as being separate activities which may be more cost effectively provided by specialist firms that provide such services for a number of ELBs, and even for network businesses which require similar functions in other industries (e.g., gas, water, roading and telecommunications).¹⁹

¹⁷ When applied to ELBs, open access here refers to the encouragement of retail competition. Unlike essential facilities, King and Maddock do not see bypass as being inefficient for regulatory monopolies, but open access is suggested as the first step to lowering entry barriers because it is recognised that any new infrastructure forming the basis of long-term competition “cannot be built overnight”. Allowing competing firms open access to existing facilities is seen as a way to help competitors establish a market presence and reputation, while they prepare to construct their own infrastructure. As noted later (§3.3.1), somewhat ironically, the current ownership separation of lines and energy businesses may make bypass more difficult. A competing ELB, although it may have an established reputation, cannot familiarise potential consumers with its product offerings without making a substantial outlay on infrastructural assets.

¹⁸ This ELB is now named Vector Ltd., and the name ‘Mercury Energy’ is used by what was previously the incumbent electricity retailer of the bundled EC called Mercury Energy Ltd.

¹⁹ Considerable economies (§3.4.2) are suggested by consolidating meter readings, billing, customer information, and customer response activities across network businesses. Similarly, many network businesses also require expensive trenching of equipment (e.g., power and communications cables, and pipes), and Mercury suggests that considerable savings could be made through some rationalisation of these activities.

On the other hand, functions (iv) and (v) are cited as containing elements of natural monopoly. Notwithstanding these characteristics, Mercury points out that the control function could co-ordinate assets owned by different companies, such as assets owned by third parties, for instance residential subdivisions (§3.4.1). Consequently, Mercury also proposes that further rationalisation of network co-ordination might be possible, with the implication that such a function could be handled by a separate firm from the ELB. The most efficient number of co-ordination firms would depend on at what point economies of co-ordination might reach their limit. As for access to corridors for siting assets, Mercury (p. 12) indicates that generally ELBs neither own nor have exclusive rights to such corridors, since these rights are held by local government authorities.

Finally, Mercury assesses the network assets themselves, and in a similar approach to Boiteux and Stasi (1952) described above (§3.1.1), categorise assets into three distinct elements: (a) direct high voltage (HV) connection, assets directly connecting large HV consumers to the Transpower point of supply; (b) the medium voltage (MV) network, which both directly supplies medium to large consumers as well as the LV network; and (c) the low voltage (LV) network itself, which supplies small consumers. Since direct HV connection involves readily identifiable dedicated assets, and is sometimes provided by the consumer itself rather than an ELB, Mercury (p. 13) suggests that these elements of the network are clearly contestable. Large HV consumers can seek proposals from alternative suppliers nationally or internationally, and compare these with proposals for constructing the assets whereby the consumer retains ownership. (For instance, some ELBs do not own any HV connection assets). As for the other two elements of the network, Mercury assesses their natural monopoly characteristics on the basis of examining their average cost curves, and the validity of this approach is discussed below (§3.4.3-§3.4.4).

3.3 Distribution Competition in New Zealand

3.3.1 Possible Modes of Distribution Network Competition Identified in New Zealand

As noted in the opening paragraph of this Chapter, some competition for distribution network services has been observed in New Zealand, even though this was not explicitly envisaged as a reform outcome.²⁰ For instance, in their pre-reform study of economies of scale in New Zealand's electricity distribution sector, Wyatt *et al.* (1989, p. 7) referred to work in the US demonstrating that viable competition in electricity distribution could exist for large industrial loads at the boundaries between the areas of two neighbouring distribution companies (i.e., Weiss, 1975). However, this rather mild implication that distribution competition might become feasible in New Zealand was with respect to competition for *bundled* network and energy services. Similarly, other pre-reform studies of distribution cost and price structures, where addressing the possibility of distribution competition at all, again only

²⁰ Much of the discussion in this sub-section, as well as in §3.3.3 and §3.4.5, is drawn from earlier work by the present author (i.e., Gunn, 1996).

considered bundled competition, or competition in retailing alone (e.g., Kask, 1988; Saha and Sell, 1990).

Likewise, once reform began in the distribution sector the focus of attention was placed firmly on retail competition. Bergara and Spiller (1996) provide one of the rare academic works explicitly acknowledging that competition for construction and ownership of new or rebuilt network assets became possible in New Zealand after the removal of the franchise boundaries. Bergara and Spiller were focusing on modes of competition possible under the open access regime which emerged after the full removal of franchises in April 1994 (§2.4.2), but before the legal separation of lines and energy businesses in July 1998 (§2.4.4). These modes comprised: (i) bypass of the “retail dimension” of the incumbent (bundled) distributor (i.e., retail competition); (ii) innovative energy product offerings (i.e., energy services competition); (iii) large consumers disconnecting from the local HV distribution network and taking supply from the EHV transmission network instead (i.e., transmission bypass); and (iv) expansion of an existing (bundled) distributor’s network by a different (bundled) distributor (i.e., subnetwork competition), particularly in areas of rapid growth.

Retail competition for electricity supply, the first of these modes of competition, has of course finally become a reality in New Zealand, if still on a somewhat limited scale. The 1998 reforms saw the number of consumers switching their energy retailer increase from 3% to 18% by the end of 2000 (Energy Markets Policy Group, 2001b). Although usually the energy retailer bundles energy and network services together in its commercial relationship with the consumer (i.e., a delivered energy agreement), it is not able to own any part of the network itself. Large consumers, particularly if they have special security and quality of supply requirements, may prefer to enter into distinct agreements with the energy retailer (i.e., a supply agreement) and with the ELB to whose network it is connected (i.e., a line function services agreement). The first commercial relationship is often termed a ‘retailer interposed’ arrangement, and the second, a ‘conveyance’ arrangement (e.g., Lockhart and Mallon, 1995). Energy services competition is a special form of retail competition relating to the provision by an energy service company of energy end-uses, such as heating and cooling, rather than electricity *per se*.

To their list of possible modes of competition, Bergara and Spiller could have added ‘distribution bypass’, the mode of competition identified by the Ministry of Commerce itself (§3.2.2). This comprises the boundary competition discussed above, but also the case where an external ELB provides the link between the consumer and the Transpower point of supply. The difference between this latter mode of competition and the transmission bypass mode outlined above, is that it involves an external ELB. As is discussed elsewhere (§3.2.1 and §6.1.1), the asset duplication which can result from either transmission or distribution bypass (depending on how ‘duplication’ is defined), is not necessarily dynamically inefficient. However, unlike bypass modes, subnetwork competition generally does not require any duplication of assets, and therefore appears both less costly and more efficient than bypass.

In discussing subnetwork competition, Bergara and Spiller were writing prior to the 1998 ownership separation of line and energy businesses. Consequently, their description of this competitive mode relates to an energy retailer charging consumers for connection to a new subnetwork that it constructs and owns (and is embedded in the incumbent distributor's network), as well as for electric energy conveyed. Since a line business is currently prohibited from retailing electricity, and vice versa, such a retailer would now have to find an external ELB willing to construct (and own) the new subnetwork for them. Hence, the current regime might conceivably make subnetwork competition less viable, or at least less attractive than previously, since the contractual relations are more complex, and the associated transaction costs could possibly be higher (§2.1.6). Pre-1998, a bundled delivered energy agreement between consumer and (bundled) retailer, as well as both a 'use-of-system agreement' and an 'interconnection agreement' between the retailer and incumbent line business, would have been required.²¹ Currently, a supply agreement would need to exist between the (unbundled) retailer and consumer, and two use-of-system agreements would be required between the retailer and both the incumbent and external ELBs. In addition, the incumbent and external ELBs would also have to enter into an interconnection agreement. Figures 3.1a-3.1d show a number of the various modes of distribution and retail competition in New Zealand that are possible, and the associated commercial or contractual relationships (modified from Lockhart and Mallon, 1995).

3.3.2 Evidence of Distribution Bypass and Subnetwork Competition

Mercury suggests that the threat of distribution bypass has been taken seriously by some ELBs, including themselves, as the following quote indicates.

Shortly after the franchise area restrictions were removed, a large customer located in an industrial area visited Mercury with a proposal. The proposal was simple—either meet the prices offered by a competing supplier of line services, or the customer and some other firms located nearby would move to the competitor. The proposal was supported with credible plans to install the required capacity. This is competition in action, unleashed by regulatory reform, and it works! Mercury adjusted its product offering to satisfy its customer, and continues to do so in order to remain the network provider of choice in the Auckland region (Mercury Energy Lines Business, 1999, p. 15).

²¹ Of course, since two contracts exist between the same two parties, they could be combined into a single agreement. However, the key point to recognise is that such an agreement would relate to two different functions: firstly, the conveyance of energy purchased by the bundled retailer across the incumbent's network (i.e., the use-of-system agreement); and secondly, the terms governing the interconnection of the incumbent's network with the bundled retailer's subnetwork (i.e., the interconnection agreement). This latter agreement is required, because the incumbent would need to be assured that the operation of any equipment connected to the downstream subnetwork would not have the potential to damage its own network. Every network has a 'distribution code' specifying the characteristics of equipment that can be safely connected to its system, and the terms of the interconnection agreement gives the code contractual weight.

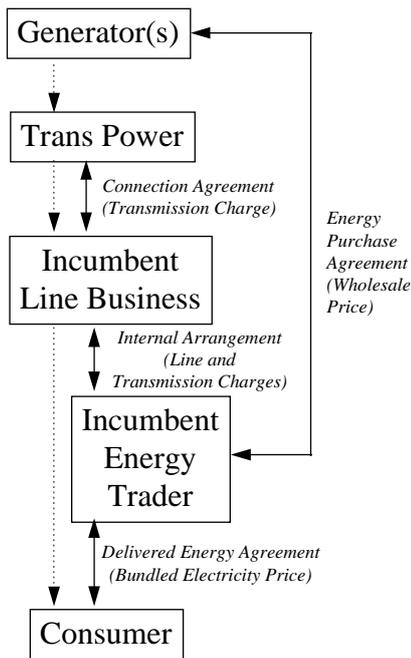


Figure 3.1a: No Competition

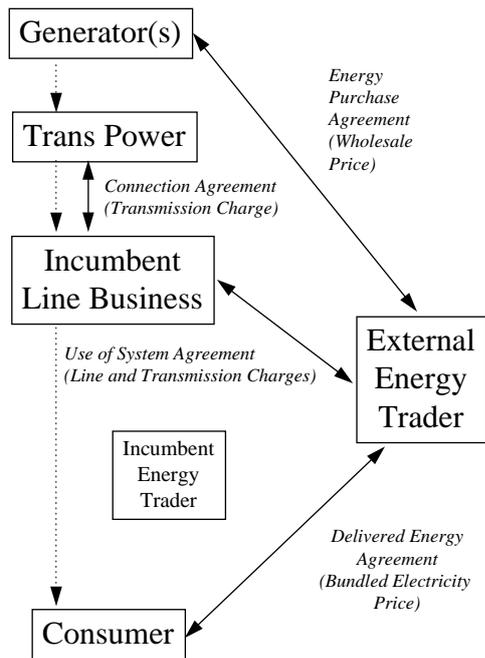


Figure 3.1b: Retailer Interposed Agreement

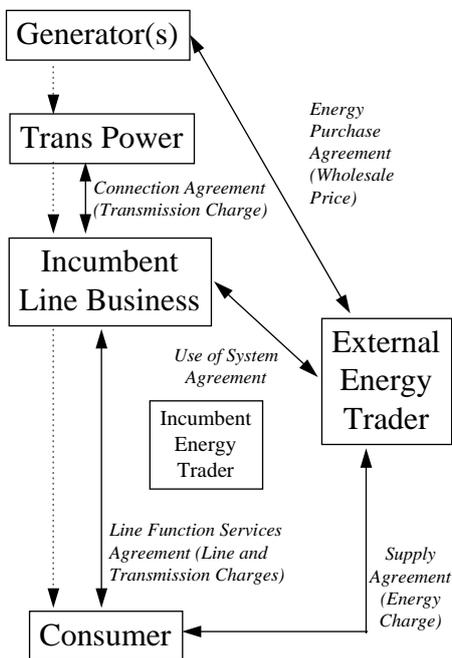


Figure 3.1c: Conveyance Agreement

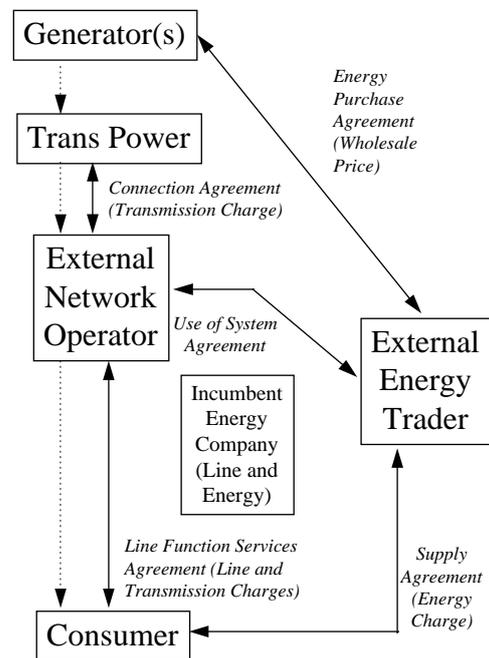


Figure 3.1d: Bypass Scenario

.....> **Energy Flow** <-----> **Contractual Relationship**

This quote was part of an assertion that the threat of new entry had been constraining ELB pricing decisions for HV and MV consumers, and as such, network services were becoming “contestable”. The context for making this assertion was the Government’s 1998 proposal for

introducing more heavy-handed regulation under the later shelved “thresholds” regulatory scheme (§2.4.6).²² Mercury suggested that competition was increasing for MV consumers, and that larger consumers were actively seeking out alternative supply proposals to use in price negotiations, in order to “drive prices toward long run average incremental costs” (§4.1.5). However, it was conceded that “competition is still in its infancy” for existing LV consumers (Mercury, p. 25).

As discussed above (§3.3.1), whereas distribution bypass does not involve a competing ELB requiring connection to an incumbent ELB’s network, subnetwork competition does. New LV consumers potentially can be exposed to some competition under such a mode. In such cases, where the competing ELB reticulates new subnetworks embedded within the incumbent’s network, the competitor itself becomes a customer of the incumbent. The competitor must pay a line charge to the incumbent for the privilege of connecting to its network. For transparency, the *Electricity Act 1992* implicitly requires that this line charge be the same as that which any regular customer of the incumbent would pay for being connected at that point. Although not citing any specific examples, Bergara and Spiller (1996, p. 7) state that such “competition has begun to occur, particularly in conjunction with new development”.

Mercury (p. 16) contends that not only is bypass competition a reality, but so too is subnetwork competition (although they do not use that terminology to describe it). Once the legislative barriers to competition were removed, line businesses began to compete to build networks for new residential subdivisions as well as industrial and commercial office parks. Such competition “challenged established practises, drove down costs and signalled who was best placed to undertake these projects”. However, the outcome of this initial burst of competition appears to have been a realisation that many line businesses were not efficient at building networks. Consequently, as noted above (§3.2.3), Mercury states that most construction work is now carried out by third party contractors.

However, bypass and subnetwork competition which actually result in an external ELB owning assets within a competing ELB’s old franchise area now appear to be exceptions to the rule, and the fact that some competition for network services has been occurring does not necessarily prove that electricity distribution is not in itself a natural monopoly within a particular area (or for a particular group of consumers). Although the notion that direct HV connection consumers are contestable is logically quite compelling, a big question mark still looms over the natural monopoly status of the MV and LV network.

²² Consequently, such a claim might be greeted with scepticism; Mercury could potentially be protecting monopolistic behaviour that was possible under the existing regulatory regime. However, as noted in the Preface and Acknowledgements of this thesis, the present author was actually employed by Mercury at the time the cited proposal was made, and observed Mercury’s response first hand. Although not mentioned in the document quoted here, Mercury themselves engaged in distribution bypass outside their traditional franchise area on at least one occasion. However, this was a direct network

Based on anecdotal evidence, the initial surge of bypass and subnetwork competition at MV and LV level has now died down.²³ One reason could perhaps be that competition for ELB *ownership*, through mergers and acquisitions, has been seen as potentially being more profitable than competition for *consumers* connected to localised subnetworks. Secondly, as indicated above (§3.3.1), the prohibition on joint ownership of line and energy businesses, although making retail competition more feasible, may potentially have dampened incentives for either bypass or subnetwork competition. And the general uncertainty in the future regulatory environment is likely to have increased the perceived risk of engaging in small-scale competitive activity. Possibly such “competitive” activity—or “market discovery” (§3.5.3; §6.1.3, fn. 7)—was the industry simply testing the natural monopoly assumption itself: testing ELB average costs within a particular area as a precursor to ELB amalgamation, or perhaps “diversifying into largely unsuccessful areas” as part of the “transition from protected monopoly to market competition” (Gardner and Gilson, 1994).²⁴ Another possibility is that greater barriers to entry have been able to be raised in the meantime, such as those noted below relating to access to points of supply (§3.4.5). Finally, opportunities to undercut line charges may have arisen from distortions caused by the regulatory regime, rather than from any inherently oligopolistic characteristic in the industry’s underlying cost structure. Nevertheless, the evidence suggests that, at least for a while, the *threat* of network competition—particularly transmission or distribution bypass—was sufficient to have the quasi-contestable effect of restricting line charges in New Zealand, at least for some consumers, if not for all.

3.3.3 “Unnatural” Competitors: the Case for Unsustainability and Greater Regulation

Yet this still leaves open the question: what actually made the observed bypass or subnetwork competition feasible in the first place? There are a number of possibilities that can be considered to provide the mechanism for the observed competition. Initially, assume that distribution network services do exhibit the cost characteristics of a natural monopoly, but only over a particular geographical area. As such, at some point, any economies of scale, scope and density are likely to become exhausted (§3.4.2). There are three feasible mechanisms for competition in such a naturally monopolistic distribution sector: (i) *rationalisation* of ELBs to their naturally monopolistic size and shape; (ii) *cross subsidies* in the line charge structure; and (iii) *unsustainability* due to inherent or regulated costs.

connection between a Transpower point of supply and a new large consumer—the Lichfield dairy plant—rather than a new connection to an existing consumer.

²³ For instance, in the mid-1990s, Bay of Plenty Electricity Ltd. constructed four embedded subnetworks (serving residential and light industrial consumers) within Mercury’s traditional franchise area. However, shortly after Bay of Plenty Electricity became Horizon Energy Distribution Ltd., it sold those subnetworks to Mercury (personal communication, Vector Ltd., 23 July, 2001).

²⁴ Irwin (2000) suggests that the light-handed regulatory regime has given ELBs “opportunities to experiment” with contracting out, mergers, pricing structures and other measures, in an attempt to determine what does or does not work. Irwin considers that, in the long run, these experiments will improve innovation and thus efficiencies.

The first of these mechanisms may arise because the old franchise boundaries do not equate with the boundaries which would truly make each ELB a natural monopoly. Not only might the boundaries be different, but the number of firms and the number of naturally monopolistic areas might not coincide. The Government, supported by some evidence (§3.4.3), felt that there were too many distributors, and therefore industry average costs were unnecessarily high. For England and Wales, for instance, at that time there were only twelve electric distribution companies (e.g., Kennedy School of Government Case Program, 1998). If there are too many distributors in New Zealand, given the customer base, the current process of ELB *rationalisation* could be partly explained as the industry finding its own “natural” size for efficient ownership of a distribution network. The number of ELBs has currently reduced to around half the original number of ESAs (Energy Markets Regulation Group, 2000d, p. 3). However, some ELBs in trust or local government ownership have been tenaciously holding onto their old areas, and their ownership structure makes competition by takeover infeasible (§2.4.2). Consequently, a logical explanation is that the only option open to potential entrants is to compete through bypass or subnetwork competition, and that this competition is feasible because the incumbent firms are generally small, with higher average costs.

Logical or not, this tidy explanation does not present the full picture. The competition described in the previous sub-section related to competitors entering the old franchise area of one of the largest ELBs (although being trust owned at the time, it was protected from takeover). Possibly the viability of entry was not due to higher average costs, but rather because the incumbent had not yet removed all the cross subsidies inherent in its pricing (i.e., line charge) structure. Allowing competition in ownership as well as for connecting consumers may place some contestable pressure on ELBs to move toward a productively efficient size as well as to offer allocatively efficient prices.²⁵

But another rather intriguing possibility exists however. As discussed earlier (§2.1.8), competition can exist even where a natural monopoly firm is productively efficient. This arises if the natural monopoly’s cost structure is *unsustainable*. In their treatise on contestability theory, Baumol and his colleagues present a model of intertemporal unsustainability, and make the disturbing conclusion that unsustainability can be the rule in growing markets where capacity construction costs are sunk and subject to increasing returns to scale (BPW, pp. 359 and 473). BPW’s concerns in this regard have been taken seriously by some of those providing public policy advice on network industries in the European Union (e.g., Heald, 1997). Furthermore, Teplitz-Sembitzky (1990, p. 47) concludes that BPW’s description of potentially unsustainable industries is applicable to electricity distribution networks. As

²⁵ Cross-subsidies may also exist in the transmission charge component of an ELB’s line charge. Different Transpower points of supply may have quite different “average” costs associated with them, but the ELB may average these costs across its consumers. A potential bypasser might be able to take advantage of Transpower points of supply with below average transmission charges to give it a quasi-competitive edge.

customers are lured away from the incumbent natural monopolist, its average costs increase, making it successively more vulnerable to additional entry, resulting in a vicious circle of unsustainability. Perhaps this suggests that ELBs need protection from inefficient competition, and that once rationalisation has reduced the number of ELBs to an “acceptable” level, the franchises should be re-imposed to avoid any wasteful duplication of facilities. In particular, BPW (p. 476) indicate that a source of unsustainability can be regulatory rules on depreciation policy and rate of return (Chapter VII). Perhaps the implicit rate of return regulation resulting from the application of the ODV methodology actually distorts ELB cost structures, and makes inefficient entry viable.²⁶

Regulatory restrictions may make prices unsustainable in theory, but what about in practice? On the one hand, the fact that significant erosion of ELB service areas resulting from inefficient entry is not being observed is not in itself evidence that unsustainable cost structures are not present. As Sharkey (1982a, p. 165) explains, the conclusion that a natural monopoly may be unsustainable is somewhat unsatisfactory, because it is not derived in an equilibrium context. Even if prices are unsustainable, the identification of successful entry strategies may be far from obvious (BPW, p. 10). But rather than pointing toward the need for more heavy-handed regulation as a result, this suggests that even if unsustainable line charges do actually exist in practice, they will not necessarily cause inefficient competition. As BPW (p. 217) express it: “unsustainability does *not per se* constitute a case for governmental restrictions on entry”. Baumol, in particular, seems to have become less concerned regarding unsustainability with the passage of time. For instance, in his monograph with Gregory Sidak on competition in local telephone markets, the possibility of unsustainability does not even warrant a single mention. Rather—with respect to telecommunications—Baumol and Sidak (1994a, p. 121) conclude that, “since we cannot be certain which arenas of local telephone service, if any, are natural monopolies, ... the most rational way to distinguish the arenas into which entry is feasible is to let the market decide”.

Some critics of contestability theory have dismissed the likelihood that unsustainability can exist in the real world, because they consider the perfectly contestable assumption that firms will be unable to use responsive pricing to deter entry to be unrealistic. One of the common criticisms of the concept of perfect contestability is that “prices are, in general, not likely to be more difficult to adjust than output” (Hazledine, 1992, p. 14), and that the Bertrand-Nash assumption underlying BPW’s models is not realistic (§2.1.7). Yet, New Zealand’s information disclosure regime acts to make ELB line charges relatively “sticky”, and difficult to adjust quickly in response to competition, at least in regard to existing consumers. ELBs must publicly disclose their line charge methodology, structure and price schedules.

²⁶ Some initial work on the possibility that the financial principles and valuation rules inherent in New Zealand’s regulatory regime might result in distorted ELB cost structures, and thus potentially unsustainable price structures, was performed by the present author (i.e., Gunn and Sharp, 1999).

Therefore, an external ELB, attempting to win away one or more consumers from the incumbent via distribution bypass, knows what the incumbent is charging that consumer. As long as the competitor can choose a line charge lower than the incumbent's, but sufficient to cover the costs of bypass, it can potentially steal away the consumer(s). The incumbent can only respond and change its price offering to that consumer by re-disclosing its entire line charge methodology and price schedule. To offer a lower price might require altering the existing methodology in a way that would appear inconsistent and/or inequitable to its remaining neighbouring consumers, and thus subject to challenge. By contrast, the external ELB is operating on a greenfields basis—its new line charge to the “stolen” consumer has no basis for comparison to any of the line charges it already offers to its traditional customer base. Even if the bypass relates to connecting a new rather than an existing consumer, disclosure causes the same problem of price comparability for the incumbent. Therefore, the prospect of inefficient competition for existing consumers arising from distribution bypass cannot be written off completely.²⁷

But competition for new consumers—particular through subnetwork competition (which always relates to the supply of new consumers—is somewhat different from competition for existing consumers. Such competition only restricts an ELB from growing (or reduces its network expansion rate), it does not eat into its existing customer base, possibly leading to the contraction and eventual unsustainability (and bankruptcy) of the incumbent. Furthermore, as noted above (§3.3.1), subnetwork competition does not require any duplication of assets, and therefore appears more desirable than competition for existing consumers via bypass. Panzar (1980) has developed a model of similar downstream entry into a vertically-integrated market (e.g., upstream and downstream electricity distribution), and found that such entry is implicitly more sustainable.²⁸ Intuitively, this result makes sense. Once an external ELB has won the right to connect the new subnetwork, the incumbent is now only able to respond through bypassing the new subnetwork entirely (or by offering to directly purchase the subnetwork from its competitor). Hence, the successful competing service provider is sustainable once it has constructed its subnetwork.

In fact, subnetwork competition, rather than appearing as a symptom of unsustainable natural monopoly, to some extent exhibits the key characteristics of a perfectly contestable market (§2.1.7). If the incumbent and entrant face the same line charges for connecting to the incumbent's network (and the incumbent does not raise barriers to entry through the terms of its interconnection agreement), then both firms have symmetric costs. Consequently, the barrier to entry associated with bypass—the sunk cost

²⁷ In discussing network industries, Heald (1997) also notes that “while technology and demand conditions may be changing rapidly, prices may be very sticky, probably because of legal and political constraints upon the scale of year-on-year tariff changes”.

²⁸ A similar model is developed by Burnell *et al.* (1999) in relation to telecommunication networks. However, this relates to partial bypass to compete for existing consumers.

involved in duplicating some of the incumbent's assets—is not present. Moreover, there are no barriers to exit for the incumbent; neither firm is yet actually serving the new market. Finally, the disclosure regime, combined with the transformation of the “new” consumers to “existing” consumers of the entrant (once the subnetwork is constructed), make it difficult for the incumbent to successfully engage in retaliatory price changes—thus providing a quasi-Bertrand-Nash environment.

3.3.4 “Unnatural” Monopolies: the Case for Less Regulation

Therefore, another possibility is that the assumption electricity distribution is a natural monopoly over a particular geographic area is completely invalid. Perhaps, using King and Maddock's (1996) taxonomy, ELBs are regulated monopolies or even fully competitive facilities. Such a viewpoint is taken to its logical extreme by the New Zealand Business Roundtable. Government intervention is seen as being responsible for the default monopoly structure for network services, and this would be eroded if the State simply stopped intervening in the industry.

If the industry were carefully analysed it would become apparent that the opportunities for competition and bypass of lines vary considerably across the system. ... It is likely that some of the activities of existing line companies are more contestable than others. ... In a more competitive environment we could expect lines to be bypassed on a piecemeal basis as the industry evolves. ... It might become economically feasible for cable TV, telephone, water, rail, roading or natural gas companies to lay cables that compete with local line businesses for electricity. ... The fact that they have not done so this century owes more to government ownership, regulation and control than to ‘natural monopoly’. Obviously, the incentive of private investors to devote resources to exploring such possibilities depends on their ability to enter the market and their assessment of the risk that the profits from any such investments might attract regulation (Kerr, 1999, pp. 5-6).

In sum, between the extreme perspectives of electricity distribution as an essential facility versus that of a competitive facility, lie many possibilities each which prescribe various degrees of regulatory intervention. Perhaps rather than being *geographically* specific, the naturally monopoly characteristics of distribution might be more *functionally* (or consumer class) specific. If so, this could cause some problems for designing effective regulation, since if only some parts of an ELB's operations are contestable, an ELB could attempt to compete by cross subsidising its contestable activities from the “captured” base of customers connected to its natural monopoly facilities. The next section examines the naturally monopolistic characteristics of power distribution networks in more detail.

3.4 Natural and Regulated Monopoly Characteristics of Distribution Networks

3.4.1 Number of Distribution Firms

The discussion in the previous section has highlighted that the assumption electricity distribution is a natural monopoly cannot be taken for granted. But determining whether a particular industry is

naturally monopolistic or not requires a detailed analysis of that industry's underlying cost structure, rather than of any evidence of competitive behaviour, since such observed competition could potentially be inefficient and/or short-lived. However, before examining the costs inherent in the provision of distribution network services in New Zealand, it is worth pointing out that, for an industry largely labelled as a natural monopoly, even post-deregulation there still seem to be a large number of market participants.

The usual objection to allowing competition for conveyance and network connection is that it would result in the wasteful duplication of facilities (§3.2.1). But this really only suggests that more than a single *network* is uneconomic, as opposed to a single *firm*. Surprisingly, a distinction between the number of firms and the number of networks serving consumer demand for network connection is not often found in the literature. For although electricity distribution is usually labelled a natural monopoly, even a cursory examination of most electricity supply industries worldwide would reveal that usually a multiplicity of firms exist to distribute electricity from bulk transmission level down to the end-consumer. Moreover, separate markets already exist for some functions or activities associated with the provision of network services, in particular, network construction (§3.2.3).

Firstly, more than one firm tends to exist 'horizontally' in the entire distribution market, serving different geographical areas, with little if any overlap. However, since electricity conveyance to a specific location can be considered a distinct product, a geographical sub-market could be defined within which the supply of all products—the connection of all consumers to the network—is most inexpensively performed by a single firm. The notion of distributors being *regional* natural monopolies is not an uncommon one. Nevertheless, the question remains, at what point do any economies in the network run their course over a particular geographical area?

Secondly, even within a particular region, more than a single firm might be involved in providing distribution services 'vertically'. For example, a distribution company might supply a large consumer such as an airport, and the airport management company in turn supply 'subconsumers' within its own network.²⁹ Even households own the majority of their own wiring. In a truly radial network, electricity conveyance via each successive segment in the network from the point of supply to an end consumer, past various branches in the network, is dependent on the previous segment of the network. The question remains, at what point from transmission network down to final appliance, rather than to the consumer, are all economies exhausted? One view is that the ownership of dedicated assets (§3.1.1)—those which can be explicitly identified as being associated with a particular consumer (or groups of consumers)—can

²⁹ Power distribution networks at major airports within New Zealand are both owned by the airport management company itself, such as in Auckland, and by external utility companies, such as Infratil for Wellington International Airport (NZPA, 2001b).

be owned by the consumers or the distributor without any loss in efficiency. If such is the case, then it could be argued that the same would apply if an external distributor, rather than the consumer, owned the assets.

As David and Li (1991a) phrase it, optimal decision-making in electricity supply involves three related questions, pricing, investment and control (or co-ordination). Although these issues are more complex in relation to operating a transmission network, issues of co-ordination could arise at distribution level if more than two firms are associated with the distribution services provided by a single network. One question would be, how is the cost of backup supply (i.e., electricity conveyance) allocated between the two or more firms? Nevertheless, economically allocating the costs of assets to consumers which they seldom use might also be a complex problem even were the network owned by a single firm. From a natural monopoly standpoint the question is: do the costs of conveyance increase due to the transaction costs (§2.1.6) incurred as a result of information transfer between the two firms?

Mercury Energy Lines Business (1999, p. 16) suggests that, if the co-ordination has naturally monopolistic characteristics, then that function could be separated out (§3.2.3). Mercury also provides one reason why two vertical firms may in some circumstances be more efficient than a single ELB. Apartment and industrial park owners are cited as having “site specific knowledge” in respect to what is involved in distributing electricity within the confines of their own complex.³⁰ In any event, such a vertical arrangement between two firms does not involve any duplication of network assets.

3.4.2 *Distribution Network Economies and the Significance of Common or Joint Costs*

The presence of numerous entities involved in electricity distribution is by no means necessarily least cost; the suggestion that any observed system must be optimal simply because institutions are assumed to be transaction cost minimising, is somewhat of a “fallacy of consequence” (e.g., Gale, 1989). Rather than examining the actual number of firms, evidence for the existence of natural monopoly has traditionally been provided by an assessment whether the industry in question exhibits ‘*economies of scale*’ (e.g., Teplitz-Sembitzky, 1990, p. 4). Even these days, the assertion that declining average costs and natural monopoly are equivalent can be found in some of the debates concerning power sector reform in New Zealand. For instance, Mercury (p. 13) states that: “a natural monopoly will exist if the cost of

³⁰ Mercury (p. 16) also cite this as evidence that there is an absence of economies of scale at this level of this business. How this conclusion relates to a vertical arrangement is unclear, since the question is the number of firms, and not the scale of operations. The installed capacity could be identical whether the complex owner or the ELB owned and operated the subnetwork.

constructing each component of the reticulation assets shows decreasing average costs for a typical network configuration”.³¹

In their most general sense, economies of scale relate to declining average (or unit) costs as output increases. For example, Wyatt *et al.* (1989, pp. 8-10) state that a number of technical and organisational factors influence economies of scale in electricity distribution, and that these arise from characteristics of distribution equipment which lead to economies of density, economies in capacity expansion, and economies in the provision of capacity to meet peak power demand requirements. Consequently, economies of scale are associated with decreasing unit costs for any increase of output, irrespective of the reason. However, a narrower definition of economies of scale arises when output is increased, but all other factors, such as the number of products, are held constant. Other economies are found by holding output constant, but changing other parameters, such as the number of products, or the number of consumers.

Of particular importance are the economies which result from the production of a greater number of products. These are termed ‘*economies of scope*’, and occur when the cost of supplying a portfolio of products is less than the sum of the costs of supplying each product separately. (For instance, costs may be lower because demand risk is lower; §9.4.5). The term ‘economies of scope’ was originally coined by Panzar and Willig (1981), key contributors to contestability theory. Panzar and Willig link economies of scope to the existence of ‘shared inputs’, that is “inputs which, once procured for the production of one output, would be also available (either wholly or in part) to aid in the production of other outputs”. Shared inputs are caused by ‘*common costs*’ or ‘*joint costs*’. Common costs occur when a single product is provided to two or more users, whereas joint costs occur when costs are a ‘non-separable’ function of two or more outputs (e.g., Kask, 1988b, p. 20).³² Since, electricity distribution can be considered to be a multiproduct industry where every user consumes a different product, joint costs are the appropriate term to use with respect to electricity distribution shared inputs. Panzar and Willig provide specific examples of such shared inputs, which are seen as occurring for “power transmission” facilities, and any ‘*indivisible*’ item of equipment able to contribute to the production of more than one product.

Both economies of scale (e.g., Williamson and Mumssen, 2000, p. 2) and economies of scope from the production of complementary goods (e.g., Evans and Quigley, 1998) still remain the most often cited characteristics of New Zealand’s network industries in general. But for electricity distribution

³¹ And Evans and Quigley (1998): “the natural monopoly property means that minimum average cost in the activity is achieved at level of output greater than or equal to the whole market”. In fact, minimum average cost could also occur at a level of output less than the entire market, and the firm still be a natural monopoly (e.g., Sharkey, 1982a).

³² Nevertheless, this definition is not entirely standardised. For instance, Heald (1994, p. 2) defines joint cost in the manner Kask describes common cost, and vice versa.

networks, it is now becoming recognised that the greatest economies are likely to result from the latter (e.g., Weyman-Jones, 1995). As Sharkey (1982a, p. 23) expresses it: “it is the network configuration rather than the network size that is relevant in the determination of natural monopoly characteristics in the industry”. Scale economies are more likely to be found in individual items of distribution equipment, since for many items, particularly transformers, an increase in equipment capacity results in a less than proportionate rise in equipment cost (§3.6.1).

Excess or ‘*spare capacity*’ is also cited as a source of economies in distribution networks. For instance, Wyatt describes joint production economies which exist between consumers who use electricity at different time periods. Effectively, if one consumer’s demand for peak power is not constant over time, then spare capacity exists at a later period to serve another consumer’s demand. Wyatt also cites excess capacity as causing “economies of fill”, which Sharkey (1982a, p. 79) defines as a short run concept. Electricity distribution can only be expanded in discrete steps, and is a lumpy investment made of indivisible equipment sizes. This results in declining unit costs in the short run until demand for capacity reaches the actual level of installed capacity, and the cost of connecting new customers is therefore low. However, Wyatt also attributes economies of fill to the situation where capacity is installed in excess of current demand, due to an expectation of future expansion of demand, rather than simply due to indivisibilities. As is discussed later (§6.1.1), there is an important distinction between: spare capacity caused by the indivisibilities of lumpy equipment; anticipatory construction in expectation of future demand; and redundancy required for improved reliability purposes.

Hay and Morris (1993, p. 37) present a taxonomy of economies of scope sufficient to cover all these various types of economies. They identify three forms. The first arises where some factors of production are “public”, in the sense that once they have been acquired for the use in producing one good, they are costlessly available for the use of production in others. In electricity distribution such economies are generally considered to arise because consumers have peak power requirements at different times. Consequently, the total installed capacity of equipment in the semi-individual network is less than the total connection capacity of all dedicated assets, and similarly the capacity of the semi-individual network is less than that of the collective network, as a result of diversity (§3.1.1). This might suggest that distribution networks experience significant ‘*intertemporal economies of scope*’ (e.g., Crew *et al.*, 1995). However, if an electricity line business is considered to sell connection capacity (§3.1.1), then, as long as the consumer’s peak power requirements do not change, its demand for capacity remains constant. Dedicated assets therefore only exhibit intertemporal economies of scope when the consumer’s demand for capacity lasts for a shorter period than the lifetime of those assets.

The second type of economies of scope arises from Panzar and Willig’s “shared inputs”, and Hay and Morris also encompass shared inputs to mean the case where spare capacity is available due to equipment indivisibilities. Both indivisibilities and shared inputs occur in distribution networks.

Distribution equipment is manufactured in discrete sizes, and this results in indivisibilities at all input levels throughout the network. Diversity allows inputs to be shared at semi-individual and collective network input levels. However, since diversity means that required capacity of these parts of the network is less than would otherwise be the case, scale economies at these input levels are as not as greatly realised as they would otherwise be.

The final economy of scope arises from ‘cost complementarities’. Complementary goods are those which tend to be consumed together, and Hay and Morris state that cost complementarities occur when the marginal cost of producing one product falls as the output of another increases (§4.1.1). Network and energy services are clearly complementary goods, and they also exhibit cost complementarities. This is because larger capacity distribution equipment yields lower energy costs because higher voltage and/or lower resistance operation reduces system energy losses (§3.1.1). Although this is sometimes cited as an economy of scale due to equipment size (e.g., Wyatt *et al.*, p. 8), under Hay and Morris’s taxonomy, it would be considered an economy of scope in outputs (realised through an economy of scale in inputs). However, since their definition of cost complementarities is based on marginal costs, it seems unlikely that the distribution network on its own exhibits complementarities between connection capacity at different locations.

Apart from economies of scale and scope, the literature also mentions a few other forms of economies potentially relevant to electricity distribution. In concluding that it would be inefficient to serve the same geographic area with more than one distribution network, Joskow and Schmalensee (1983, p. 59) give recourse primarily to the notion of ‘*economies of density*’ (and similarly in New Zealand: Wyatt *et al.*, 1989; Culy *et al.*, 1997). These are defined as exhibiting declining unit costs while consumer density grows, *ceteris paribus* (e.g., Roberts, 1986; Vogelsang, 1994). Consumer density is typically defined using some measure of ‘network utilisation’, such as the installed connection capacity (kVA)—or the number of consumers—divided through by the total length of lines in the distribution network. Nevertheless, a contrary view is provided by Mercury (1999, p. 13). Mercury suggests that economies of density relate to the interconnection of nodes with different demand profiles, and that these economies are not an indication of natural monopoly. Mercury’s argument is presented on the basis that the interconnection of different nodes with different production profiles, as “applies to generation and the development of spot markets”, demonstrates that “common ownership is not a necessary condition” for gains from economies of density to be obtained.³³

³³ The equivalence of generation with distribution made here seems spurious, since the economies of density are realised through the interconnection of the production nodes via the *transmission network*, not through the operation of the spot market, and these economies are responsible for the usual assumption that transmission is the “classic” case of a natural monopoly (e.g., Weiss, 1975, p. 144). Interestingly, Vogelsang (1994) suggests that electricity distribution has *greater* characteristics of a natural monopoly than power transmission. He attributes distribution networks with strong economies of

Wyatt and his colleagues also discuss a characteristic associated with economies of density, that of improved security of supply. Since increased reliability occurs as a result of a larger network, lesser investment is required in assets specifically required to improve security. Such benefits could be termed ‘economies of reliability’. Supply security improves (and demand risk reduces; §9.4.5) when consumer density rises, since several radially-operated subnetworks can be interconnected with open switches to provide different paths for electricity to flow through in the event of a fault (§3.1.2). The property of a network to have greater value as its interconnectedness increases is often termed a ‘network externality’ (e.g., Evans and Quigley, 1998).³⁴ However, Wyatt and his colleagues suggest that, in New Zealand’s case, the benefits of these externalities are exhausted by the requirement to keep the networks supplied from each Transpower point of supply separate. This is an operational limitation implemented to keep any fault current below a level that might cause damage to network equipment. As such, any economies between neighbouring subnetworks connected back to different Transpower points of supply seem likely to be fairly weak. Possibly the limit to how large a subnetwork becomes before the marginal improvement in reliability from greater interconnectedness is less than consumer marginal willingness-to-pay for that improvement, is even smaller than the subnetwork connected to a point of supply.³⁵

The literature also describes ‘economies of sequence’ (e.g., Spulber, 1989, p. 113), which relate to productive efficiencies achieved through vertical integration. Of course, the electricity supply industry has been moving away from vertical integration, although the recent reintegration of generation and retailing functions in New Zealand may suggest that economies of sequence can be realised between these two functions.³⁶ However, the concept is important with respect to subnetwork competition in

density. Although acknowledging that transmission networks exhibit economies of scope, Vogelsang suggests that “it is quite conceivable that an interconnected transmission network could consist of several legally and commercially independent but interconnected parts. ... Thus, the argument against competitive and in favor of regulated monopoly supply is somewhat weaker in the case of transmission than it is in the case of distribution networks”.

³⁴ Spulber (1989, p. 54) would likely refer to such improved reliability to be an ‘internality’ rather than an externality, given that the benefit is endogenous, resulting from the market transactions themselves (§5.3.3).

³⁵ For instance, prior to reform (§2.2.3), it was not uncommon for an MED franchise of a small city or large town to be surrounded by a larger EPB. Since the MED boundary was based on the traditional local authority boundary (which stayed constant), as the city expanded new suburbs became supplied by the EPB instead. Once the franchise boundaries were removed, a common source of amalgamation was between such ESAs, particularly if they were both connected back to the same Transpower point of supply. One might expect that after amalgamation, lines might have immediately been constructed to interconnect the previously artificially separated distributors. The present author was involved in one such study for the amalgamated Waikato Energy Ltd. serving the city of Hamilton and surrounding areas (see Preface and Acknowledgements). However, although some interconnection was investigated, the key justification for the interconnection was not improved *reliability*, but the reduction in energy *losses* that would be possible as a result of reconfiguring operation after the construction of the new lines.

³⁶ Doubts have been recently raised in New Zealand as to whether a pure retailer can even survive in a market where vertical integration is allowed between generation and retailing. For instance, the electricity retailing company of the Natural Gas

particular, since such a mode of competition breaks any economies that may exist between upstream and downstream distribution assets.

The degree to which the various forms of economies will be realised in a specific distribution network will strongly depend on the nature of demand growth (§3.1.2). Where gradual demand growth only requires minor augmentation of the existing system, substantial economies can be experienced should network capacity be optimally sized *in anticipation* of long term future demand (§6.1.1). These economies will, however, be exhausted once demand reached the limits of previously installed network capacity. Investment decisions can then appear globally sub-optimal, even though they are cost effective (§6.1.4). On the other hand, rapid demand growth at the fringes of an established network will allow a greenfields design to be implemented; one that can be globally optimal, at least in the medium term. However, there will be a trade-off between any positive economies in the demand for capacity—which arise from the multi-product and diversified nature of the demand for electrical energy—and the potentially negative economies associated with the distance of new consumers from existing points of supply.

3.4.3 Studies of Electricity Distribution Costs in New Zealand

In engaging in the reforms to the distribution sector, the New Zealand Government was particularly concerned about the number of distribution/retailing firms—the ESAs (§2.2.3). This concern was underwritten by a study prepared for the Ministry of Energy (subsequently Commerce) in the late 1980s which had estimated, and then examined, the supply industry’s average cost curve. This study, the already-referenced report of Wyatt *et al.* (1989)—henceforth referred to as ‘Wyatt’—assumed that the outcome of any future reform would be the amalgamation of existing franchised monopolies into a smaller number of natural monopolies, each operating as such within a particular geographical market. Examining economies of scale, to determine how large the market size of each firm would need to be for this assumption to hold, the study found the minimum of the distribution sector’s average cost curve, where costs were averages in terms of energy sales (kWh). The study noted that the U-shaped industry cost curve which they fitted to the ESA cost data was consistent with similar studies of electricity distribution which had been performed in the US (e.g., Neuberg, 1977). The minimum of the curve, although it remained flat over a large range of output, lay at around one-eighth to one-eleventh of entire market demand for ESA energy sales. Therefore, the conclusion was that the industry structure, as it stood during 1987, should reduce to between 8 and 11 companies to gain the optimal advantage from economies of scale. The implication was that the industry as a whole was *not* a natural monopoly.

Corporation, which has no generation assets, has been described by industry analysts as “cancerous” (NZPA, 2001a). In such a market structure the pure retailer may be forced to purchase wholesale electricity from a generator that also owns a competing retailer.

However, implicitly distributors were still perceived as natural monopolies within the confines of a particular geographical area.³⁷

A number of problems can be identified with the approach taken in the study. Firstly, Wyatt (pp. 2 and 15) failed to identify a group of ESAs which performed consistently better than the others, finding no correlation between the performance of an ESA and its size (either in terms of electricity conveyed and sold, or in terms of consumer numbers), although some relationship between performance and consumer density was found. One of Wyatt's basic assumptions (p. 60) was that ESAs were cost minimisers, rather than profit maximisers, and the problem with the industry was that, *as a whole*, it was not cost minimising. However, if the industry as a whole suffered from productive inefficiencies, then perhaps individual ESAs also did. For instance, empirical studies of municipally-owned power distributors in the US would suggest that such companies are not cost minimising (e.g., Claggett *et al.*, 1995). Therefore, the lack of correlation in the data used by Wyatt might be due to uncorrelatable levels of inefficiency. If so, then the derived long run average cost (LRAC) curve for the industry is not the true one, since the LRAC curve is traditionally derived by assuming that the industry expands along an optimal path (e.g., Hay and Morris, 1993, p. 28). Moreover, the observed declining average costs may not necessarily be due to economies of scale, but to other economies (§3.4.2).³⁸ In any event, as will be discussed in the next sub-section (§3.4.4), the analysis of economies of scale and the average cost curve is *not* the appropriate test for natural monopoly.

Secondly, the conclusions which the study draws for *unbundled* distributors may be limited. Apart from the fact that the study treated the industry as providing a single product (i.e., bundled network and energy services), meaning that any "economies" related to the two aggregated activities in combination, energy sales is not the appropriate measure of a unbundled distributor's product (§3.1.1). Nevertheless, Wyatt did make an attempt to compensate for product bundling by also performing a 'multi-output' analysis, involving two products, power (kW) and energy (kWh), although this attempt

³⁷ Wyatt (p. 8) references both Sharkey's and BPW's work on natural monopoly, and made it clear that it was assumed distribution was a naturally monopolistic activity within a specific geographic area: "The existence of natural monopoly in electricity distribution would tell us that only one firm will operate in any area. The present study assumes this outcome, and so does not test for natural monopoly. Knowledge of the economies of scale determines how large that area should be, and hence the optimal size for an ESA". Wyatt's report was based on modelling work presented in Giles and Wyatt (1989), and the research was later published in the academic literature as Giles and Wyatt (1992).

³⁸ For instance, in discussing Wyatt's work, Evans (1998) cites the study by Salvanes and Tjotta (1994) of Norwegian electricity distributors. Salvanes and Tjotta concluded that distribution costs are substantially affected by economies of density (§3.4.2), and found that when these densities are accounted for, economies of scale vanish at all but very small output levels.

was hindered by inadequate data.³⁹ Nevertheless, network connection capacity was not considered as one of the products in this approach either.

Finally, the study highlights some of the problems inherent in using econometric techniques on limited, and highly scattered, historical data. The “costs” used in the study, although excluding capital expenditure, included depreciation, and primarily related to only a single year. Given the long lifetime of network investments, it is unlikely that a time slice based on historical depreciation methods could adequately indicate the true economic costs of bundled electricity conveyance and supply (e.g., Hay and Morris, 1993, p. 49; §7.2). Wyatt and his colleagues (p. 26) themselves pointed out the drawbacks of dealing with data for just a single year.⁴⁰

Wyatt’s study examined distribution costs on an aggregate basis. By contrast one of the ELBs themselves—in the discussion paper by Mercury already-referenced throughout this Chapter—takes a “bottom up” approach to costs, and uses this to conclude that electricity conveyance across a typical 11 kV (i.e., MV) network is *not* a naturally monopolistic activity (Mercury, p. 14). This claim warrants some attention, since it is one of the very few studies to attempt to address the question of distribution natural monopoly using real cost of equipment data.

Figure 3.2 presents what Mercury terms the “average incremental cost curve” for a representative 11 kV distribution network, where each homogeneous medium-consumption consumer (or group of consumers) requires 1.5 MVA of capacity.⁴¹ Hence, the *x*-axis could be readily converted from number of consumers to capacity, without any change of scale. Although each consumer is actually utilising a different product (because each network connection is at a different location), all consumer products are measured in the same units (i.e., kVA), and therefore average cost relates to the average cost across a number of similar products.

³⁹ The study looked at two products of: (i) ‘energy’ in kWh, and ‘power’ in kW; (ii) industrial and non-industrial bundled demand; and (iii) residential and non-residential bundled demand. Wyatt *et al.* (1989, p. 53) concluded that since kWh could not be provided with kW, the two “products” functions were not really separable, since one could not be provided without the other. Nevertheless, in reality, network costs are not most closely associated with energy or power, but with ‘capacity’, in kVA, as is discussed above (§3.1.1).

⁴⁰ More recent studies include Gale and Strong (1999), and Strong and Gale (1999).

⁴¹ Each consumer is also considered to possibly be a consumer substation feeding an LV network of small commercial and residential consumers, rather than a medium consumption consumer taking supply directly at MV. Hence, consumers are only homogeneous with respect to their capacity requirements, and not necessarily to their kWh consumption. Note that the *y*-axis is labelled “average cost” rather than the term Mercury uses, “average incremental cost”. The reasons for this are explained shortly.

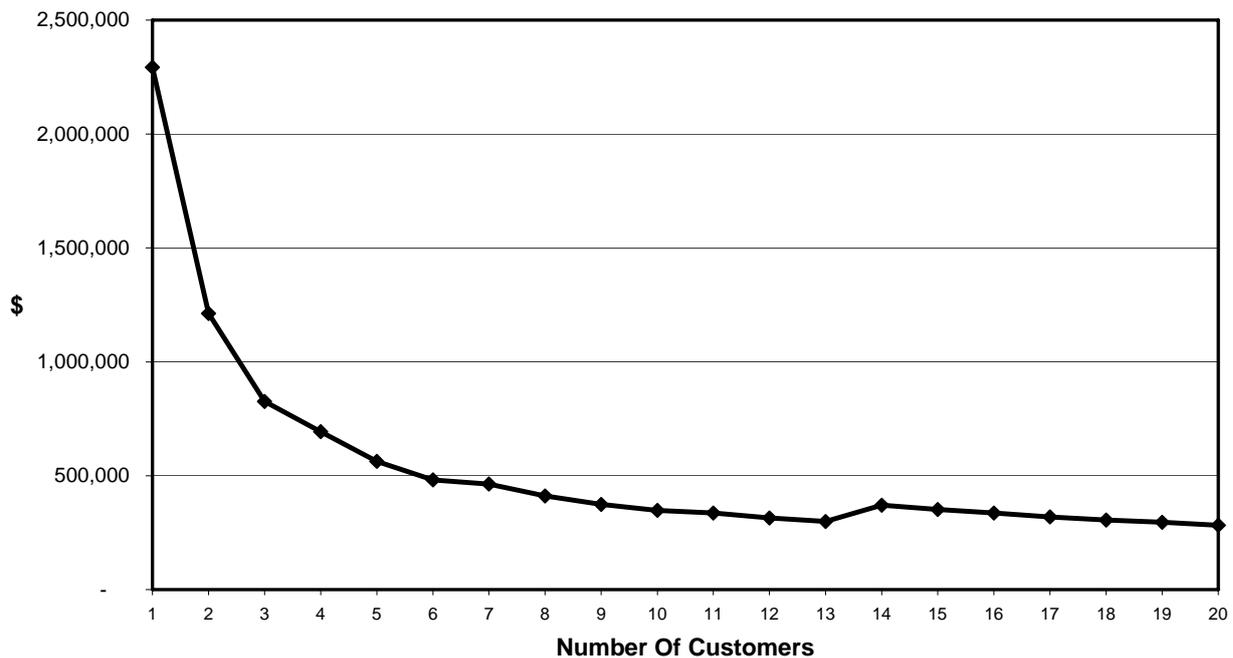


Figure 3.2: 11kV “Average Incremental Cost” Curve

Mercury state that: “if the 11 kV network is a natural monopoly, the incremental average cost could be expected to fall across the entire range of output. If it is not a natural monopoly, average incremental cost curves can be expected to flatten, or even begin to rise after a certain portion of consumption has been met”. On the basis of this assumption, and because the graph shows that “average incremental costs of capacity fall relatively sharply, flatten, and then rise again slightly [at the 14th consumer] as additional investment in transformers is required to service increased demand”, it is concluded that “the 11 kV network does not exhibit natural monopoly cost characteristics”. Provision of 11 kV network services is thus likened to any other capital intensive industry, where some “minimum level of market penetration” is necessary to recoup the cost of capital.⁴² Although the level of investment required to enter the market is demonstrated to be significant, it is stated that this in itself is not a barrier to entry in an economic sense (Mercury, p. 14).⁴³

⁴² Mercury’s concept of a minimum level of market penetration is similar to the traditional industrial indicator of ‘minimum efficient scale’ (MES), attributable to Bain (1954). The MES is the level of plant production at which costs become constant, and further economies of scale are negligible. When considered in the context of total market demand, the MES becomes the ‘minimum efficient market share’ (MEMS) required for a firm to survive in the market (e.g., Miller, 1995). Consequently, MEMS is a traditional measure of the competitiveness of a particular market.

⁴³ Mercury also examined LV networks and came up with a similar conclusion, again because the average “incremental” cost curve included an upturn in average cost at the 25th consumer.

This conclusion, and the analysis itself, have a number of shortcomings. Firstly, as discussed below, the absence of declining average costs over the entire cost curve is not sufficient to demonstrate the absence of natural monopoly. In fact, in the range of 14 consumers, the average cost curve provides a classic example of an unsustainable single product natural monopoly (§3.3.3). Secondly, the transformer configurations appear not to be optimally sized to the demand for capacity.

Finally the graph is not of average *incremental* cost, nor even long run average cost. The graph does involve long run costs, since short run average cost curves relate to situations where capacity is fixed (§4.1.5). Mercury's graph is of the average cost of constructing capacity, so by definition, it relates to long run costs. Nevertheless, it presents the cost of constructing a particular level of capacity, and does not reflect the cost of expanding capacity once some initial capacity has already been installed. Consequently, the graph is not a true LRAC curve, since as mentioned above, it would have to account for the optimal expansion path.

The cost data is the total average cost of connecting, on a greenfields basis (§3.2.1), the stated number of consumers. But network capacity is not perfectly adjustable, because it is not perfectly fungible, and therefore costs are intertemporally interdependent, as is discussed shortly (§3.5.1). Should the original subnetwork be constructed to connect say nine consumers, assuming that the investment is non-fungible (i.e., it has no value in any alternative use), the incremental cost of adding the tenth consumer at some later date is actually the average cost shown on the graph of constructing an MV network for a single consumer. Hence, the overall LRAC is much higher than that indicated in Figure 3.2 for ten consumers. (In fact, an analysis of actual long-run incremental costs actually adds weight to Mercury's argument, as is shown later in Figure 6.1).

3.4.4 *Appropriate Tests of Natural Monopoly under Contestability Theory*

New Zealand's power sector reform was to a large extent driven and justified by contestability theory (§2.1.6). However, the theory's recommendations as to the appropriate tests for natural monopoly has not stopped industry analysts in New Zealand from attempting to determine the optimal distribution company "size" by examining industry *average* costs (both in terms of energy sales and numbers of consumers). But as Baumol and his colleagues explain: "A multiproduct cost function possesses no natural scalar quantity over which costs may be 'averaged'. That is, we cannot construct a measure of the magnitude of multiproduct output without committing the sin of adding apples and oranges" (BPW, p. 47). Further, Hay and Morris (1993, p. 57) instruct that: "there can be no excuse for regarding scale as the sole significant or systematic determinant of firm costs".⁴⁴

⁴⁴ Yet this recognition took some time to filter through into the literature, which may explain why the concept is still misunderstood by pricing practitioners. For example, even Blaug (1990), in discussing marginal cost pricing (§4.1.1), states that: "If there really are 'natural monopolies'—that is, public enterprises in which costs continue to decline monotonically

As long ago as 1977, prior to its integration into contestability theory, William Baumol demonstrated that scale economies are neither necessary nor sufficient for monopoly to be the least costly form of productive organisation (Baumol, 1977).⁴⁵ Rather, the critical concept is strict ‘subadditivity’ of the cost function, meaning that the cost of the sum of any specified number of output vectors is less than the sum of the costs of producing them separately. Proving subadditivity requires testing every possible combination of output vectors, a task that becomes increasingly more difficult as the number of products increases. As BPW (p. 170) later expressed it: “the intuitive appeal of the subadditivity concept is counterbalanced by its analytic elusiveness”. For a firm providing a set of products $N = \{1, 2, \dots, n\}$, with price and output vectors $p = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_n)$, and a technology described by the multiproduct cost function $C(Q)$, subadditivity of the cost function requires (3.1) to hold, for every and all product sets.

$$C\left(\sum_{j=1}^k Q_j\right) < \sum_{j=1}^k C(Q_j); \quad \exists \forall Q_1, \dots, Q_k \quad (3.1)^{46}$$

Baumol (1977) demonstrated that, even for a *single* product firm, evidence of scale economies is always *sufficient* but *not necessary* to prove subadditivity. Conversely, increasing average costs are not sufficient evidence that a single product natural monopoly does not exist. However, Baumol noted that, even in the single product case, the analysis makes it “harder to prove that a particular monopoly is natural”, because “proof of subadditivity requires a *global* description of the shape of the *entire* cost function from the origin up to the output in question, thus calling for data that may lie well beyond the range of recorded experience”.

For a *multiproduct* firm, the notion of declining average costs becomes problematic, because the concept of scale economies requires input to change proportionately (meaning that the production function is ‘homogeneous’), and in general this will not minimise the cost of expansion. On the other hand, if outputs do not expand proportionately, then an index of aggregate output by which total cost can be divided cannot be determined. Neither is there a way of apportioning joint and common costs in order to calculate a single product’s average cost. However, Baumol found that sufficient conditions for subadditivity must include some sort of ‘complementarity’ in the production of the different outputs, a

for all levels of output—it is of little help to be told that short-run marginal costs will be equal to long-run marginal costs when capacity is optimally adjusted, because the optimum level of capacity of ‘natural monopolies’ is infinitely large” (§4.1.3). As explained earlier (§2.1.5), the optimum capacity level for a natural monopoly is not necessarily “infinitely large”.

⁴⁵ Baumol’s (1977) cost test assumes that input prices are constant, and therefore an industry may be a natural monopoly for one set of input prices, but not for another. Baumol, however, considered that it would be far too restrictive to require that the definition of a natural monopoly relate to anything other than current input prices.

⁴⁶ This particular definition of subadditivity is taken from Faulhaber and Levinson (1981).

concept which he described as being similar to economies of scale. However, like economies of scale, economies of scope do not in themselves necessarily imply subadditivity of the cost function (BPW, p. 173).

The earlier literature on the cost characteristics of multiproduct industries was drawn together in the standard text on contestability theory by Baumol with Panzar and Willig, in order to establish the appropriate technological conditions for natural monopoly (BPW, pp. 169-189). Baumol and his colleagues demonstrated that for any firm to be in equilibrium in a contestable market, then that firm has to be a natural monopoly with respect to *its own output level*. This result holds whether the industry is a natural monopoly, perfectly competitive, or is oligopolistic. Subadditivity thus has profound implications for the existence of equilibrium and for the optimal number of firms serving any market. Unfortunately, BPW (p. 170) also found that there are no conditions that are both necessary and sufficient for subadditivity which are analytically simpler than the definition. Instead, Baumol and his colleagues derived a series of necessary tests for subadditivity, and some sufficient tests for subadditivity, based on the multi-dimensional geometry of the cost function.

3.4.5 Barriers to Entry for Distribution Networks in New Zealand

If an analysis of distribution costs is unable to demonstrate that ELBs are (regional) natural monopolies, then it is possible that they are regulated monopolies (§3.2.1), able to retain their market dominance through historic reasons. The possibility that such pre-existing market power allows incumbent regional regulated monopolists to raise barriers to entry through the terms and conditions of interconnection agreements, has been mentioned above (§3.2.3). Another barrier to entry that can be raised, or may already exist without an incumbent needing to take any action, is either no spare capacity or insufficient space for a new distribution feeder (i.e., line) at the local Transpower point of supply. This could prevent either transmission bypass or distribution bypass. Even if the transformers at the Transpower substation have available capacity given the local security criterion, there may not necessarily be a spare transformer bay for a new feeder. If there is not, then the potential entrant would have to pay Transpower for the costs of expanding its point of supply, and these costs may be sufficient to make entry uneconomic. A canny incumbent could raise this barrier to entry by sub-optimally utilising (or perhaps reserving for future network expansion, via its contract with Transpower) any spare transformer bays. The competing ELB could of course cite such behaviour as being anti-competitive under the *Commerce Act 1986*. On the other hand, the incumbent might be able to successfully use an “efficiency defence” to justify its actions (§2.1.3).

Although pre-existing dominance may allow ELBs to raise barriers to entry, another possibility is that barriers to entry themselves contribute to the regulated monopoly status of ELBs. The concept of barriers to entry is a relatively old one, and the sources of such barriers were traditionally listed as (i) “product differentiation”, particularly in regard to the preferences of consumers for existing products;

(ii) “absolute cost advantages”; and (iii) “economies of scale” (Hay and Morris, 1993, p. 86). One of the most widely cited definitions of a barrier to entry is that of Stigler (1968, p. 67), who defined an entry barrier as any arrangement that imposes a cost upon an entrant from which an incumbent is immune.

As part of his wider critique of contestability theory, Shepherd (1995) states that, in general, entry barriers are a more formidable and widespread problem than the “Baumol group” acknowledge. Shepherd—like another critic of contestability theory, Miller (1995)—suggests that the central and most general cause is the customer base held by an incumbent firm; in other words, product differentiation. Not only might the incumbent defend this fiercely, but many consumers are also “inert” or “habitual”, and are not quick to switch suppliers. Even where consumers are willing to change, the “switching costs” involved in changing electricity suppliers may be relatively high for smaller consumers (e.g., York, 1994). Conversely, from the supplier’s perspective, Lowry and Kaufmann (1998, p. 6) suggest that incumbent electricity distributors can experience “economies of scope” associated with their “core capabilities”. Such capabilities stem from a history of providing reliable distribution services to specific consumer groups, and may provide special insights into consumer needs. This knowledge may be “tacit”, and difficult for potential competitors to reproduce (an argument which has parallels to Mercury’s observation regarding “site specific knowledge”; §3.4.1).

Absolute cost advantages may have some role to play in electricity distribution, given that demand patterns, and thus the least cost network configuration, change over time. But given that assets are long-lived, this actually works against the incumbent firm, and for the potential entrant. Firstly, in constructing a new bypass network to connect a subset of consumers, the potential entrant may be able to identify a lower cost connection option, particularly through a cheaper route. Secondly, the existing network may have local security standards above those for which the consumers are willing-to-pay, and the entrant can provide a less expensive connection based on a lower target level of reliability. Finally, depending on the resale markets for distribution equipment, an entrant may be able to source cheaper second-hand equipment, having a remaining lifetime longer than the period of existing consumer demand. Such a competitor could be able to engage in a form of medium term ‘hit and run’ entry. On the other hand, an incumbent may have an absolute cost advantage caused by regulatory restrictions which apply to new entrants, but not to the incumbent. Two key areas where regulations may increase the costs to entrants are local government requirements for any new power lines to be trenched underground, and restrictions on rights-of-way for new line routes.

Miller (1995) is only one of many authors who describes the key structural barrier to entry into any network industry as being the high threshold level of investment needed for entry into the incumbent’s market. Since network industries are usually deemed to exhibit economies of scale (as well as scope and density), the incumbent is likely to face lower average costs than possible entrants. Miller suggests that these factors provide incumbent firms with “significant strategic maneuverability that

permits market dominance”. However, for New Zealand’s ELBs, any such advantages would only help the incumbent firm with respect to fending off bypass competition, and not subnetwork competition, since as discussed earlier (§3.3.3), entrants engaging in subnetwork competition are likely to face similar costs as the incumbent, up to the point of the new subnetwork. Furthermore, there are already a large number of players in the distribution market. ELBs with lower average costs might be able to successfully compete by cross-subsidising the cost of bypass or subnetwork competition from their existing customer base.

However, within the context of contestability theory, neither economies of scale—nor fixed costs which result in decreasing average costs—are in themselves necessarily considered barriers to entry. Scale economies need not constitute a source of market power. The key factor is ‘*sunk costs*’. Scale economies may exacerbate the effects of sunk costs on entry deterrence, but economists such as Spulber (1989, s1.3.1) claim that it is the sunk costs which actually raise the barrier in the first place.

3.5 Sunk Costs, Asset Specificity and Fungibility in Electricity Distribution Networks

3.5.1 Barriers to Entry from Sunk Costs, Asset Specificity and Imperfect Fungibility

Spence (1977) was the first to suggest that an incumbent monopolist might deliberately incur the costs of *excess* capacity to make credible a threat to increase output, and/or to lower price, in order to deter entry. Dixit (1980) took Spence’s reasoning to the next stage, and showed that, in general, sunk costs associated with durable capital goods, could be sufficient to deter entry even without the incumbent having to incur the excess capacity. Capacity is often classified as “sunk” if there is no effective “second-hand market” for capital goods. Entry becomes deterred because, by “sinking” capital into durable assets, the incumbent creates a fundamental *asymmetry* between itself and the potential entrant. The incumbent has sunk its fixed costs of capacity, so its behaviour is only motivated by the desire to achieve a surplus over variable costs (i.e., ‘quasi-rent’; §5.3.2). Potential entrants are thus subjected to the risk of retaliatory behavior on the part of the incumbent, which may result in part of the costs of their investment becoming ‘*irrecoverable*’ (e.g., Hay and Morris, 1993, pp. 90-91).

Bollard and Pickford (1995) have described the key characteristics of electricity distribution networks as comprising lumpy, immobile, indivisible, and *sunk* assets. And Teplitz-Sembitzky (1990, p. 47) states that investments in a power distribution system are almost entirely sunk, because they involve “transaction-specific” commitments in assets which have no use other use than that of distributing electricity. This description harkens back to the theory of transactions governance (§2.1.6), which suggests that the principle dimensions of any market transaction are uncertainty, frequency and ‘*asset specificity*’ (Williamson, 1986b). Asset specificity refers to a situation in which “durable investments” are undertaken in support of particular transactions, and the value of those investments would be much lower in best alternative uses *or by alternative users*, should the original transaction be prematurely terminated.

Williamson distinguishes four types of asset specificity: (i) human asset specificity; (ii) physical asset specificity, which relates to mobile assets of a specialised nature, similar to the essential input concept (§3.2.1); (iii) site specificity, relating to assets which are difficult to relocate, and where successive stages of production are located in proximity; and (iv) dedicated assets. Clearly, in distribution networks, the collective and semi-individual networks (§3.1.1) have strong site-specific characteristics. Williamson suggests that common ownership—in other words, vertical integration—is usually the most appropriate relationship in circumstances of site specificity. On the other hand, since investment in dedicated assets is performed on behalf of a particular buyer, common ownership is “rarely contemplated”. The conclusion that such a classification has for electricity distribution networks is that dedicated network connections, at any voltage level, could potentially be provided on a contestable basis. Asset specificity therefore provides weight to Mercury’s conclusion that the provision of dedicated HV connections is a contestable activity (§3.2.3), as well as to Mercury’s explanation for why large embedded, but dedicated, subnetworks can be efficiently owned by third parties (§3.4.1).

In discussing asset specificity, Williamson states that it is common practice to distinguish between fixed and variable costs, but considers that this is simply an “accounting distinction”. Both fixed and variable costs can have varying degrees of specificity, and what is more relevant is whether assets are “redeployable” or not. This concept bears a close resemblance to that of ‘*asset fungibility*’, where ‘*perfect fungibility*’ requires that “the value of any item of capital equipment in its best alternative use must be equal to that in the firm which initially undertook the investment” (BPW, p. 377). Miller (1995) states that network industries, such as electricity distributors, are characterised both by assets which are substantially ‘*non-fungible*’, as well as by substantial sunk costs. Given that non-fungibility implies the absence of a second-hand market for capital goods, this may at first appear to be a tautological description. However, a possible distinction is made shortly (§3.5.4). A key characteristic of imperfect fungibility is that it makes costs ‘*intertemporally interdependent*’ through the linkage provided by the non-fungible capital. In other words, costs are not ‘*intertemporally separable*’, since this would require that costs incurred during any one period are unaffected by the output quantities in any other period (BPW, p. 373); consequently, there are no intertemporal economies of scope (§3.4.2). As is noted above (§3.4.3), the imperfect fungibility of distribution network assets, and the associated intertemporal interdependence of network costs, makes it difficult to determine an LRAC curve even when just a single product is provided.

Williamson indicates that the transactions governance approach requires the study of contracting to include both *ex ante* and *ex post* features. He suggests that *ex post* competition following an initial arrangement to supply a particular good will only be effective if that good or service is not supported by investments in “durable transaction-specific assets”. Where specialised investments are not involved in market contracts, the winning supplier realises no advantage over non-winners. Although the winner may continue to supply for a long period of time, this is only because it is, in effect, continuously

meeting the competitive bids of rival suppliers. By contrast, reliance on transaction-specific assets introduces “contractual asymmetry” between the winning and non-winning suppliers. Both the incumbent supplier and the consumer(s) have a clear incentive not to terminate the existing arrangement. On the one hand, the supplier would be unable to realise the specialised asset’s value if it were redeployed, and on the other, the consumer(s) would have to induce potential entrants to make the same specialised and non-redeployable investments.

Sharkey (1982a, p. 37) states that: “a sunk cost is the difference between the *ex ante* opportunity cost and the value that could be recovered *ex post* after a commitment to a given project is made”. Sharkey also explains that whether a cost can be considered sunk depends on the definition of the market. For instance, Sharkey (1982a, pp. 37-38) notes that while the capital embodied in an airplane is sunk in the sense that the airplane can only be used for transportation, the capital embodied in an airplane serving a particular market is not a sunk cost, because it is easily transferred to other markets.

If a firm is not successful in recovering the entire pre-commitment opportunity cost that has been sunk in an asset, at some later point in time following the commitment, then that asset may be considered wholly or partially ‘stranded’. The notion of ‘stranded assets’ is often associated with ‘irrecoverable costs’ in real markets, particularly in electricity supply industries that are in the process of being deregulated (e.g., Black and Pierce, 1993; Baumol and Sidak, 1995a-b).⁴⁷ The notion of stranded costs can be considered equivalent to earning below a normal (or zero) economic profit.

That inability of utility shareholders to secure the return of, and a competitive rate of return on, their investment gives rise to the condition known as stranded investment or *stranded costs* (Sidak and Spulber, 1997, p. 29).

Hay and Morris (1993, p. 94) state that sunk costs are an essential element in the analysis of potential competition, and depend on “exit costs” reflecting the disposal value of the capital installed. Should the disposal value be very low, then the sunk cost element will be very high, and vice versa. Under Williamson’s transactions governance framework, the exit costs of terminating an existing arrangement would be high both for supplier and consumer, thus reinforcing the argument that sunk costs are associated with strong barriers to entry. Spulber (1989, s1.3.3) also seems to imply a similarity between sunk costs and asset specificity by stating that: “the market response to sunk cost and attendant risk is the long-term contract”. He suggests that “judging by the large number of long-term contracts,

⁴⁷ However, Sidak and Spulber (1997, p. 29) go on to distinguish between “stranded costs” and “stranded investment” and to link the concept to deregulating previously regulated markets: “More precisely, stranded investment is a subset of stranded costs. The latter includes expenditures (such as the mandatory purchase of energy at the utility’s avoided cost but above the market price of such energy) that are not capital investments in physical plant per se, but that nonetheless reflect outlays required by regulators that firms cannot recoup in the presence of competitive entry”.

sunk costs are a common phenomenon”, whereas Williamson might cite this as evidence that asset specificity is common. On the other hand, this perspective of sunk costs and asset specificity provides its own explanation for the traditional product differentiation barrier to entry (§3.4.5)—the bond between supplier and consumer can be strong even without a legally-binding contract.

3.5.2 *Sunk Costs and Barriers to Entry in Contestability Theory*

Sunk costs are a fundamental concept in contestability theory, since it is the absence of sunk costs which allows a market to be perfectly contestable. One of the wider circle of contestability theorists describes the significance of sunk costs as follows.

The single most important element in the design of public policy for monopoly should be the design of arrangements which render benign the exercise of power associated with operating sunk facilities (Bailey, 1981).

Hay and Morris (1993, p. 91) go as far as stating that Baumol and his colleagues assert the only barriers to entry are those which arise from the existence of sunk costs. However, Hay and Morris (1993, p. 94) also note that one of the key contributions of contestability theory is its focus on ‘*barriers to exit*’, which had previously received little attention in the literature.⁴⁸

Clearly the Baumol group’s perspective of sunk costs is crucial when it comes to considering the business of electricity distribution within a contestability framework. Initially, BPW’s standard text on contestability theory defines sunk costs as being costs that are both ‘*irrecoverable*’ and ‘*irreversible*’.⁴⁹ BPW (pp. 6 and 377) equate ‘reversible entry’ with ‘costless exit’, and define a perfectly reversible investment as one for which there exists perfect rental or resale markets, which requires the perfect fungibility of all capital utilised in production. Like Sharkey, BPW link the concept of sunk costs with that of *ex ante* imperfect fungibility, by stating that, if “some portion of capital is imperfectly fungible” it “constitutes a sunk cost once it is installed” (BPW, p. 377). However, BPW also use the term “sunk” simply to describe outlays of capital that are tied up in the construction of durable asset capacity.⁵⁰ The issue appears to be that BPW use the term sunk both to apply to capital which is sunk *physically*, and to

⁴⁸ “Determination of the structural contestability of a market requires evaluation of the costs of entry and exit and of the magnitude of unavoidable sunk costs. In particular, the availability of resale markets for durable inputs and their usability in other activities (fungibility) must be investigated since, clearly, the less the financial loss incurred in such a transfer, the lower will be the costs that are truly sunk and the smaller will be the costs of exit” (BPW, p. 469).

⁴⁹ “When entry requires the sinking of substantial costs, it will not be reversible because, by definition, the sunk costs are not recoverable. However, if efficient operation requires no sunk outlays, then entry can, by and large, be presumed to be reversible, and the market can be presumed to be contestable” (BPW, p. 7).

⁵⁰ This arises because BPW (p. 412) make “the crucial distinction between costs sunk in capacity construction and fixed costs incurred in the establishment of a firm”. As is pointed out in the next sub-section, the fixed costs associated with non-durable capital could also, in BPW’s sense of the term, be considered sunk.

capital which is sunk *economically*.⁵¹ Capital which is *not* sunk economically is defined by BPW (p. 382) as having intertemporally separable costs.

Baumol and Willig (1981) state that barriers to entry can arise from the need to “sink costs”. The barrier arises should such costs create an asymmetry between the incumbent and competing firms. In such a scenario, sunk costs are any imperfectly fungible expenditures incurred by the incumbent in the past, which an entrant would have to duplicate upon entry. Further, Baumol and Sidak (1994a, p. 129) suggest that, where “large sunk investments are required ... the second firm to enter a market is likely to face higher funding costs than its predecessor’s because the entrant faces an added risk of strategic entry countermeasures by the incumbent. This is not an imperfection in the capital market, but represents its accurate pricing of risk”. In either case, entrants would incur a cost—associated with production in the first case, and financing in the second—that is not currently incurred by the incumbent firm. Consequently, Baumol and Sidak state that they consider Stigler’s definition of entry barriers to be the most appropriate (§3.4.5). Even though an incumbent and an entrant may face identical production costs, if those production costs involve “sunk investment”, then the costs incurred by incumbent and entrant are different, *at the time of entry*. Hence, the sequencing and *timing* of costs is important (§5.3.1). Baumol and Sidak liken this entry barrier to the traditional notion that there is some minimum efficient scale required for an entrant to be able to compete effectively (fn. 42).

The use of the term “sunk” by Baumol and his co-authors in the rather weak sense of simply pertaining to “historic” capital expenditures is rife in the literature relating to electricity pricing (as is discussed in the next Chapter). Unlike the initial definition of the term “sunk costs” outlined in the standard contestability text, Baumol does not appear to be implying that the incumbent’s costs are necessarily irrecoverable. If past capital outlays are clearly *already* irrecoverable (i.e., stranded), irrespective of whether a competitor enters, then one way of looking at this scenario is that the incumbent actually has no more to lose by exiting the market.

By contrast, a firm that believes some of its costs may *become* irrecoverable upon exit, in response to competitor entry, is certainly likely to retaliate. It will realise that if its prices need to drop temporarily to successfully deter entry, then any loss of revenue is also only temporary, and could possibly be recovered later. But there is no reason why an incumbent monopolist with perfectly fungible durable assets would not also retaliate to the threat of entry, in an attempt to maintain any pre-existing

⁵¹ BPW (p. 381) state that, at the time when some investment or salvage is undertaken, older capital stock of the same type, which would otherwise be “sunk”, becomes transformed into capital which is “liquid at the margin”. That type of capital then “becomes perfectly substitutable for investment (or salvage) and the two must, therefore, have exactly the same financial (rental) value. ... For even though the capital is sunk physically, it is not sunk economically on the margin at a date at which its owner chooses to augment its amount by means of gross investment” (BPW, p. 382).

monopoly rents. Imperfect fungibility is by no means the only motivation for retaliatory response. Although the possibility of retaliation, like barriers to entry, is inconsistent with a perfectly contestable market (§2.1.7), in Baumol and Willig's view the threat of retaliation is not itself a barrier to entry. The asymmetric costs, which themselves are incompatible with perfect contestability, act as the barrier to entry. Hence, in contestability theory, if an incumbent's assets are all perfectly fungible, then by definition there is no barrier to entry, even in the presence of significant economies of scale, and even though the incumbent may potentially retaliate.

3.5.3 *Sunk Costs as Sources of Intertemporal Unsustainability*

However, Baumol and his colleagues uncovered a complication in this articulation of the role of sunk costs as barriers to entry. While sunk costs are generally considered in contestability theory to act as entry barriers, in an intertemporal context, BPW (p. 412) concluded that the higher the sunk costs involved in the construction of new capacity, the less likely is the sustainability of an incumbent monopolist. In other words, when considered in a non-static sense, sunk costs do not necessarily act as a barrier to entry after all, in fact they may actually encourage new entrants, potentially leading to "destructive" or "excessive" competition" (e.g., Sharkey, 1982a, p. 25; and Spulber, 1989, s1.3.2, respectively). So, on the one hand, depending on the *timing* of investments, sunk costs may favour the incumbent, acting as a barrier to entry, while on the other, sunk costs may favour the entrant, inviting inefficient entry. In any event, BPW (p. 476) state that "sunk costs can be singled out as the villain". As discussed above (§3.3.3), BPW's concerns relating to unsustainability have led some authors to suggest that electricity distributors need regulatory protection against inefficient entry (e.g., Teplitz-Sembitzky, 1990, p. 49). Nevertheless, BPW place a strong burden of proof on those who would demonstrate unsustainability.⁵² After all, duplication of facilities and excess capacity can also occur as the result of the "discovery process" in real-world competitive markets (§6.1.3).⁵³

⁵² BPW (p. 473) state that: "one must proceed with great caution. As long as any doubt remains about the unavailability of sustainable solutions, one must hesitate before bowing to pressures for the encouragement of barriers to entry. It is understandable and natural for the incumbent firms in an industry who are fearful of enhanced competitive pressures to seek the erection or toleration of protective umbrellas against entry. But those who have the task of protecting the interests of society must resist such demands until the evidence for them is all but incontrovertible". On the other hand, in the chapter added in the second edition of their text on contestability theory, BPW (p. 486) appear to shift the burden of proof somewhat: "before anyone can legitimately use the analysis to infer that virtue reigns in some economic sector, and that interference is therefore unwarranted, that person must first provide evidence that the arena in question is, in fact, highly contestable".

⁵³ For instance, Sidak and Spulber (1997, p. 27) note that: "in competitive markets there is often duplication of investment, and the entry of excess or insufficient capacity can take place as a consequence of uncertainty regarding costs, technology, or market demand". Consequently, Berg and Tschirhart (1995) advise that: "With deregulation, identifying which firm (or location) has economic advantages is a task involving trial and error. Unless chronic excess capacity and duplication of

In presenting their model of intertemporal unsustainability (discussed in detail in Chapter VIII), BPW make a “sunk costs assumption” requiring that: “constructed facilities have no other valuable use outside the industry in question, so that once built, these facilities are sunk” (BPW, p. 407). This would be a weak definition of non-fungibility—since BPW also appear to define ‘*perfect non-fungibility*’ with respect to assets which have no value in *any* alternative use at any time, not just *outside* the industry (e.g., BPW, p. 378 and p. 429)—and an even weaker definition of sunk cost. On the other hand, BPW consider that the fixed costs incurred in the establishment of a firm tend to improve sustainability. This conclusion is consistent with the traditional assumption that economies of scale act as a barrier to entry (§3.4.5), since fixed costs, like excess capacity, are key contributors to scale economies in their broadest sense. Yet in BPW’s model of intertemporal unsustainability, fixed costs are also assumed to be “sunk”, because no resale market exists for establishment costs, and as such they are perfectly non-fungible. Hence, in contestability theory, static sunk costs contribute to sustainability, whereas intertemporal sunk costs detract from sustainability.

As is discussed later (§8.3), there are a number of reasons for considering that the Baumol group’s concerns regarding unsustainability are unfounded in real markets, and asset specificity provides another possible justification for being less concerned about the unsustainability of incumbent monopolists. If potential entrants see that consumers are open to breaking an existing commercial relationship with their current supplier, then potential entrants may fear that the same fate awaits them.

3.5.4 *The Spectrum of Sunk Cost Definitions from Historical to Irrecoverable*

The discussion in the previous sub-sections indicates that the various applications of the term “sunk costs” lie on a spectrum ranging from: capital simply “sunk” in durable physical assets, at any time; through to assets which are “sunk” because they are imperfectly fungible and were installed in the past (and hence involve some level of “irreversible *pre-commitment*”); to investments which are “sunk” because their costs are *potentially* not fully recoverable; then to those assets which are “sunk” and their costs will definitely not be recovered, or have actually not been recovered, now that the full lifetime of the investment has been reached.

Difficulties arise especially when the term “sunk” is used in a number of different ways. For example, Teplitz-Sembitzky (1990, p. 37) states that sunk costs always result in economies of scope among outputs supplied at different dates, which simply equates sunk costs with the intertemporal interdependence caused by non-fungibility, irrespective of the time period in question. Also, Teplitz-Sembitzky (1990, p. iv) defines sunk costs as those which are “partly or completely irrecoverable in the short-medium term”. This does not appear to imply that those costs are not necessarily recoverable at all,

facilities is likely to result, a strong case can be made that this task be left to the marketplace, rather than to administrative procedures”.

just in the short term. But in a later work, Teplitz-Sembitzky (1992, pp. i and 4) defines sunk costs as those which “cannot be recovered in the long run”, and as occurring when “the worth of the sector’s assets cannot be recovered entirely upon exit”. However, if an entrant’s costs are guaranteed to be “recoverable”, then any “sunk” costs required to enter the market do not act as a barrier to entry: “Sunk costs do not raise entry (and exit) barriers if, after entering the incumbent’s market, an entrant is able to recover the sunk costs incurred before exiting the market” (Witteloostuijn, 1990).⁵⁴

Notwithstanding his multiple uses of the term, Teplitz-Sembitzky (1990, p. 46) correctly points out that, apart from any unavoidable transaction costs related to installing or reselling equipment, the costs of durable assets are not necessarily sunk. Both Teplitz-Sembitzky and Shepherd (1984) point out that sunk costs need not necessarily be embodied in durable inputs. Sunk costs may be associated with less tangible forms of costs, such as research and development (e.g., Stiglitz, 1987), advertising, and training. Hence, just as Williamson suggests the notion of asset specificity should be disassociated from fixed costs, so too should sunk costs from durability, and vice versa.

Miller appears to distinguish between “sunk costs” and “non-fungible assets” (§3.5.1), suggesting that the two concepts are not necessarily equivalent. This distinction occurs where the concept of sunk costs is linked to that of irrecoverability, rather than perfect non-fungibility (which Baumol and his colleagues equate with irreversibility). Baumol and Willig’s (1981) description of sunk costs only implies that such are *potentially* irrecoverable, not actually irrecoverable, whereas Loubé (1995), for example, declares that sunk costs are costs which could not be recovered if a firm goes out of business. The possibility of irrecoverable costs clearly raises incumbent barriers to exit. Incumbent firms retaliate because they fear entry will cause them to incur irrecoverable costs, and those costs would not be irrecoverable had the investments not been made in sunk assets.

However, simply because assets have no value in any *alternative* use does not necessarily imply that such assets have no value in their *current* use. For instance, in so much as R&D, advertising, training, installation and transaction costs can be incorporated into charges to consumers for asset utilisation, these non-durable expenditures are not necessarily irrecoverable or stranded. The same applies to durable assets. Investments in immobile and durable assets may be both physically irreversible, but still economically fully recoverable. Assets will be economically fully recoverable, even

⁵⁴ Similarly, Sidak and Spulber (1997, p. 25) state that: “To establish a network, industries such as telecommunications and electric power must make substantial nonrecoverable, market-specific investments, known as *sunk costs*. Networks represent the quintessential sunk cost. The transportation and reticulation facilities in telecommunications, electricity, railroads, oil and natural gas pipelines, and water services are tied to specific geographic locations. The capital from the facilities has little if any scrap value, and the facilities cannot physically be transferred to another market, unless alternative uses can be found *in situ*. ... The recovery of sunk costs is a critical aspect of network industries that complicates deregulation”.

where no alternative use exists for those assets, as long as the stream of payments to capital received from asset users over the lifetime of those assets provide a return *on* the capital employed, and ensure the return *of* the original capital (§7.1).⁵⁵

Whereas “irrecoverable” is a concept which requires *ex post* evaluation to determine whether the costs of investment in question actually were not recoverable, “irreversible” simply relates some ‘*pre-commitment*’ to a non-fungible investment (e.g., Hay and Morris, 1993, p. 101), and hence relates to the asset specificity of the investment. Yet, although describing sunk costs as being *ex ante* “irreversible” (Sidak and Spulber, 1997, p. 423), Sidak and Spulber do not see irreversible costs as necessarily being barriers to entry.⁵⁶

Competition nonetheless remains feasible even with sunk costs. The need to sink costs should not be viewed as an insurmountable barrier to the entry of new competitors. All competitive markets involve some degree of irreversible investment, whether in capital equipment, marketing, or research and development. Entrants commit capital resources in markets where they expect to earn competitive returns on their investments (Sidak and Spulber, 1997, pp. 26-27).

The notion that past sunk expenditures have *no* bearing on current and future investment decisions is wholly invalid where such historic investments are imperfectly fungible. Baumol and Willig (1981) suggest that an incumbent’s historic “sunk” costs—in BPW’s sense of *potentially* irrecoverable costs—are viewed by possible entrants as part of their incremental cost of entry, and therefore an incremental risk which must be covered by post-entry revenues. Consequently, a barrier to entry arises because the incumbent firm, in its current and future decisions, does not take into account expenditures that it has already incurred. Yet, as noted above, Baumol and his colleagues cite one of the key characteristics of imperfect fungibility as being the resultant intertemporal interdependence of production. Therefore, in that respect, historic expenditures do have a bearing on current and future expenditures.

⁵⁵ Williamson (1986b) suggests that asset specificity makes the contractual relationship between supplier and consumer strong, which implies that the arrangement will not be terminated before the supplier recovers its investment outlay. Hence, asset specificity is less likely to be associated with potentially irrecoverable costs. Moreover, while asset specificity can lead to transaction cost-minimising market organisation, sunk costs are viewed as a potential source of allocative inefficiency.

⁵⁶ Similarly, Cairns and Mahabir (1988) argue that contestability can even be increased (i.e., barriers to entry be lowered) where sunk investments are required, particular if potential competitors exist in some other “prototype of the market to be contested”. For instance, the geographical basis of each electricity distributor’s “market” could be seen as a “prototype” of any other distributor’s market. The example of “identical products” in a “geographically distinct market” is specifically addressed by Calem (1988). In his scenario of a “penetrable market”, ease of entry derives from “low transport costs”. Boundary or fringe competition in power distribution networks (§3.2.1) could be a case in point.

Nevertheless, this interdependence does not necessarily mean that the barrier to entry is removed entirely, since the bearing past costs have on present and future costs may not be as great to the incumbent as they are to the potential entrant. On the other hand, the opposite may be true, as BPW's model of unsustainability suggests. An incumbent's past investments in non-fungible assets may mean that the incumbent is unable to meet current or future demand at as low a cost as a potential entrant, particularly where the industry is subject to strong economies of scale and scope. Historic expenditures in durable non-fungible assets can both raise and lower barriers to entry, in the sense of altering costs of production in favour, or to the detriment, of an incumbent firm. In a static world, the concept of symmetric costs is easy to define. But when those statically symmetric firms act in an intertemporal world, that symmetry often dissolves.

3.5.5 *The Fungibility of Distribution Network Assets*

Perfect rental or resale markets exist for very few components of an electricity distribution network. BPW (p. 377) state that in such markets the price for the use of any capital good is determined *outside* any particular industry, with no role played by allocative behaviour within that industry. As noted above, Baumol and his colleagues state that perfect fungibility is a necessary condition for perfect resale markets, but are less specific on whether the converse is true (§3.5.2). Although intertemporal interdependence is considered to be a key characteristic of imperfect asset fungibility (§3.5.1), if assets have the same value in an alternative use *inside the industry*, or even *within the firm*, then costs are not so strongly intertemporally interdependent after all.

Some distribution assets, for example underground cables, can generally considered to be non-fungible. The costs involved in digging cables up, and the possibility of damage to them in the process, make the potential resale market for second-hand cables, even as scrap, very thin. Lines are less difficult to remove, but again resale markets are generally limited. On the other hand, distribution transformers can be easily and relatively inexpensively relocated within an incumbent's own network. Although there are relocation costs, these will only become partially stranded if line charges are unable to cover those costs.

Obviously the possibility for alternative uses within the firm, or the industry as a whole, increases with the extent of the network. When extended to all distribution firms, clearly the potential for relocation and resale of distribution assets increases. For instance, there are fewer zone substation transformers than consumer-level distribution transformers. Consequently, the number of alternative uses, even on a national level, would appear to be much lower for HV transformers, than for MV distribution transformers. Alternative uses are also likely to occur more frequently in a rapidly expanding network as opposed to one experiencing only gradual growth.

3.6 Electricity Distribution Costs in New Zealand

3.6.1 Distribution Network Asset Costs

Distribution network assets in New Zealand are typically classified into four major categories: (i) system assets, mainly comprising the system control and data acquisition (SCADA), communications, and load control equipment (such as that used to switch off water heaters at times of peak load); (ii) the high voltage (HV) network, comprising the zone substations and the subtransmission network which connects them back to the Trans Power points of supply; (iii) the medium voltage (MV) network, comprising the main distribution feeders, and distribution substations supplying large MV consumers directly or groups of LV consumers; and (iv) the low voltage (LV) distribution network, applicable to individual LV consumer installations. Data from New Zealand ELBs suggest that, on average, around one quarter of all capital costs are associated with the system and HV assets, just under 55% are associated with the MV assets, and one fifth for LV assets.⁵⁷ While the MV system accounts for the majority of distribution network costs on average, there is a large range in the data. For instance, HV asset costs range from 14% to 40%, with the lower end of the range relating to ELBs serving predominantly rural areas, and the higher end relating to dense urban areas.

MV distribution transformers and HV zone substation transformers are the main assets which potentially have resale value, and hence may be partially fungible (§3.5.5). It is estimated that MV transformers account for around 7%-16%, and HV transformers for about 5%, of total distribution network costs respectively.⁵⁸ Therefore, the fungibility of distribution network assets, whether considered outside or inside the industry, or even within a particular firm, appears to be relatively low.

Both types of transformers are subject to strong economies of scale. Standard distribution transformer cost data in the Ministry of Economic Development's Optimised Deprival Valuation (ODV) Handbook (Energy Markets Regulation Unit, 2000c) indicate that the costs of three phase ground and pole mounted distribution transformers fit a first or second order function (with constant or declining marginal costs) well, and the costs of one and two phase distribution transformers closely fit a linear function with a large non-zero intercept.⁵⁹ Standard costs for zone substation transformers are not listed

⁵⁷ This data comes from the asset registers of a number of ELBs for the year ending March 1998 (see Bibliography). Total replacement cost data for network assets were available for 12 ELBs (including the two largest).

⁵⁸ Total replacement cost data for MV distribution transformers were available from the asset registers of a number of ELBs for the year ending March 1998 for 12 ELBs (including the two largest). The cost of distribution transformers as a percentage of total network replacement costs ranged from 6.6% to 15.9% with an average of 11.3%. Total replacement cost data for HV transformers were available for 7 ELBs. As a percentage of total network replacement costs these assets ranged from 3.2% to 6.7% with an average of 5.0%.

⁵⁹ The cost functions are as follows: three phase (ground), Cost (\$) = $4,554 + 34.5kVA - 0.0143kVA^2$, with an r^2 of 0.996; three phase (pole), Cost (\$) = $4,554 + 34.5kVA - 0.0143kVA^2$, with an r^2 of 0.996; three phase (pole), Cost (\$) = $2,557 + 45.3kVA -$

by the Ministry of Economic Development. However, one ELB provides costs for a number of standard sizes, and the linear function below (3.2) exactly fits the transformer cost data.⁶⁰

$$\text{Individual Transformer Cost (\$)} = 160,000 + 16,000 \times \text{Transformer Capacity (MVA)} \quad (3.2)$$

Additional “products”—in other words, additional consumer connections—result in economies of scope in a distribution network, because of the economies of scale inherent in such joint capacity as distribution and zone substation transformers. Although lines and cables are also subject to economies of scale with respect to capacity, distance is a much more significant factor in driving circuit costs. Because zone substations are the “lumpiest” investments in any distribution network, and are part of the “collective” or “common” network shared by many consumers, the cost characteristics of zone substations warrant closer examination.

3.6.2 Zone Substation Design and Costs

Vector (1999) describes how zone substation design is constrained.⁶¹ Fault level constraints limit the number of transformers in any single zone substation to three, and generally each transformer should be of the same capacity. Vector also notes that each substation is limited to a ‘firm capacity’ of 50 MVA. Given that Vector’s network serves primarily urban areas, the design contingency criteria (§3.1.3) is $n-1$ where n is the number of transformers in the zone substation. Hence, the firm capacity of any zone substation is not just the sum of its installed transformer capacities. In the event of a single transformer failure, the remaining operational transformer(s) can be temporarily overloaded for a number of hours. During this period, load normally supplied by the zone substation with the failed transformer can be transferred to neighbouring substations. This is possible because although the network is operated radially, feeders are interconnected in a mesh with some switches normally open (§3.1.2). On the other hand, load transfer between substations may not be cost effective for isolated substations in rural areas. Consumers may therefore have to accept a lower contingency level (i.e., $n-0$) in order to make the cost of supply viable, unless they are willing-to-pay for the higher security.

0.0294kVA², with an r^2 of 0.996; and single phase, Cost (\$) = 1,851+45.7kVA, with an r^2 of 0.986. The three phase (ground) curve is for transformer sizes less than 1MVA. Including sizes up to 1,500kVA, the three phase (ground) cost function is Cost (\$) = 6,549+20.95kVA, with an r^2 of 0.985. Raw cost data taken from Energy Markets Regulation Unit (2000c, Table B.1).

⁶⁰ This curve is derived from the raw data underlying the distribution cost study (Mercury Energy Lines Business, 1999) described above (§3.4.3).

⁶¹ This is not to suggest that all ELBs in New Zealand have similar approaches to network design as Vector. However, this example provides a useful example for understanding the key cost drivers of electricity distribution, particularly at HV and MV levels.

The firm capacity of the (urban) zone substation is thus the minimum of the overloaded remaining capacity (because it is a temporary measure) and the remaining capacity plus the transfer capacity. The zone substation firm capacity (X_Z) is thus as shown in (3.3), where θ is the overload factor (cited by Vector as 150% for a two hour period), Θ is the load transfer capability (cited as 10 MVA), and X_T is the installed capacity of each of the n transformers. The contingency criterion is represented by n_C , and is set to $n-0$ or $n-1$ and so on, as appropriate. Vector suggests that the only discrete transformer sizes available are 2.5MVA, 5MVA, 10MVA, 15MVA and 20MVA, which means that the largest capacity zone substation would have a configuration of 3x20MVA transformers. Total installed capacity would be 60MVA, but the firm capacity would be 50MVA given Vector's overload factor, transfer capacity and contingency criteria. Hence, the firm capacity for two zone substations with identical transformer capacities is not necessarily the same, since firm capacity also depends on factors other than substation design.

$$X_Z = \text{Min}(\theta n_C X_T, n_C X_T + \Theta) \quad (3.3)^{62}$$

Returning to the cost model discussed earlier (§3.4.3), which also relates to Vector's network, a cost function for a zone substation can be derived from the underlying raw data. Apart from fixed costs, zone substation construction costs are dependent on two key factors, the installed capacity of the transformers, and the number of transformers. The second factor is important, because even if the sizes of the transformers change, costs only associated with the number of transformers remain constant. Fixed costs comprise: one-off administration costs such as easement and planning consents; land cost; and the zone substation building cost. Costs dependent on the number of transformers comprise: transformer bays and earthing, as well as metering and SCADA equipment. In addition, although not part of the zone substation itself, the number (and thus the costs) of incoming subtransmission lines/cables, and protection and communications equipment at the upstream Trans Power point of supply, are dependent on the number of transformers at the zone substation. The cost of the zone substation transformers themselves, and the cables between them and the incoming circuit breakers, are dependent on both the number and installed capacity of the transformers.⁶³ The total costs associated with the zone substation (K_Z), which include the costs of the upstream HV network, are therefore as approximated in (3.4).

⁶² This equation is derived from information presented in Vector (1999, p. 15).

⁶³ Zone substations also include switchgear equipment. However, these costs are more dependent on the number of outgoing (i.e., downstream) distribution feeders, and so are ignored here.

$$K_Z \cong (F_A + F_L + F_B + n[F_P + zC_{ST} + F_E + F_M + (a + bX_T)])(1 + f_p) \quad (3.4)^{64}$$

where: F_A is the administration cost; F_L is the cost of land; F_B is the cost of the building; n is the number of transformers; F_P is the cost per transformer of modifications at the Transpower point of supply; C_{ST} is the cost per unit of distance of individual subtransmission lines and cables, which itself could be a function of capacity; z is the distance from the zone substation to the point of supply; F_E is the individual transformer bay and earthing cost; F_M is the cost of metering and SCADA; $a + bX_T$ is the individual transformer cost, similar to (3.2), but also including associated linking cables; and f_p is a project management cost factor.⁶⁵

Figure 3.3 shows the total zone substation costs ignoring the fixed costs of administration, land and buildings, as well as the cost of the subtransmission lines/cables. The cost per transformer of Transpower point of supply modifications, transformer bays and earthing, as well as metering and SCADA, are based on Mercury Energy Line Business's (1999) data, and are \$30,000, \$68,000 and \$9,000 respectively. The project management factor is 10%. Individual transformer costs include the cost of the linking cables within the zone substation, and approximately satisfy $\$168,333 + \$16,333X_T$. The case of $n-1$ contingency with 10MVA transfer capability is contrasted with a case where no transfer is possible (under $n-1$), and another case with $n-0$ contingency (and no load transfer). Firm capacity is based on equation (3.3) with n_C set to $n-1$ and $n-0$ as appropriate. This demonstrates that requiring $n-1$ contingency has a substantial impact on capacity costs, and this is even more marked where no transfer capability exists from neighbouring substations.

Table 3.1 presents the firm capacity associated with different transformer configurations. Although it is possible to have a configuration of 3x2.5MVA, 3x5MVA or 3x10MVA transformers, given that many costs are dependent on the *number* of transformers, there is always a less expensive option available with just two transformers. It is possible to have three of the two larger sizes of transformer simply because 20MVA is the maximum transformer size assumed to be available. Hence, options marked "na" (i.e., not applicable) indicate that if it is assumed that any configuration is sufficient to meet the projected demand for capacity over the entire lifetime of the zone substation, then the *na* options would never be constructed. Given the particular contingency criterion, the overload factor and the available transfer capacity, there would also be a lower cost option that could provide the same level of firm capacity. However, as is discussed later (§6.2.2), the situation is different where demand is growing, and installed capacity must be increased over time.

⁶⁴ Note that the total cost is only considered to be approximately equal to the equation in (3.4); it does not provide an exact fit to the raw data.

⁶⁵ All costs include installation costs. For example, in the case of subtransmission cables, trenching costs should be included.

The average cost curves associated with each of these cases are presented in Figure 3.4. As noted for the average cost curve above (§3.4.3), this cannot be considered a true long run cost curve, because it does not take into account the optimal capacity expansion path matched to future demand. Hence, the fact that the average cost curve begins to rise at a number of points does not necessarily imply that two firms could successfully supply the demand served by this single zone substation. Moreover, future demand might require an existing substation to be expanded by another transformer, might require the replacement of existing transformers, or might require an entirely new transformer to be constructed (which could perhaps be undertaken by a different firm). Optimal investment and the capacity expansion path based on this cost function are developed later (§6.2.2). However, before returning to the issue of distribution *costs* over time, the next two Chapters turn to the issue of electricity *pricing*.

Contingency Criterion Overload (θ) Transfer Capacity (Θ)	<i>n</i> -1	<i>n</i> -1	<i>n</i> -1	<i>n</i> -0
	150%	150%	150%	150%
	10MVA	5MVA	0MVA	0MVA
1x2.5MVA	<i>na</i>	<i>na</i>	<i>na</i>	2.5
2x2.5MVA	3.75	3.75	2.5	<i>na</i>
1x5MVA	<i>na</i>	<i>na</i>	<i>na</i>	5
2x5MVA	7.5	7.5	5	<i>na</i>
1x10MVA	<i>na</i>	<i>na</i>	<i>na</i>	10
2x10MVA	15	15	10	<i>na</i>
1x15MVA	<i>na</i>	<i>na</i>	<i>na</i>	15
1x20MVA	<i>na</i>	<i>na</i>	<i>na</i>	20
2x15MVA	22.5	20	15	30
2x20MVA	30	25	20	40
3x15MVA	40	35	30	45
3x20MVA	50	45	40	60

Table 3.1: Indicative Zone Substation Firm Capacities (MVA)

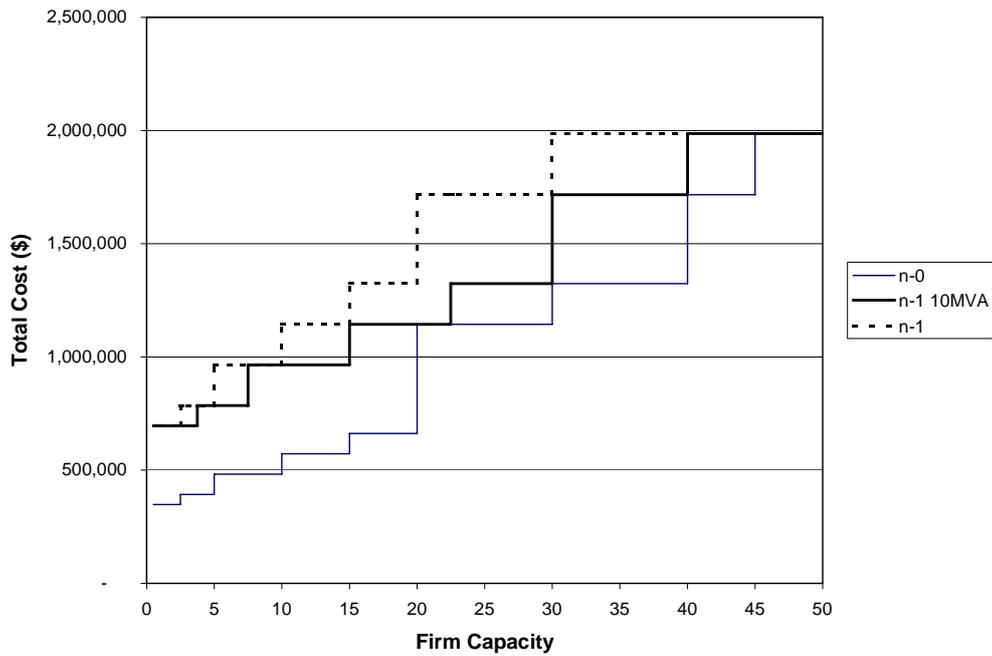


Figure 3.3: Zone Substation Total Costs (by Contingency Criteria and Transfer Capacity)

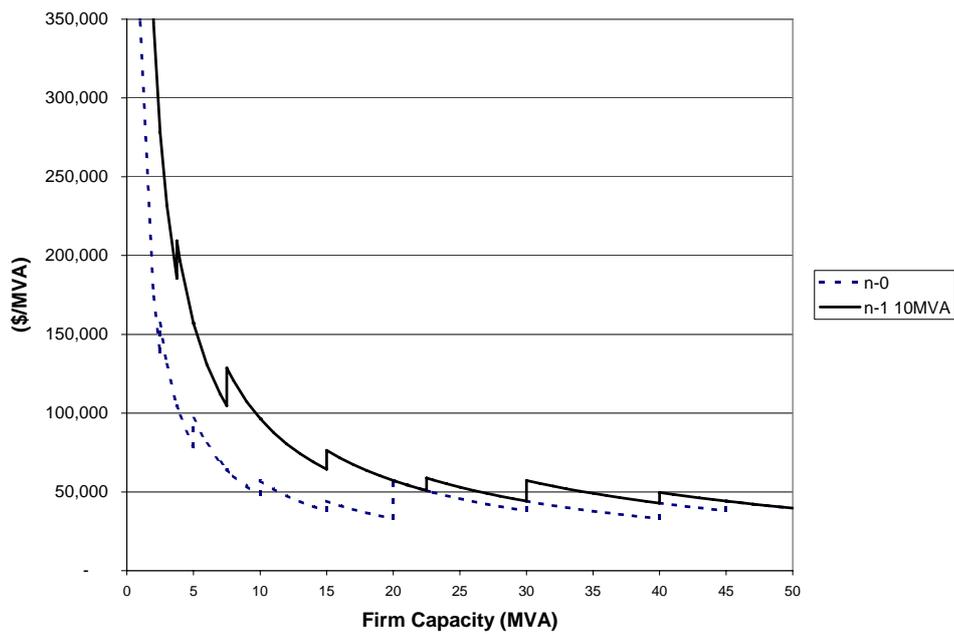


Figure 3.4: Zone Substation Average Costs (by Contingency Criteria and Transfer Capacity)

CHAPTER IV

EFFICIENT PRICING OF ELECTRICITY DISTRIBUTION: SHORT RUN OR LONG RUN CONCEPT?

A pricing policy based on long run marginal cost effectively permits the burden for the cost-benefit analysis to be placed on the electricity consumer, because he signals the justification of further investment by his willingness to pay the marginal cost of electricity supply: World Bank energy economists, Mohan Munasinghe and Jeremy Warford (1982, p. 27)

Long run marginal cost is an entirely irrelevant cost concept in decision-making. Resources invested in capacity expansion cannot be recovered. We find it advisable to dispense with the long run marginal cost concept altogether and rely on pricing based on short run marginal cost: Swedish economists, Roland Andersson and Mats Bohman (1985)

A rule that price must be set equal to marginal cost is a prescription for financial disaster: US exponents of contestability theory, William Baumol, John Panzar and Robert Willig (1988, p. 504)¹

“Getting the price right” is at the heart of striving for allocative and dynamic efficiency in the electricity supply industry, or in any sector of the economy, and Trebing (1984) for one has suggested that neoclassical economics has made its greatest contribution in the area of tariff design. There are three key dimensions of “efficiency” or “fairness” which are associated this goal. Firstly, is the *absolute price level* correct? Is the revenue recovered from all consumers commensurate with the opportunity costs of electricity supply, or in the case of electricity distributors, with the costs of network connection? In other words, do neither consumers subsidise firms nor vice versa? Secondly, are the *relative price levels* “fair”, such that no consumer cross subsidises another? Thirdly, answers to both of the previous two questions should explicitly acknowledge the *time dimension* involved. This is because both past capital expenditures, as well as expected future prices and demands, influence the efficiency of present day investment decisions *ex ante*, while actual future prices and demands will affect the optimality of investment *ex post*. (This dimension needs to be specifically highlighted, since it is not common to see research on efficient pricing explicitly discuss the issue of *intertemporal* cross subsidies). In light of these questions, this Chapter reviews the literature on efficient pricing of electricity supply in general, and examines some of the issues specifically applicable to setting line charges for distribution network connection.

¹ All quotes have been abridged for clarity, and acronyms spelt out in full.

4.1 Efficient Pricing of Electricity Supply

4.1.1 Marginal Cost Pricing of Electricity

Marginal cost pricing (§2.1.2) has long been advocated by many economists as the most efficient approach to the pricing of electricity supply.² Given that marginal cost pricing is so closely bound to the concept of Pareto optimality, it is interesting that Hay and Morris (1993, p. 635) suggest that few serious attempts have been made to apply the concept to industries other than power. The prescription that the price of electricity be set to marginal cost—regardless of how “marginal cost” might actually be defined—remains to the present day, even with respect to the provision of unbundled electricity distribution network services. For instance, in a recent presentation of the ‘essential facilities’ concept (§3.2.1), King and Maddock (1996) state that the basic principle of essential facility access pricing is simple; services should be provided at the *short-run marginal cost* of production, and if the facility is *congested*, then the price should be allowed to rise in order to ration existing supply. By ignoring the problem of ‘*indivisible*’ or ‘*lumpy*’ investment, capacity expansion is triggered when the value to consumers—as measured by the congestion price for the facility—exceeds the marginal cost of expansion. King and Maddock’s prime example of an essential facility is a residential electricity distribution network, which they maintain is characterised by substantial scale economies, high fixed costs, and a low marginal cost of operation. Consequently, they concede that, given the conflict between optimal short run facility pricing and ensuring an adequate return on investment, the potential application of the short-run marginal cost pricing principle may be difficult.

King and Maddock’s affirmation of this pricing principle is a re-iteration of the time-honoured approach for the efficient pricing of electricity service provision under capacity constraints, which goes back to the work performed during the 1940s and 1950s by Marcel Boiteux (§3.1.2) of Electricité de France (including: Boiteux, 1949; and Boiteux, 1956).³ It was not long before marginal cost pricing became the dominant theoretical paradigm for electricity tariff setting, and numerous authors expounded upon the approach (e.g., Turvey, 1968; Cichetti *et al.*, 1977; Scherer, 1977).⁴ In various guises, it subsequently became part of the standard texts on the subject of electricity economics (e.g., Turvey, 1975; Turvey and Anderson, 1975 and 1977; Munasinghe and Warford, 1982; Munasinghe, 1990; Berrie, 1992).

Marginal costs in the electricity supply chain can relate to both short run marginal costs (SRMC), when capacity is fixed—and costs are primarily driven by the second-to-second costs of operating

² Blaug (1990) traces the history of marginal cost pricing and its application to public utility pricing.

³ These papers, as well as Boiteux and Stasi (1952), were later republished in English in Nelson (1964), and Nelson’s versions are the references cited here (although the dates provided correspond to the original publication dates).

⁴ Joskow (1976) provides a summary of contributions to the theory from Boiteux through to the early 1970s.

generating plant, which are highly dependent on relative fuel costs and plant efficiencies—as well as to long run marginal costs (LRMC), when the capacity of generation, transmission and distribution facilities can all be adjusted (§2.1.4). If capital investments in electrical equipment were continuously adjustable, and not discrete (or indivisible) as they are in reality, then it can be shown that where capacity is optimally adjusted to demand, SRMC and LRMC prices are equal at equilibrium (e.g., Blaug, 1990). But, as Tabors (1994) points out, and as is elaborated upon shortly (§4.1.3): “because equilibrium occurs only in the calculations of engineers and economists, it is important to note the differences in application of these two methods in a live environment”. Moreover, indivisibility is not the only problem. The equality of SRMC to LRMC also requires that optimal production plans be independent of decisions taken in previous periods (e.g., Kay, 1971). But because the majority of assets involved in electricity supply have substantial non-fungible components (§3.5), this will never occur in practice.

The basic marginal cost pricing result is that price equals SRMC, but where there is a capacity constraint, price acts as a rationing mechanism and the supplier obtains a ‘scarcity rent’. The appropriate time to invest in new capacity occurs when SRMC exceeds LRMC, where LRMC is SRMC plus the ‘*marginal cost of capacity*’—often traditionally considered to be the ‘*annualised cost of capacity*’.⁵ Once the capacity increment is installed, in the traditional view, it becomes a sunk cost, and SRMC falls back to its old trend line (e.g., Munasinghe and Warford, 1982, Ch. 2). The cost is now sunk, in the sense that the investment in new capacity is now a *historic* irreversible cost not affecting the calculation of the efficient SRMC price. The accumulated scarcity rents would contribute to covering the cost of the new capacity. Williamson (1966) extended this basic approach to deal with one of the complications which King and Maddock (1996) assume away, namely indivisible capacity, and found that where capacity is under-utilised, the efficient price lies *between* the SRMC and the LRMC. Foreshadowing some of the later debates on marginal cost pricing, Johnson and Brown (1970) disputed this general annuitised cost of capacity approach to calculating LRMC, declaring that sunk or historic costs could not contribute to LRMC at all, since by definition, historic costs could not be marginal.

In summing up his review of the evolution of marginal cost pricing principles, Blaug (1990) concluded that marginal cost pricing “is a method not a dogma. It is grounded in Pareto optimality and the maximization of consumers’ and producers’ surpluses, but, then, so are all the policy views of economists”. He felt that proper application of the method “requires empirical judgements on a product-by-product basis about market structure, indivisibilities, externalities and elasticities of demand and supply; in short, it is a systematic check-list of what to look for in pricing a public service. It does not

⁵ The annualised cost of capacity is effectively the constant payments to capital which recover both a return on and return of capital over the lifetime of the asset which provides that level of capacity (§7.2.3).

therefore furnish any simple pronouncements about public pricing, except perhaps that almost any pricing rule is better than average cost pricing”.

4.1.2 Peak Load Pricing, Time-of-Use Pricing and Spot Pricing

The original and most basic approaches to marginal cost pricing assumed a relatively fixed demand curve, and both Boiteux (1949) and Williamson (1966) also examined the case where demand changes, by simplifying time-varying short run demand into ‘peak’ and one or more ‘off-peak’ periods.⁶ The standard solution to this ‘peak load pricing’ problem is the often-cited result that consumers in off-peak periods pay solely the marginal cost of operation (i.e., SRMC), whereas consumers in the peak period pay both the marginal costs of operation plus *all* the marginal costs of capacity (i.e., SRMC+LRMC). Peak load pricing theory primarily emerged from an analysis of the electricity supply industry, because demand for electrical energy varies moment to moment and electrical energy as a product is, to a large extent, non-storable. The underlying assumptions of the initial peak load pricing models are that: (i) the costs of capacity are homogeneous (i.e., output varies linearly in proportion to inputs—in other words, there are constant returns to scale); (ii) capacity expansion is optimal; (iii) capacity is fully utilised; (iv) there is no intertemporal elasticity of demand (i.e., demands between periods are independent); and (v) the system is at long run equilibrium. This set of assumptions, and the “conventional” solution to the peak load pricing problem—peak consumers pay β , the marginal (i.e., annuitised) costs of capital, plus b , the marginal costs of energy, while off-peak consumers pay b alone—is often attributed to Steiner (1957).

The application of peak load theory to real world power industry problems was initially directed toward power generation. Turvey (1968) relaxed the first assumption of homogeneous production capacity with respect to investments in power generation plant, and demonstrated a key result for generation planning purposes, namely, the efficient provision of a periodic demand for electricity generally requires a mix of different types of generating capacity where operating costs are inversely related to the costs of capacity. Turvey emphasised the calculation of the long run cost aspects of providing generation assets to meet a fixed set of demands for electrical energy.

Wenders (1976) followed Turvey’s approach, but relaxed the second assumption as well. By not requiring capacity to optimally match the typical load duration curve, he showed that in such more realistic cases, off-peak consumers contribute to some of the capacity costs after all. Crew and Kleindorfer (1975) relaxed the third assumption and produced a similar finding. They extended this

⁶ Faulhaber and Baumol (1988) trace the development of peak load pricing theory, and find its origins among the industry forums of electric utility engineers more than a century ago. They cite the earliest reference in the engineering literature as being 1892, and in the economic literature as 1911. Crew *et al.* (1995) summarise the evolution of the various *models* used to solve the peak load pricing problem.

approach to consider the impact of stochastic (i.e., random) demand, and found that off-peak periods should include some contribution to the ‘marginal rationing costs’, which for an optimal system can be expressed in terms of marginal capacity costs (Crew and Kleindorfer, 1976). Moreover, they proved that the presence of uncertainty in demand requires the level of optimal capacity to exceed the efficient level of capacity in the corresponding deterministic case, as Boiteux (1949) had surmised earlier. Similarly, Panzar (1976) also found that where the short run costs of operating individual plant exhibit *decreasing* returns to scale, plant optimal production requires excess capacity in every period, including that of the peak. (Although both Turvey and Wenders had relaxed the assumption of homogeneous capacity, they had assumed that each heterogeneous plant itself exhibited constant returns to scale with respect to operating costs). Finally, even Steiner (1957), although ignoring capacity constraints and assuming a single proportional cost technology, had found that relaxing the fourth assumption—by allowing for the interdependence of demands—also required off-peak consumers to contribute to capacity costs.⁷ Nevertheless, even within the last decade, some economists have still associated peak load pricing theory with the prescription that peak consumers should pay *all* capacity costs (e.g., Banks, 1994).

In practice, peak load pricing became implemented in the electricity supply industry through pre-published ‘time-of-use’ (TOU) rates. In their simplest form, TOU rates related only to a single peak and off-peak period, but the number of periods could be increased to incorporate electricity price changes on a daily, weekly, and/or seasonal basis. This generated a substantial body of research in both the economic and engineering literature from the 1970s through to the early 1990s.⁸ As Morgan and Talukdar (1979) observed, the decision to implement TOU pricing requires: “a year’s worth of appropriately designed and executed load research, development and justification of appropriate rates, completion of an analysis of the potential costs and benefits of the proposed rates, and, if an implementation decision is reached, the completion of a thorough customer information and education program”. Although Joskow and Noll (1981, p. 17) have lauded TOU pricing as “the one great practical triumph” of monopoly pricing theory, to some, such approaches—predominantly as implemented by vertically-integrated power utilities—suggest a return to “central planning” (e.g., Black and Pierce, 1993).

⁷ The complexities involved in modelling cross-time elasticities, have meant that this has been a relatively neglected area. Nevertheless, David and Li (1993b) have provided a rigorous theoretical model for resolving this problem within the context of spot pricing theory (discussed below). They separate consumers into those who optimise their behaviour with respect to current prices only, and those who also consider forecasts of future prices. However, the key problem with developing such a model is determining the appropriate elasticities.

⁸ This research encompasses: the determination of period prices (e.g., David *et al.*, 1986; Monts, 1991); implementation case studies, particularly relating to the use of enhanced metering technology (e.g., Chambers *et al.*, 1990; Sheen *et al.*, 1994); analyses of the response of various consumer classes to TOU pricing, both theoretically (e.g., Aigner and Leamer, 1984;

If so, then somewhat ironically, the intensive analysis of costs and prices associated both with peak load pricing theory, and the implementation of TOU power tariffs, may have itself provided a spur to the deregulation of the electricity supply industry. Even though the capacity costs of individual plants might be subject to economies of scale, the marginal costs of generating electrical energy were clearly *above* average costs for long periods. For instance, in their early model, Crew and Kleindorfer had shown that when welfare is maximised, utility revenue could far exceed its total costs. Similarly, Munasinghe and Warford (1982, Ch. 1) considered that implementing strict LRMC pricing would result in a financial *surplus*. Consequently, it was even suggested by some authors (§4.1.6) that, for a vertically-integrated utility, the rents extracted from pricing based on marginal generating costs could be used not just to cover the costs of capacity associated with energy generation, but also to cover the capacity costs associated with energy conveyance (i.e., the transmission and distribution network). But such a result no longer lent support to the presumption that power generation in its own right was a naturally monopolistic activity.

TOU pricing reached its zenith in the theory of spot pricing—expounded comprehensively by Schweppe *et al.* (1988) and tracing its origins to initial work by Vickrey (1971)—and this theory generated a huge subsequent body of research, particularly in the engineering literature.⁹ When taken to its extreme, spot pricing encompasses price variation in both space and ‘real time’, rather than simply a discrete number of periods, and in parallel with developments in metering technology, the theory was instrumental to the establishment of the core component of many modern deregulated electricity markets—the ‘power pool’ (§2.2.2).

Nevertheless, marginal cost pricing was not originally associated with proposals relating to the deregulation of electricity markets. By contrast, it was part of *public* utility pricing theory. Interestingly, even Schweppe *et al.* (1988, pp. 111 and 126)—who were possibly more responsible than any other researchers worldwide for the transformation to competitive wholesale power markets—only noted

Chamberlin, 1992) and econometrically (e.g., Caves *et al.*, 1984; Park and Acton, 1984; Wirl, 1991); and assessments of the welfare benefits of implementing the approach (e.g., Wenders and Taylor, 1976; Aigner, 1984; Parks and Weitzel, 1984).

⁹ Much of this research focuses on the issue of consumer response to spot (or real time) prices, but includes work on: theoretical refinements (e.g., Farmer and Bannister, 1987; David and Li, 1991a-c; McDonald *et al.*, 1994a); theoretical response of large customers (e.g., McDonald and Lo, 1990; McDonald *et al.*, 1994b); industrial consumer case studies (e.g., David *et al.*, 1986; Neal *et al.*, 1987); commercial consumer case studies (Daryanian *et al.*, 1991a-b; Daryanian and Bohn, 1993); and residential consumer case studies (e.g., Uusitalo and Yrjölä, 1990; Renz, 1993). Outhred *et al.* (1988) attempted to compare global welfare under spot pricing, spot pricing combined with forward contracts, as well as TOU tariffs. They indicated that pure spot pricing can be considered “inequitable”, because a monopoly supplier is rewarded with higher profits during outages (§4.1.4), and thus has no incentive to be dynamically efficient; a conclusion also reached by Hunt (1994). Nevertheless, Outhred and his colleagues concluded that welfare is in most cases the same for spot pricing with and without forward contracts, but would be higher than under a comparable TOU tariff scheme.

deregulation as an “excursion on the main journey” of their work, stating that “the spot price based energy marketplace is designed to operate in a regulated environment (regulated private company, or government owned)”. They considered that the introduction of the electricity marketplace required the utility and its customers to be viewed as a single integrated system, and that due consideration needed to be taken of the engineering requirements for controlling, operating and planning that system as a whole. Consequently, they made it clear that they did not advocate deregulation, and were not clear “whether there is ‘a lady or a tiger’ behind the door” of deregulatory policies. To make spot pricing work in a deregulated and fully competitive environment requires marginal costs to be revealed by generators bidding into the pool. As such, spot prices can only be considered “quasi-SRMC”, and poorly designed pool operation and/or generation sector structure, may invite (and has invited) criticisms of “gaming” and the “abuse of market dominance”—a concern which Schweppe *et al.*, (1988, p. 117) themselves expressed.

4.1.3 The Short Run versus Long Run Marginal Cost Debate: The Case for LRMC

The application of marginal cost pricing techniques to electricity supply has often involved an implicit assumption that information sufficient for both optimal consumption and investment decisions—relating to the entire supply chain from generators through the transmission and distribution system to consumers—can be packed into a single kWh-based price. Further, this single indicator supposedly allows both allocative and dynamic efficiency considerations to be satisfied. A symptom of the problems inherent in this attempt to squeeze information relating to both short run operational and consumption decisions, as well as long run decisions on expanding generation, transmission, distribution and end-use capacity, all into a single price, has been the short run versus long marginal cost pricing debate. The peak load pricing literature did not resolve this debate, which arises because—as noted above—the SRMC and LRMC of electricity provision are never equivalent in reality. Under peak load pricing, the marginal costs of operation and the marginal costs of capacity are both typically calculated separately and distinctly, and then, depending on the underlying assumptions, variously allocated to peak and off-peak periods.

Berg and Tschirhart (1995) conclude that the reason for the tension between SRMC and LRMC pricing is that, in reality, the two approaches focus upon completely different objectives. Short run marginal cost serves as the benchmark for *pricing* decisions, and thus *allocative* efficiency, while long run marginal costs are relevant for comparing alternative *investment* patterns, and therefore relate to *dynamic* efficiency. For instance, Wiseman (1957) criticised marginal cost pricing principles in general as providing no basis for a decision on the length of the “planning period” to which marginal costs are applicable. In discussing Wiseman’s critique, Blaug (1990) concludes that Wiseman implied that the strict application of marginal cost pricing principles would always have to be supplemented by an exercise in investment planning. This was in fact Boiteux’s (1949) own conclusion: “the principle of sale at marginal cost is applicable to existing plant, but cannot alone govern investment policy. It is

investment policy which must decide plant expansion and reconcile the apparently incompatible needs of long-term and short-term demand”. Nevertheless, Boiteux felt that oscillating prices caused by the “accidental excess capacity” associated with discontinuous investments in indivisible assets was not acceptable. Critics would later interpret this as an endorsement by Boiteux of LRMC-based pricing.

Turvey (1969) felt that the definition of marginal cost as the first derivative of cost with respect to output is too simple to be useful. Both cost and output have time dimensions, and both may be subject to uncertainty, hence cost structure should depend on *current* techniques and factor prices alone, only in the case of greenfields plants (or industries). Where a supply industry already exists, Turvey felt that “a cost analysis which is to be useful in decision-making needs to be historical dynamics, not comparative statics”. Turvey observed that consumer decisions in response to electricity price are often long run ones, particularly where they involve purchasing durable goods (e.g., manufacturing plant for industrial consumers, or whiteware for residential consumers). Therefore, he considered that the appropriate marginal cost concept for electricity pricing relates to “permanent” changes in demand and, as a consequence, output. Prices therefore should act as *signals* regarding *future* conditions.

Incentives should be directed at tomorrow’s consumers rather than today’s, except when today’s problems are particularly acute (Turvey and Anderson, 1977, p. 217).

The crux of Turvey’s (1969) recommendation for efficient pricing of electricity was that “the cost of producing any given flow of output can be reduced by postponing the period in which delivery is made”, and therefore “marginal cost is the present worth of the cost consequences of bringing forward the installation of new capacity and of postponing the scrapping of old capacity”.¹⁰ As Cichetti *et al.* (1977, p. 8) rephrased it: “moving an expansion plan forward or back amounts to the same thing as adding or removing small increments of capacity”. Long run marginal cost is thus formulated by determining the difference in the present worth of the future stream of all costs associated with producing an additional unit of output, since a (permanent) change in the level of current output is considered to alter the future optimal capacity expansion plan. The LRMC is thus the excess of the present value of the costs of the future system configuration with the permanent increment in output, over the present value of system costs for the system configuration with output *postponed* to the following year. In Joskow and Schmalensee’s (1983, p. 89) view, there was no alternative. Unless electricity prices tracked long run marginal costs, there would be “poor incentives for long-run decisions”. Consequently, the prescription for LRMC-based pricing in electricity supply became a staple of the majority of standard electricity economics texts during the 1970s and 1980s (§4.1.1).

¹⁰ To some extent, Turvey’s (1969) approach foreshadowed much later work on the ‘option value’ of irreversible investments under uncertainty (§5.3.4), and the link between option value and depreciation (§7.1.1). Moreover, Turvey’s views on marginal cost were not dissimilar from that held by Baumol at this time (§7.2.1).

4.1.4 Critiques of LRMC Pricing and Advocacy of SRMC Pricing

One of the many early criticisms of Turvey's prescription for electricity pricing came from Kay (1971), who maintained that "the concept of marginal cost relevant to tariff policy is essentially a short run one", and judged that Turvey's support for a long run concept of pricing arose primarily because it provided relative price stability in contrast to a strict SRMC approach. Yet Kay concluded that: "the most suitable method of deriving optimal prices is as a solution to some specified maximisation problem in which the relevant constraints and assumptions are made explicit, not as the by-product of some arbitrary definition of marginal cost". Such a conclusion was actually similar to Turvey's (1969) own verdict that pricing decisions require: "a specification of the objective function which is to be maximised and of the decision variables which are involved. Whether the objective function is a private profit or some concept of social gain, it is clear that some of the decision rules will involve marginal cost. The nature of the marginal-cost concept required can thus be ascertained only with reference to the objective function, to the constraints and to the amount of information available". Turvey, rather than being a dogmatic adherent to some strict abstract LRMC principle, was well aware of the complications arising from the particular cost attributes of a specific technology's cost function as well as cost and demand uncertainty.¹¹

Turvey and Anderson (1977, pp. 355-358) suggested that the rules for optimal resource allocation should comprise a *pricing rule*—price should equal the higher of marginal operating costs and the price necessary to restrict demand to capacity—and an *optimal investment rule*—the costs of a marginal increment in capacity should equal the present value of the expected benefits from it. The misunderstanding of Turvey's work appears to have arisen because Turvey considered that the price

¹¹ For instance, Joskow (1976) suggested that Turvey focused on "fixed demand". However, Turvey (1969) noted that "When uncertainty concerning demand is coupled with uncertainty in production, cost minimisation ceases to be a simple concept. ... It is difficult to imagine an analysis which, for example, contains all the complications relevant both to the electricity supply system and to the manufacture of chocolate. Thus, what is needed for decision-making in any particular industry is a cost model specific to that industry, and an understanding of the industry's technology is obviously required for that purpose". Moreover, Turvey's affirmation that SRMC and LRMC are equal when capacity is optimally adjusted to demand neither implied that capacity had to be "perfectly" adjustable, nor that indivisibilities had to be assumed away, nor that the marginal-cost based price was solely a long run concept: "Marginal cost as defined here appears to be a long-run concept. ... Nevertheless, it would be misleading to describe the concept as a long-run one. ... Optimal adjustment means that no conceivable change on the way the planned output is to be produced will lower the present worth of all future costs. Now one such conceivable change is to postpone some new capacity for a year and to work existing capacity harder the while. To say that this does not pay is simply to say that the marginal-cost saving from not producing extra output from new capacity just equals the marginal cost incurred from producing extra output from existing capacity. Thus, although it is formulated in terms of a permanent output increment, it turns out that marginal cost also relates to temporary output increments".

fluctuations resulting from the application of these resource allocation rules might be unacceptable.¹² As such, Turvey (with Anderson) recast the investment optimisation problem in order to find a pricing rule which would maximise the present worth of consumer willingness-to-pay for electricity, less the present worth of all costs, subject to the constraint that a *uniform* price (or at least a relatively *stable* price) must be charged over a considerable period of time. The resultant price from such a formulation turns out to be the weighted average of marginal costs over the relevant planning period, and Turvey and Anderson suggested that for many practical purposes it would be possible to simplify this result and use the *annuitised* value of average incremental costs as a basis for pricing policy.

The Swedes Andersson and Bohman (1985), reviewed the earlier contributions to marginal cost pricing theory of Boiteux, Brown and Johnson (1969), Munasinghe and Warford's (1982) key text, as well as Crew and Kleindorfer's various publications, and provided a strong critique of LRMC pricing (§4.1.1).¹³ Similarly to Kay, Andersson and Bohman consider the requirement of a relatively stable price to be an *ad hoc* constraint: "it is actually astonishing that Turvey's recommendations have been so widely accepted by the economic profession, which over the years has argued so much in favour of flexible exchange rates and so much against fixed rates". They made the somewhat reasonable objection that: "it is difficult to see why peak-load pricing due to seasonal variations is acceptable whereas price variations corresponding, for instance, to cyclical variations over two consecutive years are out of the question".

Furthermore, like Brown and Johnson before them, Andersson and Bohman considered that as soon as an investment is made, it becomes an indivisible, irreversible and durable unit (i.e., sunk). Consequently, LRMC becomes an irrelevant concept in price setting: "In cases with indivisibilities, the concept of LRMC cannot be operationally defined. The approximation used in reality is that the calculated total annual costs for a marginal plant are divided by the expected annual production. This is by definition not a marginal cost concept. It is the *average* cost per kWh for a marginal plant that is something quite different".¹⁴ Andersson and Bohman thus concluded that LRMC boils down to nothing more than average cost pricing in practice, and interpreted most proponents of LRMC pricing as incorrectly basing their advocacy on the pre-supposition that SRMC and LRMC are equivalent—an assumption not relevant to real world lumpy electricity markets in disequilibrium. On the other hand,

¹² For instance, Turvey (1969), more as a matter of practicality given the constraints of the times (i.e., the "institutional hurdles that have to be gone through") considered that: "prices or tariff structures cannot be changed frequently but have to be fixed for two or three years".

¹³ Munasinghe (1990) reiterated the support for LRMC pricing of electricity originally outlined in Munasinghe and Warford (1982), but did not address the criticisms of Andersson and Bohman and others.

¹⁴ Teplitz-Sembitzky (1992, pp. v-vi) also considers that calculating LRMC as a "scalar measure defined as a ratio of annuities not only is a far cry from the concept of marginal costs but is also useless in the design of multiproduct tariffs". He points out (p. 30) that annuity-based pricing satisfies the duality between pricing and depreciation decisions, but it remains an open question whether annuity-based pricing is welfare optimal (an issue addressed in §7.2.3).

Andersson and Bohman considered that LRMC can become a relevant concept in the formulation of an investment rule when an increase in capacity is under consideration, since a customer's demand for a marginal kWh of energy can hardly be pointed to as the reason for building new lumpy plant. They proposed that Williamson's (1966) approach to investment decision-making be applied, namely that it is necessary to calculate the total willingness-to-pay from a collective of customers for the total output of such a lumpy investment over its expected lifetime, in order to see if it is more or less than the investment's total costs. This is equivalent to Turvey and Anderson's (1977) optimal investment rule—marginal capacity costs equal marginal benefits (i.e., consumer surplus)—and this approach to evaluating investments was also adopted by the spot pricing theorists (Schweppe *et al.*, 1988, p. 241).

Sharing Kay's views, Andersson and Bohman agreed that price is a short run concept. Andersson and Bohman disagreed with the view held in common by Boiteux, Turvey, as well as Munasinghe and Warford (1982, e.g., p. 27), that the current price needs to signal future long term price changes in order for consumers to appropriately invest in durable goods: "Boiteux was himself aware that to reach an optimum use of an unintentional over-equipment, SRMC pricing ought to be used. But he was too occupied by the idea that such a 'low' price will be used by the customers in forecasting future prices and lead to wrong investments in appliances. It is, however, quite possible for customers to pay, for instance, the actual oil price and still predict a higher one in their forecast for investments. It is hard to see why such rational investment behaviour would not also characterise electricity consumers". Tabors (1994), for one, even felt that LRMC-based pricing is actually more unpredictable than SRMC pricing.

While primarily rejecting LRMC pricing on the grounds that it relies on accurate forecasts of future demand and investment, whereas short run marginal costs are directly measurable—and therefore SRMC pricing will be more reliable and less distorting—Della Valle (1988) concurred with the view that consumers did not require signals about the future to be embedded in the current price. She noted that LRMC proponents implicitly assume that there is no way to transmit information about future costs other than by setting current prices to reflect those future costs, and stated that: "There is no dichotomy between pricing correctly in the present and pricing correctly in the future". Simply telling consumers what to expect in the way of future prices would not result in the dynamic inefficiencies that would be caused by recovering future costs in advance. Similarly, Andersson and Bohman deemed consumers capable of basing their longer-term investment decisions on the prediction of future prices, and that long term contracts could resolve the information problem for those customers who would be willing to specify in advance their time profile of energy consumption. Such an approach to electricity pricing is embodied in spot pricing combined with long term hedge contracts, and is basically the framework underlying the operation of many of today's competitive wholesale electricity markets (§4.2.5). Weisman (1991) tried to reconcile both approaches and concluded that SRMC is the first-best marginal cost measure for spot market sales—since the buyer commits to purchase only after capacity costs have

been sunk—whereas by symmetric reasoning, LRMC is the first-best measure of any such long term contract sales. (The shortcomings inherent in such a viewpoint are discussed later; §5.3.2).

Like peak load pricing, spot pricing theory—which is generally associated with SRMC pricing taken to its extreme—had its origins with non-economists. Possibly as a consequence, in its early formulations in the engineering literature (e.g., Schweppe, 1978; Schweppe *et al.*, 1980), marginal cost pricing, either derived from short run or long run cost, was by no means a required basis for the spot price, but only one approach from a range of options. Where marginal cost was considered as a possibility, this “would be calculated by means of long-term system expansion studies”, implying an LRMC approach. However, spot pricing’s proponents grappled with the dilemma that “capital charges would tend to be much higher if based on future replacement costs rather than on historical construction costs” (§7.3.3).

Caramanis *et al.* (1982) tightened the presentation of spot pricing theory considerably, and asserted that it provides for both optimal short run pricing and long run investment decisions. However, efficient long run signals were considered to be provided by the anticipation of future spot prices, and any uncertainty regarding future prices could be handled through risk hedging via long term ‘forward contracts’ (e.g., Kaye *et al.*, 1990). Proposed investments should be subjected to the standard test that the cost of the investment is less than the expected value of net revenues, discounted back to the present year. The prices themselves were to be derived from current short run costs, adjusted upward for the value of unserved energy (§4.5.1), and for any power system constraints. Such constraints were attributed to generation and transmission capacity, as well as to voltage (and system stability) limits. Effectively, the theory was now a version of SRMC pricing modified to take account of constraints on demand specific to modern power system operation and—like Della Valle (1988)—Schweppe *et al.* (1988, Ch. 3) came to criticise models of LRMC pricing as being conditional on a particular expansion scenario, and on a predefined probability distribution of demand and factor prices. On the other hand, Schweppe and his colleagues suggested that a proxy for the “network quality” component of the spot price—the component designed to cover the unserved energy and constraint adjustments—could be the annualised marginal cost of network expansion, and this was seen as being a pragmatic and reasonably accurate method for LRMC calculation (Schweppe *et al.*, 1988, p. 253).

Notwithstanding their comprehensive theoretical articulation of spot pricing theory, Schweppe and his colleagues took a fairly pragmatic approach to actually putting theory into practice. Perhaps evincing their engineering background, they stated that: “Arguments on the definition of the true spot price can be left to university professors who need to publish academic papers in order to maintain or advance their professional status”. Nevertheless, the academic literature on spot pricing and real time pricing has swelled to vast proportions, and has to some extent superseded the work on peak load pricing, particularly in relation to electricity generation. However, at its most sophisticated (e.g., Crew *et al.*,

1995), peak load pricing theory—like spot pricing theory—has similarly concluded that the price in each period should be set equal to SRMC, plus expected rationing costs, plus the marginal “disruption costs” of power outages (§6.1.1).

4.1.5 Long Run Incremental Cost Pricing and Opportunity Costs

A completely polar critique of LRMC pricing comes from Schramm (1991). Like his World Bank colleagues before him—namely Munasinghe, and Anderson (the co-author with Ralph Turvey, not to be confused with Andersson, co-author with Bohman)—Schramm also favoured including “long run” costs in the price of electricity.¹⁵ However, Schramm argued that, although both the traditional LRMC and SRMC schools of thought were right in certain aspects of their claims, both were wrong in the application of their respective principles. Schramm stated that the efficient allocation of resources requires that their users be confronted with the *full* costs to society of using those resources, and that this is the fundamental principle underlying the marginal cost pricing doctrine. Although consumers should be confronted with the costs incurred at the margin, the costs at the margin form only part of the overall costs of supply and therefore of the total ‘*opportunity costs*’ to society of supplying the service. Schramm says that the traditional arguments attempt to turn a partial equilibrium argument into a general equilibrium one, without taking into account the totality of opportunity costs. Where the use of a resource in the production of a good precludes its use in the production of an alternative good, then the opportunity cost to society of the resources is its value in the ‘*best alternative use*’ (e.g., Hay and Morris, p. 27). As Joseph Stiglitz explains:

When rational firms make decisions—whether to undertake one investment project rather than another—they take into account *all* of the costs, the full opportunity costs, not just the direct expenditures (Stiglitz, 1993, p. 44).

Schramm’s objective was to ensure that the desirable principles of “*forward looking* marginal cost pricing” were preserved, while avoiding the high opportunity costs of over- or under-pricing non-marginal components of total supply. His solution was to determine the ‘*long run incremental cost*’ (LRIC), which would take into account the operation costs of the existing system, the costs of new capacity additions, as well as the ultimate *replacement costs* of the system components that already exist. All costs (as well as future demands) would be discounted back to the present. In recognition that such

¹⁵ One of the first signs of dissent from within the World Bank to the adherence to LRMC pricing principles is apparent in the work of Teplitz-Sembitzky (1990 and 1992). He concluded that “long-run marginal costs (LRMC) are a misleading benchmark for electricity pricing” (Teplitz-Sembitzky, 1992, p. v). On the other hand, he did not come down firmly in favour of SRMC pricing either, taking the more pragmatic view that “pricing has to be relieved of sacrosanct efficiency objectives and should come to grips with more mundane and immediate commercial ends”. Nevertheless, his two texts provide an extensive summary of different approaches to electricity pricing, and as such are drawn upon frequently in the subsequent (and earlier) Sections of this Chapter.

an approach might draw the same criticisms to which LRMC pricing was subject—namely that it is an average cost approach—Schramm countered that the proposal took account of the total future costs of supplying the service, whereas average cost pricing is based on historic costs. Under Schramm’s proposal, historic costs would only intrude upon the proposed pricing scheme if past financial obligations for former investments had to be covered, but would not be fully covered from the portions of the price collected and set aside for future replacements. However, this was only seen as occurring should there be a mismatch between the repayment period of the borrowed funds used for capital investments, and the lifetime of the associated assets (§6.4.3). Such obligations could not be considered sunk because if they were not paid any shortfall would have to be covered through some means, leading to the possibility of opportunity costs elsewhere in the economy.¹⁶

4.2 Revenue Reconciliation

4.2.1 *The Revenue Reconciliation Problem*

In discussing marginal cost pricing, Panzar (1976) found that, unless the long run technology (i.e., capacity) is linearly homogeneous, total revenues will not equal total costs even if capacity is optimally matched to demand. This ‘*revenue reconciliation*’ problem has been recognised by authors on both sides of the SRMC versus LRMC debate, and is a key feature of the traditional ‘natural monopoly problem’ (§2.1.5). However, those authors who appear to have viewed electricity supply as a public sector activity (e.g., Andersson and Bohman, 1985), do not seem to have shown much concern. On the other hand, although the original spot pricing theorists, ever the pragmatists, stated that “the world would be nicer if revenue reconciliation could be ignored”, they bravely attempted to face up to the problem anyway, by including a revenue reconciliation term in their spot price equation. Schweppe *et al.* (1988, Ch. 5) presciently recognised that, in a deregulated power market, this problem would only relate to capital cost recovery for transmission and distribution (§4.2.5). Pricing transmission or distribution network services at marginal cost would most likely result in a loss, but they also noted that the problem is exacerbated when revenue recovery is reconciled on an *annual* basis.¹⁷

¹⁶ Banks (1994), took issue with all these more recent contributions to marginal cost pricing theory, contending that neither Andersson and Bohman, Della Valle, Weisman, or Schramm had actually proven their contentions regarding SRMC and LRMC, but merely stated their point(s) of view: “Where reality, in the form of the ... electricity market, does not fit the neoclassical models that ... economists studied as undergraduates, a call goes out for an alteration of reality, despite the well-known, and extensive, shortcomings of the neoclassical paradigm”.

¹⁷ Schweppe *et al.* (1988, p. 248) observed that, for *new* investments, revenue reconciliation should not be a problem. Investments in capacity should be made until the last increment of capacity earns an expected stream of new income whose present value equals the incremental cost of investment—in other words, they were advocating the standard optimal investment rule (§4.1.4). With the exception of uncertainty, all other complications, such as indivisibilities, are explicitly included in such a framework. This does not guarantee that some under- or over-recovery of costs will not occur, since investment decisions are always based on *expected* revenue values.

Schweppe and his team suggested various methods of allocating revenue shortfalls across consumers, and a weighted least squares averaging approach was favoured on pragmatic grounds. However, for investments in new dedicated network assets, they proposed that the associated user should cover all costs. Two of the many solutions to the revenue reconciliation problem proposed by Schweppe and his colleagues can be considered as being unpalatable to accountants (as well as regulators), and to economists, respectively. The first approach, which was their preferred approach, was to use a “revolving fund” which recognises that, economically, revenue should cover cost over the *lifetime* of the project, and not from year to year (§6.1.3). The other approach arose from the fact that a basic presumption of spot pricing theory was that it applied to a vertically-integrated utility. The economic architects of New Zealand’s power sector reform would most likely shudder to read the proposition made by the spot pricing theorists—most of them engineers—that the excess revenue of a utility’s “Generation Department” could *subsidize* the loss-making “Network Department”. Such a proposal goes against the entire thrust of New Zealand’s reforms, since concerns relating to vertical cross subsidisation were the reason behind the legal separation of line and energy businesses (§2.4.5).

4.2.2 ‘Second Best’ Ramsey Pricing

In Blaug’s (1990) view, marginal cost pricing is inherently a ‘*second-best*’ problem, and “must involve the problem of maximising output in the presence of an added constraint”. Blaug declared that, the views of the early advocates of marginal cost pricing that Pareto optimality requires prices to be equated to marginal costs “is nowadays dismissed as extraordinarily naïve”. Blaug criticised those advocates of marginal cost pricing who “treat historic costs as bygones that are forever bygones” and “insist on keeping questions of allocation and pricing analytically separate from questions of finance and equity”. The second best case for marginal cost pricing requires not that price *equals* marginal cost, but that price *deviates systematically* from marginal cost. As noted earlier (§2.1.7), the contestability theorists entirely agree with such a position.

No regulator can be expected to follow the precept of marginal-cost pricing that is integral to the model of perfect competition, for to do so would either drive the regulated firm into bankruptcy or force government permanently to subsidize the resulting deficit. If the model of perfect competition cannot offer the regulator useful guidance on price regulation, it is virtually worthless as a model for an agency charged with regulating prices (Baumol and Sidak, 1994a, p. 34-35).

Boiteux came up with a solution to the static revenue reconciliation problem that was later formalised in collaborative work led by Baumol (i.e., Baumol and Bradford, 1970), and this solution was generally applicable to the problem of insufficient returns for a natural monopoly. (The dynamic case of the problem is presented by Braeutigam, 1983). Acknowledging the parallels between this work, and research performed on optimal taxation by the philosopher/mathematician Frank Ramsey in the 1920s, Baumol and his colleagues labelled the approach ‘*Ramsey pricing*’. The simplest form of the Ramsey pricing rule occurs when cross-elasticities of demand are zero. Revenue is raised to a level sufficient to

cover costs by setting each consumer's price so that its percentage deviation from the true marginal cost-based price is inversely proportional to the product's price elasticity of demand; accordingly, the approach is often termed the 'inverse elasticity rule'. Consumers incurring the same supply costs could therefore face different prices—those with highly elastic demands would face prices close to marginal cost, whereas those with highly inelastic demands would be faced with much higher charges. In its most general form, where demands are interdependent, Ramsey pricing can be shown to Pareto dominate any other linear price schedule (e.g., Crew *et al.*, 1995), particularly straightforward average cost pricing, and is therefore widely considered to be the 'second best' approach to strict marginal cost pricing.

Where different classes of consumers can be characterised by different elasticities, what appears to be price discrimination (§2.2.3) can, in theory, be allocatively efficient. However, it is acknowledged that measuring elasticities in practice is extremely problematic, and that the situation is even more complex where cross-elasticities of demand are non-zero. Consequently, the application by regulators of Ramsey pricing in practice has been limited due to concerns—such as those held by the Interstate Commerce Commission in the United States—about the level of data and degree of analysis required. This problem has been acknowledged by Baumol himself in later works, including his standard contestability text (i.e., BPW, pp. 504-506; Faulhaber and Baumol, 1988; Baumol and Sidak, 1994a, p. 38).¹⁸ The spot pricing theorists derived the conditions for applying Ramsey pricing to spot pricing, but rejected its application in real networks for similar reasons.

Furthermore, Schweppe and his colleagues considered that Ramsey pricing only accounts for short run elasticities of demand, and that, in the long run, these elasticities would be very different (Schweppe *et al.*, 1988, Ch. 8). And Sharkey (1982a, p. 102) concluded that “the Ramsey rule is primarily a tool that is useful in the context of centralized planning in a well-defined industry, where competitive entry is not a serious policy question”. Hence, for better or for worse, Ramsey pricing arguments have in the past been more often used to justify price discrimination inherent in *existing* electricity tariff structures, rather than for calculating *new* tariffs (§4.3.6)—both internationally (e.g., Scherer, 1977, p. 39; Baumol and Sidak, 1994a, p. 39; Loube, 1995; Heald, 1997), as well as in New Zealand (e.g., Kask, 1988b, p. 30). In practice, however, some research on the US power industry has suggested that where regulators have allowed different consumer classes to be charged different prices, these prices have borne no relation to even the relative price levels that would be attributed to implementing Ramsey pricing (Nelson and Roberts, 1989).

In any event, these “second best” approaches to pricing did not resolve the SRMC versus LRMC debate either. For instance, Della Valle (1988) upholds that the appropriate second best solution is to set

¹⁸ By contrast, the Federal Communications Commission in the US rejected Ramsey pricing on the basis that implementing the method would violate existing US law—namely, the *Telecommunications Act 1996* (Sidak and Spulber, 1997, p. 43).

price above short run marginal cost according to the Ramsey pricing rule. By contrast, Black and Pierce (1993) recommend that regulators allow utilities to price electricity at long run marginal (social) cost—“social” meaning the inclusion of externalities from environmental damage—adjusted via Ramsey pricing methods to ensure cost recovery. They also expect that, in a competitive market, most utilities would voluntarily adopt this form of Ramsey pricing because doing so would “maximise their revenue”. As Berg and Tschirhart (1995) observe, “short-run versus long-run considerations confuse the issue” of implementing Ramsey pricing, and there are those who are sceptical of using the approach, given the limitations of knowledge regarding short and long run demand elasticities (e.g., Meyer and Tye, 1985).

4.2.3 Multiple Component Tariffs and Cost Allocation Methods

Another approach to the revenue reconciliation problem is the use of a ‘two-part’ or *non-linear* tariff to generate enough revenue from marginal cost pricing to cover both operational and capacity costs. In fact, regardless of the theories behind electricity pricing—which have often treated electricity supply as a bundled product recovered through a single ‘linear’ kWh-based price—in practice, electricity tariffs usually comprise multiple components. Historically, the concern regarding the ability of installed generating capacity to meet peak demand led to the introduction of the ‘demand charge’ (measured in kW) to supplement the energy charge in kWh (Neufeld, 1987). But capacity charges in kVA (to cover dedicated asset costs), as well as fixed administrative charges (to cover general overheads), were also often applied, making the final electricity price a tariff with up to three or four components, an approach that dates back to the late nineteenth century (e.g., Cichetti, 1977, p. 88).

Spot pricing theory proposed that the demand charge be dispensed with, since the kW component was simply a proxy for signalling the need to reduce demand as the system peak rose toward the limits of available capacity (Schweppe *et al.*, 1988, s3.7). The problem with demand charges was that, due to the effects of diversity (§3.1.1), a particular consumer’s peak consumption might occur at a different time from the system peak, potentially sending a perverse signal to the consumer to cut back on consumption at a time of surplus capacity. Spot pricing—designed as a ‘closed feedback’ approach to pricing—increases the consumer’s kWh-based price at the actual time of the system peak, and therefore addresses the capacity limit issue directly. Although examining how a “non-linear” (i.e., polynomial, rather than two-part) price structure would fit into their spot pricing framework, Schweppe *et al.* (1988, pp. 195-198) did not pursue the approach because they felt it complicated the analysis “without resulting in any proven real-world advantages”. But spot pricing already was by default a two-part pricing structure, given the inclusion in the spot price of the revenue reconciliation term (§4.1.6).

Nevertheless, this revenue reconciliation term in the spot price was introduced as a pragmatic, rather than an optimal solution to the problem. Similarly, traditional multi-part tariffs were not striving for optimality, but were simply a way of sending some “signal” to consumers regarding the need for investment in additional capacity, while at the same time recovering “allowable” costs. Traditionally, the

United States in particular either regulated the allowable level of revenue, or directly specified which “prudently incurred” costs—including a “fair” rate of return on investment (§6.3.1)—were allowed to be recovered by the power utility (Teplitz-Sembitzky, 1990, p. 22). Traditionally, those costs were then assigned to consumers on the basis of one of many arbitrary or “rule-of-thumb” cost allocation rules, collectively termed ‘fully distributed cost’ (FDC), or ‘fully allocated cost’ methods (e.g., Braeutigam, 1980). But as Heald (1994, p. 11) observed, economists have become “scathing” of FDC methods, who see them as contributing to cross subsidies and, perhaps even more disturbing, as being “an accountant’s method” for pricing (e.g., Brown and Sibley, 1986, p. 49; Baumol *et al.*, 1987; Baumol and Sidak, 1995a, p. 64; Sidak and Spulber, 1997, pp. 42-47).¹⁹ Interestingly, given the approach which the contestability theorists recommend for efficient pricing (§4.3.1), FDC methods have also been criticised as leading to prices which are “economically inefficient”, on the basis that “the practice focuses heavily on cost and little on conditions of demand” (Braeutigam, 1989).

4.2.4 Optimal Two-Part Pricing

The current theoretical literature on *optimal* two-part pricing was initiated in a paper by Oi (1971), but traces its roots to the earlier notion of lump-sum recovery of revenue deficits, which dated back to the late 1930s. Building on Oi’s work, Willig (1978) demonstrated that, as long as arbitrage was not possible between consumers, a non-linear tariff—comprising a fixed entry fee and a variable usage charge—could Pareto-dominate the linear Ramsey pricing approach. As Berg and Tschirhart (1995) point out, the ability to separate markets and prevent resale is intrinsic to utility distribution systems, which implies that a two-part tariff is preferable to a linear tariff for bundled distribution services.

According to Spulber, the solution to the revenue reconciliation problem is straightforward. Natural monopoly pricing is associated with an ‘internality’—a term coined by Spulber (1989, p. 54) himself—namely, the monopolist does not account for the loss of consumer surplus associated with not pricing its product at the competitive or marginal price. If the monopolist were taxed by an amount equal to the internality, the seller will choose the competitive price. The resultant two-part tariff will allow perfect price discrimination and is non-distortionary, since the consumer faces the short run marginal cost of its consumption decisions, leading the monopolist to the optimal output level. The internality is caused by *linear* pricing.²⁰

¹⁹ Blaug (1990), however, points out that such criticisms of FDC go back much further, to Vickrey (1955), the *de facto* father of real time pricing (§4.1.2).

²⁰ In the New Zealand context, prior to the unbundling of retailing from distribution, Bertram *et al.* (1992, pp. 137-143) discussed the SRMC versus LRMC debate and concluded that the problem is best resolved through a two-part electricity tariff, where the variable component is linked to the SRMC.

Consequently, various two part tariffs schemes have been designed to resolve the revenue reconciliation problem.²¹ Many of these schemes incorporate cost allocation methods which were originally used to solve the revenue reconciliation problem under a linear tariff. The general issue of joint costs of production generated its own body of literature, which stemmed both from the application of game theory to problems of cost allocation and cross subsidies (e.g., Peyton Young, 1985a). Curien (1991) describes that there are various approaches to cost allocation, those derived from economic considerations (the origins of which he attributes to Boiteux), from game theoretic concepts (e.g., Sharkey, 1982b), or from an ‘axiomatic’ approach to cost allocation (e.g., Mirman *et al.*, 1983). Curien recounts how the rules resulting from such approaches have led to cost-allocative prices such as: ‘Boiteux-Ramsey prices’ (second best linear prices which maximise welfare subject to a budget constraint); ‘Shapley prices’ (incremental costs averaged over all possible orderings of outputs, e.g. Littlechild, 1970a); and ‘Aumann-Shapley prices’ (marginal costs averaged over a linear path from zero to current production). Peyton Young (1985b) describes the Aumann-Shapley cost allocation procedure as “the only method that attaches no penalty to diligence, and no reward to negligence”. Nevertheless, Curien points out that there is a huge gap between the theoretical cost allocation instruments and the pragmatic costing methods that have actually been recommended by regulators. Curien attributes this to the need for “acceptability” of cost allocation methodologies. Theoretical considerations of efficiency or welfare take second place to ensuring that prices are “agreeable to all the players concerned”. A second reason is that the cost concepts developed by theorists are extremely difficult to calculate from empirical data.

Under two-part prices, the axiomatic approach is only applied to the fixed charge component required to recover cost over and above the revenue collected through strict SRMC pricing. Spulber (1989, pp. 237-8), for one, comes down firmly in favour of the modified Aumann-Shapley approach—as developed by Mirman *et al.* (1983)—for allocating the fixed component of the two-part tariff. Joint costs (§3.4.2) are allocated to prices on the basis of the weighted average of product marginal costs. Nevertheless, in summing up the literature on cost allocation, Heald (1994, p. i) states that “it is important to stress that the academic literature on cost allocation is overwhelmingly normative in design and prescriptive in its conclusions: the algorithms, however elegant, often have little in terms of behavioural or motivational underpinnings”.

4.2.5 A Partial Resolution of the Revenue Reconciliation Problem and the SRMC vs. LRMC Debate

For power generation at least, the revenue reconciliation problem has mostly been overtaken by best-practice reforms to the electricity supply industry. It appears that pricing at quasi-SRMC in power

²¹ The drawback of permitting a two-part tariff, is that if the monopolist is not successfully restricted to just recovering total costs, the two-part scheme allows the monopolist to obtain even greater ‘quasi-rents’ than would be the case for a profit-maximising monopolist using a linear price (e.g., Schmalensee, 1981; Spulber, 1993).

pools can allow (productively and dynamically efficient) generation companies to recover both their capital and operational costs (§4.1.2). In presenting spot pricing theory, Schweppe *et al.* (1988) foresaw that this would be found to be the case, once the generation market were deregulated. Consequently, the SRMC versus LRMC debate has also been resolved by default. There is no need to determine what the allocatively efficient generation prices will be, since competitive wholesale electricity markets are, by definition, competitive.

On the other hand, the risk inherent in a constantly fluctuating pool price is typically mitigated on both sides of the market—particularly the supply side, given its dependence on expensive durable assets—through parallel arrangements involving long term bilateral supply contracts and/or hedge contracts in electricity futures. Turvey and Anderson (1977) were correct in the sense that fluctuating prices are a cause of concern (§4.1.4). Since not all the wholesale electricity market will be tied up in such contracts, the marginal unit of electricity consumed will still be (mostly) priced at the marginal cost of generation, because this cost is the (approximate) basis for the pool’s moment-to-moment spot price.

Consequently, wholesale electricity markets strive to satisfy the Pareto optimality criterion (§2.1.2), although it could be argued that the allowing such parallel markets only provide a *potential*, rather than an absolute, Pareto improvement.²² This is because the long term contracts will produce both winners and losers, depending on which consumers obtain the most inexpensive long term contracts, and which generators obtain the contracts with the most substantial net revenues. Nevertheless, depending on how such contracts are structured, it is quite possible that consumers can increase their demand at the time of the peak and not incur any additional charges. As the International Energy Agency (IEA, 1991, p. 45) has cautioned: “stable pricing is the antithesis of stable markets”.

4.2.6 Issues of Electricity Distribution Pricing

Notwithstanding this partial resolution of some of the theoretical pricing debates in relation to power generation, the deregulation of the electricity supply industry has not entirely resolved the revenue reconciliation problem, nor the SRMC versus LRMC debate, when it comes to considering the remaining naturally monopolistic parts of the electricity supply industry—transmission and distribution. Moreover, electricity distribution has been the poor cousin of transmission in terms of academic attention (§1.1.2).

²² A potential Pareto improvement—also termed the ‘Kaldor-Hicks criterion’—is one where gainers *could*, but do not necessarily, entirely compensate losers for their losses, and still remain better off than they were previously (e.g., Hay and Morris, p. 572). Not all, however, agree with the philosophy behind the criterion. For example, Sidak and Spulber (1997, p. 220) suggest that the: “Kaldor-Hicks criterion was in perfect synchronicity with the metamorphosis of American constitutional law during the New Deal, a transformation that curtailed protections of contract and property and gave the central government virtually unlimited regulatory powers over economic activity”. The implication is that the criterion can be used against firms. Zajac (1985) cautions against the approach in part for the opposite reasons, although he suggests in a general sense that “a policy that attempts to approximate Pareto improving moves is not ethically harmless”.

Nevertheless, unlike many authors, issues relating to the application of marginal cost and peak load pricing techniques to electricity distribution had been given just as much thought by Boiteux as he gave to generation and transmission (§4.1).

Like Crew and Kleindorfer (1975 and 1976)—as well as other authors such as Brown and Johnson (1969), and Carlton (1977)—Boiteux and Stasi (1952) had pointed out that electricity demand is stochastic. Consequently, as Joskow (1976) expresses it, Boiteux and Stasi recognised that, while the generation and transmission systems are in a sense “common” systems, as one moves down through the distribution system from the transmission grid to the point of end use, networks—and their associated costs—become more and more “individual”. While the uncertainty in the demand profile reduces with distance from the consumer—due to the probabilistic averaging of demand across larger groups of consumers (i.e., diversity; §3.1.1)—the demand characteristics of individuals, or of small groups of consumers, are not necessarily coincident with system demand patterns.

Such individuality suggests that consumers should pay none of the costs of the supply system *downstream* of their specific location. Boiteux and Stasi proposed that this could be ensured by a ‘nodal pricing’ method of assigning and propagating marginal costs downward through the power system. This approach was later taken up by the developers of spot pricing theory, and nodal pricing techniques are now widely applied for transmission pricing (e.g., Tabors, 1994). The key intention, though, of applying nodal pricing to transmission systems has been to send strong signals regarding the optimal network locations for new generators, as well as to reflect any current bottlenecks in the transmission system and the long run costs of reinforcing the system. But, as Outhred and Kaye (1994) conclude, there may be limitations of extending nodal pricing to distribution networks. The smaller scale of the networks may mean that the activities of many participants are large enough to have significant effects on the nodal price.²³ They note that co-operative decision-making between consumers is desirable because of the economies of scale that exist in the distribution sector, and believe that “the calculation of meaningful spot and forward prices at nodes in a radial distribution network is problematic”.

Since Boiteux, the economics literature has not had much more of substance to say about the specific issues of distribution pricing. For instance, one of the traditional texts on marginal cost pricing in electricity supply declared that the transmission and distribution system can be considered as a single integrated system, both performing the function of bringing electrical energy from generators to consumers (Cichetti *et al.*, 1977, p. 19). This ignored the important, and distinct, co-ordination and stability role that is played by the transmission grid. Nevertheless, Cichetti *et al.* (1977, p. 95) did discuss how to evaluate marginal capacity costs in distribution networks: “Enormous parts of the system are comprised of distribution facilities, the costs of which, in large part, are not marginal at all (in the

²³ This conclusion was based on earlier work on generation presented in Kaye and Outhred (1989).

sense that they do not vary with the load on the system). In isolation, this aspect of utility operation would suggest decreasing, not increasing marginal costs”. Although acknowledging that the marginal capacity costs of the distribution network were dependent on the cost of energy losses, these were seen as being “small”. Consequently, Cichetti and his co-authors even suggested that, for those hours of the year when it is reasonably certain that increases in demand will not affect the distribution capacity construction schedule, “a zero marginal capacity cost is appropriate” (Cichetti *et al.*, 1977, p. 32).

Boiteux (1956) had also recognised that the SRMC of electricity conveyance is the marginal cost of energy losses (§3.1.1). Consequently, given that losses exhibit *decreasing* returns to scale—typically the opposite situation to capacity (§3.4.2)—the sale of network conveyance service for a particular line at the marginal cost of losses will involve deficits when the line has substantial spare capacity, but will be profitable when the line requires expansion. As such, any power line’s optimal loading level will never be at its maximum capacity. This result is consistent with the work of Panzar (1976), discussed above (§4.1.2); it is never efficient to “drive the marginal product of costly factors to zero”.²⁴ Energy losses can be viewed as an ‘internality’ of electricity conveyance (Spulber, 1989, p. 54) exhibiting operational diseconomies of scale (i.e., pertaining to operating rather than capacity costs), and Boiteux considered that adjusting for this internality allows the pricing of network services to be efficient. (Such diseconomies are much stronger in network equipment than is the case for generating plant). Nevertheless, Boiteux assumed away the problem that revenues taken on this basis might not exactly cover the “fixed” capital cost outlay of the line, by ignoring indivisibilities. The way to deal with indivisibilities, thought Boiteux, was to calculate the marginal cost of losses based on what *would* be the optimum capacity, if capacity were perfectly adjustable.

Consequently, notwithstanding recent advocacy of SRMC pricing for essential facilities such as power distribution networks (§4.1.1), applying strict SRMC pricing techniques to electricity distribution will result in a loss (e.g., Farmer *et al.*, 1995). Thus, as the spot pricing theorists also recognised (§4.2.1), the revenue reconciliation problem is a key issue for electricity distribution. Nevertheless, some attempts to apply SRMC nodal pricing techniques to electricity distribution have ignored the reconciliation problem entirely (e.g., Murphy *et al.*, 1994). On the other hand, it is not too surprising that distribution pricing also has its advocates of an LRMC-based approach. Although not referring to LRMC directly, Woo *et al.* (1995) define ‘*marginal distribution capacity cost*’ as the present value of the difference in costs between a capacity expansion plan initiated immediately, and the same plan delayed

²⁴ Panzar (1976) points out that there is no reason why the costs of capacity cannot exhibit increasing returns to scale, whereas the operation of that capacity is subject to diseconomies of scale. Such is the case for most items of transmission and distribution equipment, if the SRMC of operation is considered to be the marginal cost of losses.

by some unit of time. This value is also defined as the ‘*avoided cost*’ of capacity expansion, and is effectively the same as Turvey’s definition of “incremental system costs” or LRMC (§4.1.3).²⁵

A further problem remains, even were capacity optimally adjusted, and SRMC and LRMC equivalent, there may be joint costs due to shared inputs (§3.4.2). By definition, joint costs are non-separable, and some cost allocation rule must be developed to assign them to various consumers (§4.2.3). The increasing downstream “individuality” of a distribution network makes the association of costs and services difficult. As Sidak and Spulber (1997, p. 412) observe in a more general sense: “it is simply more difficult to identify the attributable costs of a particular service as one moves toward higher levels of disaggregation in the classification of services”. The problem is exacerbated by the non-equivalence of SRMC and LRMC in reality. For instance, Heald (1994 pp. 19-20) indicates that attempts to implement marginal cost pricing customarily reveal that there are powerful tensions between a policy based upon SRMC and one based upon LRMC, particularly when it comes to considering ‘*cross subsidies*’. If cross subsidy is to be measured with reference to benchmarks defined in terms of optimal pricing policy, Heald indicates that the choice between SRMC and LRMC is clearly of importance. With an SRMC benchmark, consumers who pay for marginal operating costs are not being cross subsidised. With an LRMC benchmark, consumers are expected to contribute toward marginal capital costs, otherwise there may be some basis for a cross subsidy claim. And the problem with cross subsidies, is that they are generally deemed “inequitable” or “unfair”.

4.3 “Third Best” Subsidy-Free Pricing

4.3.1 *Equity and “Fairness” in Pricing: Price Discrimination and the Removal of Cross Subsidies*

The issue of “equity” used to appear more frequently in debates associated with electricity pricing than is currently the case. For instance, Weintraub (1970) rejected the outcome of Steiner’s conventional model of peak load pricing—namely that peak period consumers cover all capacity costs—on the grounds that it was an inequitable result. Apart from the “efficient” delivery objective, the New Zealand Government’s current policy statement for the electricity industry requires that energy services be delivered to consumers in a “fair” manner (§2.3.3).²⁶ Unfortunately, however, no definition of “fairness” has been provided by the New Zealand Government).²⁷ Simply from a practical point of view,

²⁵ Again, not surprisingly, there are those who point out that Woo *et al.*’s (1995) definition of the marginal distribution capacity cost is not truly marginal. Lesser and Feinstein (1999) point out that the investments in distribution assets are lumpy, therefore, derivatives of the cost function are not well defined, and the resultant value cannot be considered truly “marginal”.

²⁶ Earlier, the more general economic statement relating to markets dominated by natural monopolies had referred to “the fair and efficient conduct of business” (§2.3.1, fn. 35).

²⁷ Sidak and Spulber (1997, p. 496) provide the following cautionary comment about such a state of affairs: “although economists may consider the definition of ‘fair competition’ to be an oxymoronic undertaking, it is nonetheless necessary to supply regulators with an operational definition of fairness that does not attempt to specify outcomes”.

Heald (1997) for one considers that introducing the issue of fairness explicitly into cost allocation exercises erodes the transparency of the process. But more significantly, Blaug (1990) provides the reminder that the separation of questions of efficiency from those of equity is a fundamental tenet of the policy prescriptions arising from welfare economics (§2.1.2).²⁸ Nonetheless, Blaug muses that: “Efficiency is necessarily a value-laden concept, and cannot be freed from the notion that it is somehow more desirable than inefficiency”. Like Turvey (1971), this dissertation sets thorny questions such as those of income distribution aside, and assumes that this is the responsibility and function of a branch of policy unrelated to the power sector.²⁹

From the point of view of the *firm*, “fairness” is typically associated with the notion that prices allow them to receive “fair return” on their investment, in comparison to alternative uses of capital employed with commensurate levels of risk (e.g., Klevorick, 1971; Sidak and Spulber, 1997, p. 499). This “fair” level of return—within the framework of economic efficiency—is that which would be achieved by a firm in a perfectly competitive (or perfectly contestable) market, with real world adjustments for risk. From the *consumer* perspective, the narrow notion of ‘horizontal equity’, or equity *between* consumers (e.g., Berrie, 1992, pp. 151 and 165) lies firmly within the framework of allocative efficiency, and this theme is a key aspect of the New Zealand Government’s power sector reforms (§2.2.3 and §2.3.3). The concept of horizontal equity generally reflects a pejorative view of price discrimination as being implicitly unfair.

However, the second best form of allocative efficiency described above (§4.2.2), may actually require price discrimination (e.g., BPW, p. 475), since Ramsey pricing will assign different prices to two consumers who face the same costs, but have different demand characteristics (particularly with respect to their elasticities). This highlights a key problem inherent in any assessment of “fair” or equitable prices. Allocative efficiency and horizontal equity not only depend on a comparison of prices, but a comparison of costs and consumer *demand* characteristics as well. Given the complexity involved in such a task, price discrimination justified on the basis of elasticity estimates can be subject to challenge from consumers (and regulators). As Heald (1994, p. i) discerns: “Injunctions to avoid ‘undue

²⁸ Blaug (1990) goes on to say that: “If we refuse, even in principle, to distinguish allocative efficiency from distributive equity, we must perforce reject the whole of welfare economics and with it any conventional presumption in favour of competitive markets—and, indeed, in favour of the price mechanism as a method of allocating scarce resources. Arguments for coordinating economic activity by markets would then have to be expressed in terms of political philosophy—for example, that markets diffuse economic power—and economics would in consequence have to become a totally different subject”. An eloquent discussion of equity and distributional issues in the context of the regulation of utilities is provided by Zajac (1985). Also see Baumol’s (1986) own discussion of “superfairness”.

²⁹ As Hay and Morris (1993, p. 569) indicate, focusing on efficiency requires making the standard “capitalist assumptions” that the distribution of income is determined by the initial distribution of factor ownership, and by the prices of factors that are thrown up by the competitive system.

discrimination’, which are typical features of statutory arrangements, sit uneasily with the kinds of price discrimination which are legitimised by Ramsey-style optimal pricing rules”.

The proponents of spot pricing theory (Schweppe *et al.*, 1988), cited “equity” as a valid pricing objective. However, they were also referring to equity in the sense of reducing or removing ‘*cross subsidies*’ between different consumer classes. Because setting prices in a manner consistent with the absence of cross subsidies generally—although not always—provides bounds within which second best prices can be found, Teplitz-Sembitzky (1990, p. 40) states that the principle of subsidy-free pricing is at least as good as a “third-best” approach to welfare maximisation. Yet, notwithstanding Blaug’s comment above, Baumol and his colleagues note that, at least prior to the development of contestability theory, arguments before regulatory agencies against cross subsidy rarely were based on the grounds of efficiency of resource allocation, rather they were discussed as matters of distributive equity (BPW, p. 354). However, New Zealand’s power sector reforms have been primarily implemented within a framework of economic efficiency, and the removal of cross subsidies has been a key objective from the outset (§2.3.1). Apart from the equity or fairness implications, the presence of cross subsidies has clear efficiency implications as well.

Cross-subsidies cause economic losses by distorting customer decisions because prices fail to convey accurate signals about costs. Services that receive the subsidy are priced inordinately low, which encourages excessive purchases of that service and displaces potentially more efficient alternative products or services. Services that generate the subsidy are priced overly high, which discourages purchases of that service and can lead customers to seek alternatives that would otherwise not be purchased (Sidak and Spulber, 1997, p. 520).

Heald (1997), who has written extensively on the subject of cross subsidies, provides an exhaustive taxonomy of cross subsidies. He defines eight cases of cross subsidy, grouped: (a) *within regulated sectors*, such as those (i) between outputs which are bundled together in a vertically-integrated industry structure (e.g., between power generators and electricity retailers, §3.4.2, fn. 36); (ii) when uniform tariffs apply across geographically differentiated supply zones (a pricing approach frequently taken by electricity distributors, §4.4.3); (iii) when different categories of consumers of the same output are treated differently in economically unjustified ways—(b) *between regulated sectors*, such as those (iv) between outputs which are bundled together in a horizontally integrated industry structure (e.g., within a combined energy utility selling both electricity and natural gas)—(c) *between regulated and unregulated sectors*, such as those (v) when producers of a monopolised output bias their choice of suppliers towards their own associated companies (e.g., between ELBs and energy retailers, prior to their legal separation, §2.4.5); (vi) when a regulated entity subsidises a potentially competitive activity (e.g., again, the pre-split of ELBs and retailers); (vii) when a competitive activity subsidises a regulated activity (noted by Heald as uncommon); and (viii) when the enterprise is compelled by government or regulator to commit resources unprofitably to activities unrelated to its own business.

In particular, Heald draws attention to cases (i) and (iv), which relate to the structure of the regulated sector, and to cases (ii) and (iii), which raise fundamental issues about the basis of pricing policy. Case (vi) is the main focus of regulatory attention in the telecommunications sector, and the literature on the regulation of network industries is firmly weighted towards this sector. As such, much of the literature cited in this and the following Chapters is drawn from work related to the telecommunications industry. Nevertheless, the focus in this thesis is on the third category of cross subsidies, those which are associated with economically unjustified actions, since this category is of most relevance to the current state of New Zealand's electricity distribution sector. This is particularly the case if not all of a particular ELB's market is a natural monopoly while the remainder is structurally contestable. To fend off competition for potentially contestable consumers—such as those large consumers with a dedicated network connection (§3.2.3)—an ELB can cross subsidise the line charges for those consumers by increasing the charges faced by “captive” consumers connected to the natural monopoly part of its network (§3.2.1).

With respect to this third category of cross subsidy, Heald suggests that claims regarding the presence of cross subsidies typically relate to: (a) the existence of common or joint costs across outputs (§3.4.2), and methods for allocating such costs particularly between different groups of consumers; and (b) the existence of monopoly power, which may be due to economic factors (such as cost and demand conditions), to political factors (such as the granting of exclusive franchises), or to some combination of these. Unfortunately, “third best” pricing experiences some of the same complexities as second best pricing. Heald (1994, p. i-ii) expresses the problem as follows: “it is virtually impossible to construct unequivocal and uncontested benchmarks of cost allocation for purposes of determining whether there is cross subsidy in particular cases. Even were such exercises feasible, their rationale would be challenged by economists who dispute that statements about cross subsidy can be made in isolation from demand conditions”. Baumol and Sidak would agree (1994a, p. 50), having stated that prices calculated without the use of demand relationship data are apt to be inconsistent with economic efficiency, and are all too likely to damage economic welfare, rather than to help it.

The contestability theorists, however, have a solution to this problem that both costs and demands must be considered in determining economically-efficient prices. The solution involves rejecting the perfectly competitive market model as the appropriate paradigm for efficient price setting (§2.1.7). By setting aside the benchmark of perfect competition, the contestability theorists also manage to sidestep the entire SRMC versus LRMC debate, since both SRMC and LRMC are derived from static concepts of the short run and long run associated with the model of perfect competition.

In their tract on competition in telecommunications networks, Baumol and Sidak (1994a, p. 50) like to suggest that the benchmark of a perfectly contestable market has almost entirely replaced the perfectly competitive model. On this basis, they provide a “floor-ceiling solution” to the efficient pricing

problem for regulated firms, known as ‘*constrained market pricing*’ (§5.1). Baumol and Sidak state that “the solution recommended by most economists engaged in the formulation of regulatory practice”—and which has actually been applied by regulators in some industries, notably telecommunications—is to divide the task into two parts. The next sub-sections describe Baumol and Sidak’s recommended approach for achieving the first part of the task—the regulatory evaluation of bounds on ‘*subsidy-free prices*’. By contrast, the second part of the task requires no action by regulators.

The first [part] consists of imposing constraints upon the setting of prices by the firm—constraints derived from the [contestable] market model ..., and which, fortunately, can be expressed in the required quantitative terms with the aid of cost information alone. The second part of the price-determination process is then left to management in the regulated firm, whose self-interest will lead it to take demand conditions into account. The regulated firm is prohibited from selecting any prices that violate the cost-based constraints adopted by the regulator; but within those limits the firm is granted the freedom to select the prices that best promote its interests (Baumol and Sidak, 1994a, pp. 50-51).

4.3.2 *Subsidy-Free Prices: the Stand Alone Cost (SAC) and Incremental Cost (IC) Tests*

The Baumol group’s approach to achieving efficient prices that would recover a natural monopolist’s total costs is firmly grounded on Gerald Faulhaber’s (1975) seminal paper regarding cross subsidisation in public enterprises. Faulhaber defined a subsidy-free price structure as one where the provision of any product or group of products by a multiproduct enterprise, making a ‘normal’ economic profit, leads to prices for all other products no higher than if those products were the only goods to be produced. Faulhaber interpreted positive economic profits as a subsidy of producers (or investors) by consumers, and negative profits as the reverse situation. Therefore the overriding subsidy-free constraint is that total revenues (TR) must equal total costs (TC)—which include compensation for the firm’s cost of capital—and thus the firm earns a normal or zero economic profit. (Hence, the revenue reconciliation problem is automatically resolved in Faulhaber’s approach). For the firm defined earlier (§3.4.4), this overall revenue/cost equality is as shown in (4.1).

$$TR = p \cdot Q(p) = C(Q) = TC \quad (4.1)$$

Faulhaber stated that the set of price vectors which are subsidy-free must lead to revenues for each subset of products no greater than the ‘*stand alone cost*’ (SAC) for that subset. For any subset of the set of products N (i.e., $S \subseteq N$), with the associated price and quantity vectors p_S and Q_S , then the price vector p is subsidy-free only if it satisfies (4.1) as well as (4.2) below, which is termed the ‘*stand alone cost test*’.

$$p_S \cdot Q_S(p) \leq C(Q_S); \quad \forall S \subseteq N \quad (4.2)$$

Combining (4.1) and (4.2), results in the ‘*complementary*’ expression (4.3) below, the right hand side (RHS) of which is defined as the ‘*incremental cost*’ (IC) of the subset $N - T$, and is termed the ‘*incremental cost test*’. (This ‘*complementarity*’ between the IC of a set of products and the SAC of all the other products, and vice versa, is a key characteristic of incremental costs and stand alone costs).

$$p_T \cdot Q_T(p) \geq C(Q) - C(Q_{N-T}); \quad \forall T \subseteq N \quad (4.3)$$

Where a firm attains a normal profit, products or groups of products that contribute revenues to the firm greater than their associated SAC of production, are considered to provide a subsidy to other products. Conversely, products or groups of products that receive a subsidy, are those which generate revenues less than the IC of providing the product or products which they consume. For the contestability theorists, the SAC and the IC values thus form the basis of the efficient ‘*price ceiling*’ and ‘*price floor*’ within which the firm should be free to set its own prices (Baumol and Sidak, 1994a, p. 51). It is important to note however, that Faulhaber’s constraints apply to *revenues*, and not to *prices*. (In the following Sections, when the term ‘subsidy-free prices’ is used, generally this term is used to refer to a set of subsidy-free bounds on revenues). When considering the firm as a whole, the SAC and IC tests together result in the total revenue/cost equality.

Applied to the full set of products supplied by the firm, [the SAC] rule dictates that the firm’s total revenues must not exceed its total costs. This is the other half of the requirement [i.e., the IC test] that the firm be permitted to obtain earnings that equal but do not exceed the competitive earnings level (Baumol and Sidak, 1994a, p. 78).³⁰

4.3.3 *Anonymously Equitable and “Fair” Prices*

Faulhaber’s initial approach to subsidy-free prices (i.e., revenues) actually evaluated cross subsidies between *goods*, rather than between *consumers*, and Governments, regulators, as well as consumers themselves, are more concerned with inequitable pricing between consumers or consumer groups (e.g., industrial to residential and vice versa; rural to urban, and so on). By expanding on this original formulation of subsidy-free prices, and also research by Willig (1979a), Faulhaber with a colleague (i.e., Faulhaber and Levinson, 1981) presented conditions for ‘*consumer subsidy-free*’ prices.³¹ Faulhaber and Levinson took the original formulation, and allowed those consumers who consume products in the product subset S , to also consume products in the subset $N - S$. They defined $M = \{1, 2, \dots, m\}$ as the set of consumers, and $Q_j(p)$ the j th consumer’s demand vector of products produced by the firm. For any subset of consumers $V \subseteq M$, Faulhaber and Levinson defined the product

³⁰ The complementarity of SAC and IC still holds in this case. The complementary set of *all* products is the null set. The SAC of the null set is zero, and therefore the incremental cost of providing all products is equal to the total cost. Similarly, the SAC of all products is also the total cost.

³¹ The term “consumer subsidy-free” is generally attributed to Sharkey and Telser (1978).

vector $Q_V(p) \equiv \sum_{j \in V} Q_j(p)$ and defined the price vector p to be consumer subsidy-free if (4.1) is satisfied, and (4.4) as well.

$$p \cdot Q_V(p) \leq C(Q_V(p)); \quad \forall V \subseteq M \quad (4.4)$$

Equation (4.4) is the consumer subsidy-free stand alone cost test. Similarly to the derivation of the incremental cost test for subsidy-free products, combining (4.4) with (4.1) provides the complementary consumer subsidy-free incremental cost test.

$$p \cdot Q_W(p) \geq C(Q_V(p)) - C(Q_{M-W}(p)); \quad \forall W \subseteq M \quad (4.5)$$

As long as the zero profit constraint (4.1) is satisfied, if all consumers have identical or even just proportional demand patterns (or consumption bundles) then all prices will be *consumer* subsidy free, regardless of whether the product prices are subsidy-free or not. The simplest example of this result is of a single consumer who purchases all the products. Clearly, as long as total revenues equal total costs, then even if there are cross subsidies between the products, that consumer will be indifferent to those cross subsidies.

Because in practice, information on individual consumer demands may not be available, Faulhaber and Levinson concluded that an equity concept related to a specific set of unobservable consumer demands is of little interest. Hence, they reformulated the consumer subsidy-free problem by requiring that any conceivable set of consumer demand vectors generates revenues no greater than its standalone costs. Thus for any arbitrary vector of demands $q = (q_1, \dots, q_n)$, where $q \leq Q(p)$, it is possible that there is a subset of actual consumers V , such that $q = Q_V(p)$. If every set of demands q generates revenues no greater than stand alone costs, then no possible consumer group could be paying more than it otherwise would. Thus, the corresponding ‘*anonymously equitable*’ SAC and IC tests are shown in (4.6) and (4.7) respectively. The property of anonymous equity was so-named by Robert Willig, because it holds regardless of the identities or the specific purchasing patterns of consumers.

$$p \cdot q \leq C(q); \quad \forall q \leq Q(p) \quad (4.6)$$

$$p \cdot q' \geq C(Q) - C(Q - q'); \quad \forall q' \leq Q(p) \quad (4.7)$$

If p is anonymously equitable, then it is also consumer subsidy-free for any demand pattern $Q_j(p)$ for which $Q(p) = \sum_{j \in M} Q_j(p)$. Anonymous equity therefore requires that any and all possible coalitions of consumers should pay at least the incremental cost to the industry that their purchases cause, but should pay no more than they would have had to, were they to produce those products themselves—namely the stand alone cost of those products. Further, if the price vector is anonymously equitable, then prices are no lower than marginal costs. Consequently, the concept of anonymous equity is seen as

satisfying the objective of “fairness” in pricing. For instance, in considering the separation of efficiency and equity considerations, Teplitz-Sembitzky (1992, p. 93) concludes that: “Unless convincing support can be given to the assignment of ‘social’ weights that tilt the balance in favor of (distortionary) redistributive measures, there is little reason to call the fairness criterion of anonymous equity into question”.

Faulhaber and Levinson proved that a price vector which is anonymously equitable will also be subsidy-free in Faulhaber’s original sense. This is because the subsidy-free price example is equivalent to the case where there are the same number of products and consumers, and each consumer exclusively purchases a single product. This is but one of the many consumption configurations which is included in the anonymously equitable set of consumption patterns, which includes all possible demand patterns. Conversely, a price vector which is subsidy-free need not be anonymously equitable, since it is only a special case. However, where the cost function is subject to *declining average incremental costs*, then Faulhaber and Levinson demonstrated that the two concepts are equivalent.

4.3.4 *Anonymously Equitable Prices in the Peak Load Pricing Problem*

To indicate the difference between subsidy-free prices, and those which are anonymously equitable, Faulhaber and Levinson revisited the conventional Steiner peak load pricing problem (§4.1.2). The revenue/cost equality for the problem is shown in (4.8), while the SAC test associated with the “peak product” (q_1), and the SAC test associated with the “off peak product” (q_2), are shown in (4.9) and (4.10) respectively (where P is revenue, per unit of demand).

$$P_1q_1 + P_2q_2 = \beta q_1 + b(q_1 + q_2) = C(q_1, q_2) \quad (4.8)$$

$$P_1q_1 \leq C(q_1, 0) = (\beta + b)q_1 = \text{SAC}_1 \quad (4.9)$$

$$P_2q_2 \leq C(0, q_2) = (\beta + b)q_2 = \text{SAC}_2 \quad (4.10)$$

It is important to note that the SAC tests are *not* formulated simply by substituting zero demand—for the product not being supplied—into the total revenue/total cost equation (4.8). For the off-peak product, such a substitution would provide an SAC of b alone which is clearly incorrect. If the off-peak demand were the *only* demand, capacity sufficient to serve q_2 would need to be constructed; thus, the SAC of serving off-peak demand must include some capacity costs. In general, stand alone costs for any subset of products must be evaluated in their own right, and not just on the basis of the cost function which relates to producing all products.³² As both Curien (1991) and Palmer (1991) point out,

³² Nevertheless, some fall into this trap. In presenting the subsidy-free approach to solving the peak pricing problem, Teplitz-Sembitzky (1990, pp. 37-40) performs the substitution directly into the overall market’s cost function, and therefore concludes that, for the prices in the conventional peak load pricing problem to be subsidy-free, peak demand should be

this makes the assessment of subsidy-free prices difficult in practice, since in many multi-product industry settings, data on the stand alone cost of producing each product (or group of products) is unavailable (as is information on demand behaviour), or even difficult to conceptualise.

On the other hand, incremental costs are not evaluated directly in Faulhaber's approach, but are always derived from the complementary set of stand alone costs. Neglecting this complementarity is not uncommon, as is shown later (§5.2-§5.3). For the conventional peak load pricing problem, subtracting the SAC expressions in (4.9) and (4.10) from the revenue/cost equality (4.8) provides the pair of corresponding incremental cost expressions in (4.11) and (4.12). The subsidy-free bounds on peak and off-peak per unit revenues are therefore as shown in (4.13) and (4.14). (Because the bounds are presented on a per unit basis, they can be considered 'average stand alone costs' and 'average incremental costs' respectively).

$$P_1 q_1 \geq C(q_1, q_2) - C(0, q_2) = (\beta + b)q_1 + bq_2 - (\beta + b)q_2 = \beta(q_1 - q_2) + bq_1 = \text{TC} - \text{SAC}_2 = \text{IC}_1 \quad (4.11)$$

$$P_2 q_2 \geq C(q_1, q_2) - C(q_1, 0) = (\beta + b)q_1 + bq_2 - (\beta + b)q_1 = bq_2 = \text{TC} - \text{SAC}_1 = \text{IC}_2 \quad (4.12)$$

$$\beta \left(1 - \frac{q_2}{q_1} \right) + b \leq P_1 \leq \beta + b \quad (4.13)$$

$$b \leq P_2 \leq \beta + b \quad (4.14)$$

Therefore, although the conventional peak load pricing result where peak consumers pay all capacity costs while off-peak consumers pay only energy costs, is consistent with subsidy-free prices, this is by no means the only set of subsidy free prices. It is important to note that the bounds on subsidy free per unit revenues in (4.13) and (4.14) must be satisfied simultaneously with the revenue/cost equality (4.8), and Faulhaber and Levinson provided a plot of the subsidy-free price locus curve which satisfies the revenue-cost equality with the bounds provided above.

Deriving the anonymously equitable prices for the conventional peak load pricing problem requires the introduction of significantly more constraints. In fact, there are an infinite number of constraints. These relate to various subsets of consumers which have demand in both peak and off-peak periods, although Faulhaber and Levinson showed that these constraints can be summarised by four sets of constraint expressions. The first pair of constraint sets relates to an arbitrary set of consumers removed from the market, and its complementary set of remaining consumers, for situations where removing those consumers does not result in the peak shifting to the previously defined off-peak period.

charged at average stand alone costs (i.e., capacity plus energy charge), while off-peak demand should be served at prices based on average incremental costs (i.e., energy charges alone).

The second pair of constraint sets relates to the situation where the peak does shift when an arbitrary set of consumers is removed. Faulhaber and Levinson proved that the bounds on anonymously equitable prices for the conventional peak load pricing problem are considerably more restrictive than the bounds on subsidy-free prices. In fact, the bounds reduce to an equality for each per unit revenue, one consistent with the solution to the conventional peak load problem, such that for anonymous equity to be satisfied, peak consumers must pay $\beta + b$, and off-peak consumers just b .

However, this restrictive result arises from the conventional peak load problem's assumption of linearly homogeneous capacity costs. By introducing a cost function that exhibits declining average incremental costs $K(Q)$, and by using the same approach as above, it is straightforward to show that the bounds on the subsidy-free and anonymously equitable (total period) revenues are as shown in (4.15) and (4.16), subject to the revenue/cost equality in (4.17).³³ There are still a large number of constraints however. Where there are not constant returns to scale, the number of constraints is combinatorially related to the number of products (e.g., Baumol and Sidak, 1994a, p. 71).

$$(K(q_1) - K(q_2)) + bq_1 \leq P_1q_1 \leq K(q_1) + bq_1 \quad (4.15)$$

$$bq_2 \leq P_2q_2 \leq K(q_2) + bq_2 \quad (4.16)$$

$$P_1q_1 + P_2q_2 = K(q_1) + b(q_1 + q_2) \quad (4.17)$$

4.3.5 *Subsidy-Free Prices: the Consumer Perspective and Game-Theoretic Approach*

Faulhaber (1975) had actually formulated the original subsidy-free pricing problem as a 'co-operative game' of various consumer coalitions demanding products from a natural monopolist, following earlier work on applying game theory to public enterprise pricing by Littlechild (1970a), among others. Consumers are considered to be free to form any and every possible coalition. The 'core' of the game is that set of subsidy-free prices which satisfies all constraints derived from the various SAC tests, including the global SAC equality requiring zero economic profit—as presented in (4.1)-(4.3) above. Significantly, Faulhaber demonstrated that there are subadditive cost functions for which no subsidy-free prices exist—in other words, there is sometimes no core to the game representing natural monopoly product provision. Such an outcome indicates that the cost function is 'unsustainable'. In the context of perfectly contestable markets, unsustainability means that an incumbent natural monopolist

³³ Like Faulhaber and Levinson (1981), declining average incremental costs are defined in the manner introduced by Panzar and Willig (1977), with variables as defined above:

$$C(Q_S, Q_T) - C(Q_S) < \frac{C(Q_S, \lambda Q_T) - C(Q_S)}{\lambda} \text{ for } 0 < \lambda < 1; \quad S, T \subseteq N, S \cap T = \Phi, Q_T \neq 0.$$

could not publish a set of prices for all products which would prevent subsequently inefficient entry (§2.1.8).

Faulhaber's original formulation, although appearing to take the consumer's perspective, matched up each consumer with but a single product, and therefore did not derive a truly anonymously equitable result. However, given the relationship between subsidy-free and anonymously equitable prices, anonymously equitable prices are subject to the same possibility of non-existence. On the other hand, Faulhaber and Levinson indicated that sustainability is a sufficient condition for anonymous equity to be achievable.

For instance, Spulber's (1989, p. 237) advocacy of Aumann-Shapley cost allocation methods (§4.2.4), arises from deriving optimal two-part tariffs as the outcome of a cooperative game of cost sharing between consumers served by a natural monopolist. If a '*second best core*' exists in a cooperative game of joint production under increasing returns to scale, then the Aumann-Shapley allocation of fixed costs—one where each price is the weighted average of marginal costs in all subcoalitions of consumers—will always lie in that core. Consequently, such an allocation will automatically satisfy the conditions of anonymous equity. Spulber suggests that, using such an approach, "regulators can design a market mechanism to achieve optimal prices without direct intervention in firm pricing decisions". His regulatory solution for cases of natural monopoly was franchise competition, in which firms compete to offer a service contract which satisfies entire market demand (§3.2.1). The general principle is that the process of firm solicitation of customers through service contracts, resembles the joint production of a number of goods by a coalition of consumers in a cooperative game. Given the possibility of recontracting, and existence of a core for the cooperative price-setting game, a firm service contract will secure an exclusive franchise if and only if the pricing policy is in the core of the associated cooperative game of joint production.

Sorenson *et al.* (1976) took such a similarly game theoretic approach to analysing the peak load pricing problem for *decreasing cost industries*. Sorenson and his colleagues ignored operating and marginal costs, and focused on the problem of allocating the fixed component of a two part tariff—namely, peak and off-peak *capacity*. Although their core takes the form of a multi-period version of that derived at the end of the previous subsection (with b set to zero), the key point to note is the way in which the problem has been articulated. Interestingly, Steiner (1957)—in presenting the conventional peak load problem—had thought that the best approach to allocating the cost of capacity was to consider consumers as acting in collusion as a monopsonist *buyer*. In fact, Sharkey and Telser (1978) took such an approach, and in doing so, introduced the concept of *consumer* subsidy-free prices.³⁴ The active players in their game are consumers, whereas firms pursue passive behaviour, simply being identified

³⁴ Also see Sharkey (1982a, pp. 102-110).

with the set of possible coalitions of buyers. By contrast, rather like Spulber (as well as Berg and Tschirhart, 1988, p. 456; and Palmer, 1991), Sorenson and his colleagues view consumers as “*self-producers*”—the consumer coalition is *itself* the market’s natural monopoly producer. In this approach, for the purpose of deriving equitable prices, the concept of the firm can be dispensed with entirely.³⁵

The equity issues associated with the cooperative peak pricing game were examined by Sorenson and his colleagues. Although the optimum level of capacity depends only on the peak consumer, they pointed out that co-operation can still be mutually beneficial. Given the fact that the industry in question is decreasing cost—it is subject to economies of scale or scope—peak consumers realise that if the excess capacity in off-peak periods is at least partially utilised by groups willing to pay in excess of operating cost, the charge to peak consumers can be lowered. Off-peak consumers, moreover, are willing to pay in excess of operating cost, since this is less than if they had to acquire the good independent of other users. Therefore, somewhat similarly to Panzar’s (1976) result, off-peak consumers do contribute to capacity costs (§4.1.2). But Sorenson and his colleagues consider what this outcome means with regard to the costs and benefits associated with the demand of particular consumers. When a peak period consumer is added to a coalition, Sorenson and his colleagues indicated that such a consumer will generally cause two things: additional benefits for existing coalition members, as well as for itself; and additional costs, to be allocated among the coalition members. However, a consumer whose only demand is in a firm off-peak period adds benefit without a corresponding cost increase. Hence, in the context of the game, this makes the off-peak consumer a very valuable player, and that consumer’s payoff must reflect this value.

4.3.6 Subsidy-Free and Sustainable Prices in Contestability Theory

Interestingly, the contestability theorists generally appear to view the issue of anonymous equity from the point of view of a firm supplying a group of consumers, a philosophical standpoint somewhat different from the consumer self-production perspective of Sorenson and his colleagues. For instance, in one collaborative paper, Baumol defines the SAC of a group of services to be “the cost of a hypothetical efficient entrant” serving those services alone (i.e., Faulhaber and Baumol, 1988).³⁶ Similarly, notwithstanding his declaration that the primary purpose of the price ceiling is to “protect consumers”, Baumol—in his work with Sidak on competition in telecommunications markets—defines the SAC of a subset of commodities (or even a single commodity) as the cost incurred by an efficient *entrant* to the

³⁵ Berg and Tschirhart (1988, p. 456) state that the price based on stand alone cost “is no higher than it would be if the group produced for itself”, and Palmer (1991) states that “a set of prices for the n different commodities exhibits cross subsidies if any group of consumers of a subset of the n commodities prefers self-production of that subset to purchasing the products from the single multiproduct producer at the stated prices”. Sorenson *et al.* (1976) did indicate that it would possible to add an “entrepreneur” to their formulation of the game, and in such a case assume that production could only take place if the entrepreneur were present.

industry in question, if it were to produce only that subset of commodities. In a perfectly contestable market, competitive behaviour is imposed upon the incumbent firm by the threat of entry, rather than by the threat that consumers will band together to produce the goods or services themselves (Baumol and Sidak, 1994a, pp. 51, 58 and 77). This is consistent with Faulhaber’s original approach, although under conditions of perfect contestability—where producers and consumers are implicitly ‘symmetric’—this might seem to make little difference.³⁷ However, in examining a real world market, the point of view taken can be significant (§6.4.2). Since regulatory restrictions on prices are intended to protect the consumer, while regulatory restrictions on entry are designed to protect the producer, it could be argued that the more appropriate perspective of not just SAC, but IC as well, is that of the consumer, rather than that of either the incumbent firm or potential entrants.³⁸

Irrespective of the relative philosophical merits of the producer or consumer perspective of cross subsidies, undoubtedly the concepts of subsidy-free and anonymously equitable prices form an integral part of contestability theory. In the standard contestability text, Baumol and his colleagues (BPW, pp. 348-354) cite managing cross subsidies as a key issue of public policy relating to efficient resource allocation, particularly should revenues fall short of the incremental costs of providing a product (or group of products). Furthermore, as noted earlier (§2.1.7), BPW (pp. 487-488) suggest that, prior to the exposition of contestability theory, probably the most “noteworthy” gap in the standard theory of policy was the lack of a defensible criterion for regulatory ceilings on prices.³⁹ Contestability theory demonstrates that, should revenues exceed the stand alone costs of providing a product (or group of products) then—in a perfectly contestable market, which ensures that firms earn only normal profits—a profitable entry opportunity is offered to potential competitors, and the firm’s price structure is not sustainable. Given the ‘symmetry’ between the SAC test and the IC test under conditions of perfect contestability, unsustainability implies inefficiency and cross subsidy (and vice versa), whereas sustainable prices are both subsidy-free and anonymously equitable. This is because, in a perfectly contestable market, prices that are not anonymously equitable would invite some consumers to self-produce, or to seek supply from another firm. Accordingly, BPW (p. 472) declare that: “Perhaps one of the more attractive features of sustainable configurations is their preclusion of hidden cross subsidies”.

³⁶ As Berg and Tschirhart (1995) observe, the *hypothetical* nature of the costs involved has limited the application of constrained market pricing in practice.

³⁷ Faulhaber (1975) states that “competitive entry, however, is the key idea of core constraints. ... The cross-subsidy approach focuses on the costs of alternative supply to determine a set of prices, all of which provide positive incentives to forestall inefficient entry, and does not directly address the welfare maximisation question”.

³⁸ Given the complementarity between SAC and IC under conditions of normal profit, prices below IC—which signal anti-competitive behaviour—must also indicate that some complementary monopolistic exploitation of consumers is occurring.

There are three key theoretical issues associated with the evaluation of the bounds on subsidy-free prices.⁴⁰ The first issue is the relationship between subsidy-free prices and second best prices. As part of their attempt to demonstrate that contestability is ultimately the “hero” of their piece (BPW, p. 479), Baumol and his colleagues presented their “weak invisible hand theorem” (originally from Baumol *et al.*, 1977). This theorem states that, under many circumstances, contestable markets will induce natural monopolists to offer prices that are both Ramsey-optimal and lie within subsidy-free bounds. In other words, prices will be no higher than the competitive earnings level (i.e., the price ceiling, or SAC). Comfortingly then, the “same invisible hand that forces Adam Smith’s perfectly competitive firms, each small in its market, to operate in a manner that is socially optimal also guides a natural monopoly enterprise to a Ramsey optimum” (BPW, pp. 357-358). Yet, Baumol and Sidak (1994a, pp. 52-54) acknowledge that while defenders of Ramsey pricing dub the method “value-of-service pricing”, its critics—such as Miller (1995)—refer to the method as “charging what the traffic will bear”.⁴¹

The fact that consumption in public utility markets is highly concentrated gives firms the ability to segment markets and differentially price, favoring consumers with higher price elasticities of demand over relatively demand-inelastic consumers. Moreover, supplier vulnerability to the exercise of monopsony power (including the threat of bypass) enables high-use customers to extract price reductions and to achieve network improvements that advantage these influential groups. Indeed, monopsony power appears to be behind many developments in public utility markets today. ... In electric markets, it seems to be a driving force behind steps to substitute price negotiation for price regulation, thus effectively achieving a system of selective discounting, and behind decisions to permit high-use consumers to pick and choose amongst sources of supply. This, in itself, gives to favored classes of consumers the ability to abandon—or threaten to abandon—the system and to leave fixed and sunk system costs the responsibility of low-volume consumers, who do not have this option (Miller, 1995).

Acknowledging work by both Sharkey (1982a, p. 101) and Spulber (1984), the Baumol group does however concede that Ramsey prices are not always sustainable, nor by inference, are subsidy-free (BPW, p. 209). Discussing this possible conflict between Ramsey and subsidy-free prices, Teplitz-Sembitzky (1992, p. 93) claims that, for a decreasing cost industry, unsustainable Ramsey prices are

³⁹ Ironically, Faulhaber (1975) attributes the concept, if not the term, of “stand alone cost” to Harry Trebing (1967), who has since become a critic of contestability theory and even of the application of the SAC test itself (e.g., Trebing, 1987 and 2000).

⁴⁰ Parsons (1998) provides a good survey of the literature relating to cross subsidy, and highlights some of these problems.

⁴¹ Hay and Morris (1993, p. 626) cite another objection to Ramsey pricing—even if it is seen as value-of-service pricing—that, consumers with no alternative, should not be expected to pay more on account of that fact, even though this results in their placing a high value on the product concerned (fn. 42).

welfare-superior to any set of unsustainable prices, all other things being equal. This might suggest that since Ramsey prices are *second* best optimal, the ‘*third* best’ subsidy free bounds can be ignored. Such a viewpoint is taken by Bös (1986, p. 194), for one, who considers the problem of cross-subsidisation as being “of no importance from the point of view of welfare economics. If optimal pricing includes any kind of cross-subsidization (of the Faulhaber type or of an extended type), then that cross-subsidization should be accepted”. However, as discussed earlier (§2.1.7), at least initially, Baumol and his colleagues affirmed that the correct benchmark even in a non-contestable market is perfect contestability. Echoing the earlier work, Sidak and Spulber (1997, p. 341) remark that, requiring prices to be below SAC is “superfluous” because “the laws of economics already prohibit a firm from charging a price exceeding stand-alone cost”. This is certainly true for a perfectly contestable market, since prices above SAC cannot exist because they would invite (unsustainable) entry.⁴²

The second issue is the effect of demand. Baumol and Sidak (1994a, p. 54) justify their regulatory prescription that demand considerations should be left to the firm, on the basis that firms have better information regarding demand conditions than the regulator ever will (i.e., the asymmetric information problem; §2.1.3). However, this ignores the fact that demand impacts the evaluation of the subsidy-free prices themselves; a problem already foreshadowed in the brief discussion of cost allocation above (§4.3.1). Heald (1997) points out that “a choice has to be made as to whether (and, if so, how) demand-side considerations should be allowed to influence cost allocation and, therefore, the benchmark from which cross subsidy is measured”. In his initial formulation of the cross subsidy problem, Faulhaber (1975) assumed that demand and supply levels remain fixed. However, like Baumol (1977), Faulhaber also considered the case with interrelated demands, in other words, where cross-elasticities of demand are not zero. Faulhaber found that in the presence of cross-elasticities, the revenues collected from a new service do not measure the true incremental revenues, since revenues from the other services may rise (if the service in question is a net complement) or decline (if the new service is a net substitute). The core of the game may thus be smaller or larger than a core derived assuming fixed demands. For example, if the goods or services are substitutes, then the bounds derived from assuming fixed demands

⁴² In his model of cross subsidies, Faulhaber (1975) assumed that consumer elasticities were based on demand for that product alone, ignoring alternative supply possibilities or any degree of competition in the market. Faulhaber admitted that the demand curve, as seen from the firm’s perspective, could in reality become much more elastic at high prices than such an assumption allowed, particularly should the price be sufficiently high to make alternative means of supply attractive to the consumer. Economic models typically use single and linear elasticity values, and thus fail to acknowledge that a change in demand can cause a change in elasticity—since elasticity is dependent on the current level of demand—and fail to allow for the possibility that even a linear elasticity may only be valid within certain bounds. An abrupt discontinuity may occur when a price change causes a consumer to switch entirely to a substitute product; one that might not even be in the same market. Consequently, the likelihood that the apparently welfare-maximising prices (above SAC)—derived from theoretical models—might be infeasible in reality, seems to have been ignored by Bös and Teplitz-Sembitzky.

only provide necessary, but not sufficient, constraints on subsidy-free prices.⁴³ Hence, prices which pass the IC or SAC test may involve a cross subsidy (although prices which fail either test will definitely involve a cross subsidy). Heald (1997) defines ‘*gross IC*’ as being exclusively cost based, while ‘*net IC*’ is the value adjusted for demand side repercussions of various supply side configurations—a term introduced by Baumol (1986, pp. 115-120).

The third issue is that, as Faulhaber (1975) himself recognised, it is possible for sustainable prices not to exist at all—in other words, that there be no core to the game. Following on from the above logic, if no sustainable prices exist, then neither do anonymously equitable prices. In a static universe, Baumol and his colleagues demonstrate that unsustainability is more likely the stronger the degree of substitution in the demands for the various products of the enterprise, the weaker the degree of complementarity in production among its outputs, and the greater its product-specific economies of scale (BPW, p. 357).

Returning to the earlier discussion of the characteristics of electricity distribution, it seems that the apparently strong economies of scope in distribution networks are likely to contribute to strong cost complementarities, rather than to weak ones (§3.4.2). Moreover, the level of substitution between the demands for capacity at different consumer locations intuitively would appear to be predominantly independent. Finally, although individual items of network equipment are subject to strong economies of scale, dedicated assets at a particular location typically serve a non-growing load, since demand for capacity primarily grows through the addition of new consumers at new locations. Hence, under such conditions of no product-specific demand growth, the product-specific economies of scale have little influence (§3.1.2). These factors would tend to imply that distribution networks might not suffer from problems of unsustainability. On the other hand, as has been discussed earlier (§3.5.3), in an *intertemporal* world, the non-fungibility of distribution assets would tend to suggest that unsustainability is the exception rather than the rule. If so, this might suggest that distributors need protection against destructive competition. But more pertinently, the presence of unsustainability, and the lack of an

⁴³ Curien (1991) considered that a less restrictive conceptualisation of cross subsidies would include welfare considerations as well. He found that the core of the “producer and consumer surplus” game lies within the core of Faulhaber’s “cost game”. Hence, like the inclusion of substitution effects, incorporating welfare-related aspects into the definition of cross subsidisation reduces the field of acceptable prices. Jamison (1996) also concluded that Faulhaber’s approach was too restrictive. By introducing the concept of “multilateral rivalry” into the determination of subsidy-free prices, Jamison demonstrated that this too narrows the core of game, meaning that Faulhaber’s bounds are simply necessary, but not sufficient conditions. Multilateral rivalry is considered by Jamison to arise from new “hybrid” utility companies, such as integrated electricity and gas companies, and from non-utility companies providing co-generation or energy conservation services. This approach can be seen as correcting for Faulhaber’s assumption that the consumer demand curve be derived by ignoring alternative supply possibilities (fn. 42). Jamison simply widens the net of possible alternatives to include supply options traditionally considered to be outside the market in question.

equilibrium, would greatly complicate any practical attempt to assess subsidy-free prices relating to the provision of network connection capacity (an issue addressed later; §8.4).

4.4 Distribution Line Charges in New Zealand

4.4.1 Government Perspectives Regarding Natural Monopoly Pricing

In a discussion paper on New Zealand’s telecommunications sector, the Ministry of Commerce and The Treasury (1995) evaluate the pros and cons of many of the pricing approaches outlined in the earlier sections of this Chapter. The presentation of various pricing methodologies is examined in the context of “access pricing” (often termed “interconnection pricing”), which relates to the efficient price to charge a competitor for access to an incumbent monopolist’s (network) assets (§5.1.2). The paper states that, while “pricing at short-run marginal cost is economically efficient it usually does not provide enough revenue for the monopolist to cover total costs” (§4.1.4).⁴⁴ Consequently, it is recognised that: “In some cases the revenue-requirement can be met through the use of two-part pricing, charging a high fixed charge to users with a variable charge equal to marginal cost” (§4.2.4). The paper also discusses peak load pricing (§4.1.2), but it only refers to the conventional Steiner solution—with constant returns to scale—thus declaring that “price is set at long run marginal cost (LRMC) in the period with the peak demand and at short run marginal cost in other periods” (§4.3.4). The problems of stochastic demand are also noted.

The discussion paper also alludes to constrained market pricing (§4.3.1) by noting that, “if the monopolist produces multiple products the average cost of the firm cannot be defined. Instead, the regulated price for each product should be somewhere above long-run average incremental costs [i.e., LRAIC] and below stand-alone cost” (§4.1.5 and §4.3.2).⁴⁵ Ramsey pricing is also acknowledged to be a “second-best” approach, although it is recognised that such prices “are not necessarily ‘sustainable’” (§4.3.6) and that “the information necessary to estimate elasticities of demand may not be readily available” (§4.2.2). Ramsey pricing is also associated with the use of constrained market pricing in a manner suggestive of the regulatory imposition of Baumol and his colleagues’ “weak invisible hand” (§4.3.6): “According to Ramsey pricing, between these upper and lower bounds [i.e., SAC and LRAIC] the price should be set higher on those products which have less elastic demands, in order to minimise the distortionary consequences of pricing above marginal cost”. Hence, the paper notes that: “We might further constrain the second-best prices by requiring that they be ‘subsidy-free’”, and Faulhaber’s (1975)

⁴⁴ All quotes taken from Ministry of Commerce and The Treasury (1995, para. 211, and Appendix D).

⁴⁵ The discussion paper notes that “no clear definition of the appropriate period” for evaluating LRAIC exists, but they cite “5-10 years to take account of time for network capacity to adjust to meet demand, and a period long enough for all costs to be avoidable”.

classic work on cross-subsidisation is directly referenced, as is Baumol and Sidak's (1994a) text that describes constrained market pricing.

4.4.2 Ministry of Commerce Recommended Distribution Line Charge Methodology

Notwithstanding this recognition of the various principles involved in efficient and fair pricing, the Ministry of Commerce's recommended pricing methodology for the power distribution sector initially opted for the "accountant's" fully distributed cost allocation approach (§4.2.3). This is interesting given that the discussion paper on the telecommunications industry (Ministry of Commerce and The Treasury, 1995) specifically addressed the fully distributed costs method and highlighted the "problems which arise from the arbitrariness with which the common costs are allocated amongst the firm's products".

In 1994, the Ministry of Commerce (Energy Policy Group, 1994a—the 'Disclosure Guidelines') distributed its first set of official guidelines on how electricity line businesses (ELBs) could comply with the information disclosure regime (§2.4.2). Section 4 of these Disclosure Guidelines was devoted to the Ministry's approach to deriving distribution business line charges from costs.⁴⁶ Although the Guidelines made it clear that distributors were "free to set line charges by any methodology they wish" (s1.12), where ELBs chose "to adopt pricing methodologies that do not directly follow the methods set out in" Section 4, "then a clear description of the alternative pricing methodology adopted" was "required to be disclosed" (s4.2).

The Ministry's recommended pricing methodology involved five major stages (s4.4): (i) revenue requirement and cost identification; (ii) disaggregation of charges; (iii) determination of statistics and properties of load groups; (iv) allocation of network costs; and (v) derivation of prices (i.e., line charges). The initial stage is similar to the traditional cost-plus approach taken by US power utilities (§6.4.1): determine revenue targets first, based on projected annual operations and maintenance (O&M), administration, overheads costs, plus depreciation provision (§7.1.1) and return on asset value (§6.3.1). Although the ELB's asset value could be based on any number of different valuation methodologies, generally the approach taken has been to use the methodology required by the disclosure regulations for benchmarking ELB financial performance, the Optimised Deprival Valuation, or ODV (§2.4.3 and §7.4). Another key part of this stage of the line charge methodology was to allocate O&M, depreciation and return on assets across the various components making up the distribution network. These costs are

⁴⁶ The line charge methodology in this Section of the Disclosure Guidelines was heavily based on the approach developed by an earlier industry working party (i.e., SOLEC, 1992).

assigned proportionately to network components according to their depreciated (or undepreciated) asset value.⁴⁷

Once costs—or more correctly “revenue requirements”—were assigned to network components, the next three stages of the process required grouping together consumers on the basis of similar (statistical and other) properties, associating the network components with these “consumer load groups”, and assigning the network costs to those groups.⁴⁸ Finally, these costs were converted into actual line charges on the basis of maximum demand (in kW), “actual usage” capacity (in kVA)—for the largest consumer groups—and a fixed charge to cover administration and overhead costs, resulting in a traditional multiple component tariff (§4.2.3). Note that while there is a kWh component in the line charge, the line charge was neither designed to cover direct energy costs (which was and still is part of the energy retailer’s business) nor to cover the indirect cost of losses. As discussed earlier (§3.1.3), losses are recovered by the energy retailer based on average loss factors calculated by the line business and assigned to various geographical areas of their network. Both before and after the separation of line and energy businesses, energy retailers have been, and still are, free to bundle the distribution line charges associated with a particular consumer, with the transmission charges as well as energy purchases and cost of losses, in any manner.⁴⁹

4.4.3 Approaches to Line Charge Setting by New Zealand’s ELBs

In the subsequent editions of the Disclosure Guidelines—including the most recent successor document, the Electricity Information Disclosure Handbook (Energy Markets Policy Group, 2000b)—no methodology for line charges was presented. The most recent disclosure regulations however, themselves appear to imply that ELBs are likely to use a similar fully distributed cost approach.⁵⁰

⁴⁷ The Guidelines (s4.11) required O&M costs and depreciation to be allocated to network components on the basis of asset replacement cost valuation, and return on assets to be allocated on a depreciated replacement cost basis (§7.4). The relevant components were: 400V lines general; 400V lines dedicated; distribution transformers; 11kV lines general; 11kV dedicated; zone substations; subtransmission lines; and dedicated networks. Hence, the Guidelines were classifying network components in a similar manner to Boiteux (§3.1.1), distinguishing between dedicated (i.e., “individual” connection) assets at any voltage level, and other assets of a more semi-individual and collective nature.

⁴⁸ The Guidelines (s4.15-s4.18) classify consumers into the following groups: 400V General ($\leq 15\text{kVA}$); 400V General ($>15\text{kVA}$); 400V Dedicated ($>15\text{kVA}$); 11kV General; 11kV Dedicated; and Dedicated Network ($\geq 11\text{kV}$). The Guidelines also describe customers who are “pseudo-dedicated” in that their consumption is “responsible” (§3.1.1) for the majority of the assets of a shared line.

⁴⁹ Recently, energy retailers have been required to offer residential consumers a “low” fixed charge option, whereby no more than 10% of the average residential consumer bill can be associated with fixed (i.e., non-avoidable) charges (Hodgson, 2000a).

⁵⁰ As of August 2001 (note the comments in §1.3.1), the most recent version was the *Electricity (Information Disclosure) Regulations 1999 Consolidated with the Electricity (Information Disclosure) Amendment Regulations 2000*. It is this version which is referenced in this subsection.

Among other information, Regulation 24 requires ELBs to disclose: (i) “the key components of the revenue required to cover costs and profits of the line owner’s line business activities”; (ii) “the consumer groups used to calculate the prices charged”, including “the rationale for the consumer grouping”, “the method by which the line owner determines which group consumers are in”, and “the statistics relating to that group which were used in the methodology”; (iv) “the method by which the line owner allocated the components of the revenue required to cover the costs of its line business activities amongst consumer groups”; and (v) “describe the method by which the line owner determined the proportion of its charges which are fixed and the proportion which are variable”.

In any event, a number of ELBs cite either the 1994 Disclosure Guidelines, or the earlier SOLEC (1992) document upon which the Guidelines were based (fn. 46), as providing the basis for their methodological approach. And whether or not they directly cite these documents, the majority of ELBs have applied a methodology that is recognisably based on the approach in the Guidelines, sometimes with the addition of refinements such as geographical zoning of consumer load groups, and the use of TOU pricing elements (§4.1.2), particularly for large consumers. A small number of ELBs refer to long run marginal cost or long run incremental cost components in the line charge. These components are typically applied to larger consumers at peak demand periods.⁵¹

4.4.4 Cross Subsidies in New Zealand’s Electricity Prices

Removing cross subsidies was one of the desired reform outcomes (§2.3.1), and it is often claimed that “cross subsidies have been removed” from bundled retail electricity prices (e.g., Lough, 1994; Wilson, 2000b). Although critical at the time of the reforms at wholesale level, by 1995 the Electricity Supply Association of New Zealand (ESANZ)—the association which represents distributors—painted a glowing picture of the Government’s reforms at distribution and retail level. ESANZ stated that success was already evidenced by a decrease in the retail price, and claimed that “cross-subsidies between domestic and commercial, and small to medium industrial customers, have been all but eliminated”. ESANZ stated that there had been a 0.49% decrease (in real terms) in the average retail electricity price, following the initial removal of retail franchises in April 1993. The use of the term “price” is somewhat misleading, as price was determined from, the total national residential and/or business revenue, divided by, the total national residential and/or business kWh consumption. Using this measure, the overall price reduction results from an increase in the real residential price of 6.82%, and a decrease in the real business price of 6.29%. These changes resulted in the residential to non-residential price ratio shifting from 0.891 to 1.015, and was the basis of the claim that cross subsidies had been “all but eliminated” (Leay, 1995).

⁵¹ These conclusions are drawn from a review of the disclosure information listed in the Bibliography.

However, it is apparent that ESANZ’s claim—one of the few to actually be backed up by any analytical work—was based on an examination of average revenues only. Yet a number of industry commentators had earlier recognised that—as the economic literature on cross subsidy highlights (§4.3.1)—examining prices independently of underlying costs allows no judgements to be made concerning cross-subsidisation (e.g., Kask, 1988b, pp. 15-18 and 38; Jackson, 1990, pp. 17-18). Consequently, some reform critics have suggested that all that has occurred is a reshuffling of the pre-existing non-residential to residential cross-subsidisation in the opposite direction. These critics consider that competitive pressures on electricity retailers will cause them to negotiate low-price and long-term contracts with their largest customers, and that these contracts can potentially be cross-subsidised by a captive pool of residential consumers (Noble, 1992). Similarly, if some parts of an ELB’s network are contestable (§3.2.3), while others are not, then this opens up similar possibilities.⁵² Such possibilities have also been raised in regard to power sector reforms outside New Zealand.

Moreover, supplier vulnerability to the exercise of monopsony power (including the threat of bypass) enables high-use customers to extract price reductions and to achieve network improvements that advantage these influential groups. Indeed, monopsony power appears to be behind many developments in public utility markets today. ... In electric markets, it seems to be a driving force behind steps to substitute price negotiation for price regulation, thus effectively achieving a system of selective discounting, and behind decisions to permit high-use consumers to pick and choose amongst sources of supply. This, in itself, gives to favored classes of consumers the ability to abandon—or threaten to abandon—the system and to leave fixed and sunk system costs the responsibility of low-volume consumers, who do not have this option (Miller, 1995).

Industry commentators point out that the ability of light-handed regulation, through information disclosure, to reveal the “fairness” of line tariffs is limited, since most consumers do not have the necessary skills to examine the data, and some would suggest that even with advanced analytical tools, benchmark cost comparisons between ELBs is extremely difficult (e.g., Gale and Strong, 1999). The Government’s requirement under the most recent regulations that ELB’s disclose their detailed asset management plans are considered by some to be “a desirable step forward. Properly prepared, the asset management plans will disclose also each line company’s power system planning criteria, its strategies and plans as well as its projected capital and operating expenditures, enabling a more detailed review of the future requirements to be carried out by the regulatory body or observers” (Wilson, 2000b). Wilson, however, goes on to caution that the scrutiny of these plans requires “considerable technical expertise and

⁵² Wilson (2000b) however suggests that “distribution lines companies are no longer subject to competitive forces in the same way as they were before the retail-lines split. They can and are therefore taking a more dispassionate view of network development strategy, untrammelled by competitive pressures to the same extent as before”. Wilson, however, implies that this lack of competitive pressure might be beneficial and has actually “led to improvements in the power planning practices of many of the distributors”.

experience, raising the question of who is expected to review such documentation or carry out such assessments”. As Wilson points out, “from the viewpoint of customers at large, the majority is neither interested in investigating their electricity charges in detail, nor would be competent to do so”. In any event, there appears to often be some misunderstandings regarding how cross subsidies should be evaluated, and the remaining Chapters address the issues of determining efficient and fair subsidy-free prices—both from an absolute and relative perspective, with the time dimension explicitly taken into account—in greater depth.

CHAPTER V

SUBSIDY-FREE PRICES IN POWER DISTRIBUTION NETWORKS: BACKWARD LOOKING OR FORWARD LOOKING?

Those who advocate marginal cost pricing treat historic costs as bygones that are forever bygones. Furthermore, they insist on keeping questions of allocation and pricing analytically separate from questions of finance: UK economic methodologist, Mark Blaug (1990)

While users should be confronted with the costs incurred at the margin, the costs at the margin form only a part of the overall costs of supply and, therefore, of the total opportunity costs to society of supplying the service. Total long run incremental costs consist of the actual running costs of the existing system, plus the costs of current and future additions within a reasonable planning time horizon, plus the ultimate replacement costs of the presently existing system: World Bank energy economist, Gunter Schramm (1991)

The Romans built temples to Janus, the most ancient king who reigned in Italy, who was often represented with two faces because he was believed to know both the past and the future. Like Janus, regulators alternate between past and future perspectives on markets as doing so serves their purpose. The result, which we shall call the Janus artifice, is an inconsistent economic analysis of competition and pricing. When evaluating the prospects for competition, regulators often look to the past, emphasizing the sunk costs of the incumbent. For pricing purposes, however, regulators look to the future, promoting their notion of forward-looking costs: US writers on law and economics, Gregory Sidak and Daniel Spulber (1997, p. 425)¹

The contestable market theorists do away with the traditional allocatively efficient pricing benchmarks of short run marginal cost (SRMC) and long run marginal cost (LRMC)—derived from within the perfectly competitive market paradigm (§4.1-§4.2)—and replace them with the perfectly contestable concepts of stand alone cost (SAC) price ceilings and incremental cost (IC) price floors (§4.3). The practice of implementing such price ceilings and floors in actual regulatory regimes is known as ‘constrained market pricing’. Although constrained market pricing is not applied in New Zealand’s regulatory regime for the power sector, the method bears closer examination. This is because both New Zealand’s power sector reforms and the implementation of constrained market pricing (a) are theoretically underpinned by contestability theory, and (b) acknowledge objectives of “fairness” in addition to efficiency (§2.1.6, §2.3.3 and §4.3.1).² Consequently, this Chapter examines the issues involved in implementing constrained market pricing in practice. In addition, a number of fallacies

¹ All quotes have been abridged for clarity, and acronyms spelt out in full.

² Moreover, as discussed earlier (§4.4.1), the New Zealand Ministry of Commerce and The Treasury (1995, para. 211) express familiarity with the concept of constrained market pricing, which they discuss in the context of possible access charges for competitors to an incumbent’s telecommunications network.

associated with calculating stand alone costs—and particularly incremental costs—are highlighted. It is submitted that one of the key causes of such fallacies is the way in which the time dimension is incorporated into the calculation of SAC and IC values, particularly with respect to whether costs are backward-looking (i.e., historic values) or forward-looking (i.e., replacement values).

5.1 Constrained Market Pricing in Practice

5.1.1 Subsidy-Free Pricing as a Regulatory Tool: Willingness-to-Pay and Demand Side Issues

Notwithstanding some of the theoretical complications involved in determining subsidy-free prices (§4.3.6), Baumol and his colleagues point to a number of real world applications of constrained market pricing, a term coined by the Interstate Commerce Commission in the US. Constrained market pricing dictates that prices should lie between stand alone costs and incremental costs, or at the very least be below SAC (e.g., BPW, pp. 504-510; Faulhaber and Baumol, 1988; Baumol and Sidak, 1994a, p. 81). Moreover, Baumol and Sidak (1994a, p. 43) provide the reminder that bounds on subsidy-free prices are determined neither to explain nor to predict actual market behavior, but rather to provide guidance to regulators where the real world falls short of the perfectly contestable ideal (§2.1.7).

The first-best lesson of the perfect competition model, calling for prices to be set equal to marginal costs, has no doubt contributed to the common regulatory ethos which *equates* price to *some* measure of cost. This doctrine has been used frequently where it is completely inappropriate and without logical foundation, that is, in cases where prices should be based on demand as well as cost considerations, because of the presence of economies of scale and scope. ... In contrast, contestability theory suggests cost measures that are appropriate guideposts for regulated pricing—incremental and stand-alone costs. ... One cannot legitimately infer that monopoly power is exercised from data showing that prices do not exceed stand-alone costs, and stand-alone costs constitute the proper cost-based ceilings upon prices, preventing both cross-subsidization and the exercise of monopoly power (BPW, p. 508).

The theoretical issues, though, do have some important implications for the practical application of constrained market pricing. In particular, the welfare basis for the SAC bound, and the impacts of demand side considerations, are related. Irrespective of more complex issues such as the effect of cross-elasticities of demand, it is clear that consumer demand considerations can strongly impact any assessment of price ceilings. As noted earlier (§4.3.6), Sidak and Spulber (1997, p. 341) affirm the legitimacy of SAC as the upper bound to acceptable prices. Intuitively, irrespective of the perfectly contestable benchmark, one might argue that it would be irrational for the willingness-to-pay of any group of consumers to exceed the stand alone costs of producing the goods which they consume. On the other hand, willingness-to-pay could be higher than SAC, because the utility of consumers might be increased through not having to engage in the nuisance of self-supply, such as any transaction costs

involved in forming a coalition.³ Moreover, consumers might not be aware that they could self-produce at a cheaper cost, either on their own, or as part of a coalition. As Heald (1997) points out, the recipients and sources of cross subsidy may not necessarily correctly perceive their own position. He suggests that recipients of cross subsidy “have a habit of believing that they are sources!”. Hence, the willingness-to-pay of consumers might be dependent on their access to accurate information regarding the costs of supply, as much as to their interest in the nature of prices. An advocate for domestic electricity consumers in New Zealand brings the issue firmly back to earth as follows.

To begin with, consumers are being asked to grapple with relatively complicated principles. Arguments about cross-subsidisation and separation of fixed and variable costs can entertain the theoreticians, ideologues, academics; and even some politicians can contribute to the debate. ... [D]omestic consumers essentially want to know how much they have to pay a month for their electricity; and for most, the machinations of economic theory don't matter (Russell, 1991b).⁴

Yet, if perfect contestability is to be the benchmark, then it could be argued that the subsidy-free bounds on prices should be derived assuming that consumers have access to perfect information. For instance, Spulber (1989, p. 4) considers that for the Bertrand-Nash assumption to hold—which is a precondition of a perfectly contestable market (§2.1.7)—a “perfect information” assumption must also be made. Therefore, Spulber states that a perfectly contestable market requires that: “perfect information on prices is available to all consumers and firms, and firms have complete demand information”.

If it is similarly assumed that consumers are also subject to the “rationality postulate” of perfect competition (e.g., Blaug, 1992, Ch. 15), then all that stands in the way of SAC bounds acting as a cap on willingness-to-pay are the transaction costs of forming a consumer coalition—in other words, any “transaction specific” aspects would need to be assumed away (§3.5.1). A key assumption of perfect contestability is that incumbents and entrants are “symmetric”, and this is consistent with the Baumol group's generally firm-oriented perspective of market behaviour (§4.3.6). In considering issues of anonymous equity, however, it does not seem to be stepping too far out of the perfect contestability framework to propose that firms and consumers are themselves quasi-symmetric—at least in the sense of having access to the whole range of options for serving demand, even where the incumbent firm is not actually using the least cost option. This would mean access to information regarding *all* possible supply alternatives, even those outside the normal definition of the consumer's “market”. The least cost

³ The converse could also be true; willingness-to-pay could be less than SAC, in which case there would be no demand at a price sufficient to cover stand alone costs.

⁴ This viewpoint, based on casual empiricism, is borne out by (albeit dated) international research into the response of consumers to energy prices: “although people are typically assumed to respond to marginal prices, they are more likely to notice average prices, and the limited evidence suggests that what people perceive most clearly is neither of these, but rather the total cost (for example, the monthly electric bill rather than the marginal price per kilowatt hour)” (Stern, 1986).

alternative would then be considered to constrain the SAC. Considering alternative supply options in determining stand alone costs is an approach that is implied by Jamison's (1996) notion of "multilateral rivalry" (fn. 12), and in the New Zealand Government's restriction on the valuation of electricity line businesses (§7.4.3).

5.1.2 Regulation of Absolute, Relative and Access Prices for New Zealand's Electricity Distributors

There are three key applications of constrained market pricing to network industries: (i) *absolute price levels*; (ii) *relative price levels*; and (iii) 'access pricing'. Regulators have traditionally focused on the first of these. Rate-of-return regulation and price-cap regulation of utilities (§6.4.1), as well as other approaches, have all had the objective of ensuring that monopolies are restricted to acceptable *absolute price levels*, which from a cross subsidy perspective would mean a zero economic profit. (This issue, as it relates to New Zealand's electricity line businesses is discussed further in Chapters VII and IX). As noted earlier (§4.3.2), in Faulhaber's view, any deviation from normal profit levels signals the cross subsidy of producers by consumers, or vice versa, although this conclusion is based on the assumption that technology and the optimal asset configuration do not change with time. The SAC of the *entire* network therefore, would appear to be the appropriate basis for assessing equitable total levels of revenue. However, this thesis finds (subject to certain assumptions) that the appropriate regulatory ceiling on prices is not the "vanilla" SAC, but the "*net intertemporal*" SAC, which is assessed on the hypothetical greenfields basis of all current and *future* consumers forming a coalition to serve their demands at least cost, taking into account changes in optimal network design.⁵

Assessing equitable *relative price levels* involves the considerably more complex issue of determining bounds on subsidy-free prices for each consumer or group of consumers in the network, given the joint (and some common) costs associated with network assets (§3.4.2). As Irwin (2000, p. 15) for one observes, even if the absolute price level is "efficient", significant allocative inefficiencies can arise due to relative price distortions, (and Irwin implies that Ramsey efficient price discrimination be applied to resolve this problem; §4.2.2). In a power distribution network, if each consumer is considered to demand a different product, namely the connection capacity at their particular location, then the issue of anonymous equity is automatically addressed through an examination of the joint cost of providing capacity on an individual, semi-individual and collective basis (§3.1.1). However, the number of possible ways that the network could be configured to serve the multiple combinations of different demands is immense. As Heald (1997) highlights, it is quite possible that a particular output which passes the SAC test when it is performed on that output alone, will fail the test when it is performed for some combination of outputs. In practice, because of this combinatorial complexity, regulatory tests of

⁵ The notion that price ceilings should be determined considering the demands of *future* consumers as well, is not often acknowledged in the literature. This is discussed shortly (§6.4.2), and forms a large part of the analysis in Chapters VIII and IX.

cross subsidies between consumers have focused on very large groups of consumers, such as residential versus business, or rural versus urban.

By contrast, *access pricing* has become a key area of regulatory focus in recent years, particularly for telecommunications networks (e.g., Cave and Doyle, 1994; Laffont and Tirole, 1994; Armstrong *et al.*, 1996; Valletti and Estache, 1999). New Zealand has itself offered up one of the classic textbook case studies of access pricing in telecommunications networks, due to the disputes between Telecom New Zealand and Clear Communications (e.g., Baumol and Sidak, 1994a, pp. 13, and 96-97; Sidak and Spulber, 1997, pp. 347-388; Harvard Business School, 1998). Nevertheless, during the period between initial franchise removal and complete separation of line and energy businesses, access pricing was also an important issue for electricity distribution in New Zealand. Similarly to telecommunications, the main concern was inequities between incumbent and external firms (i.e., electricity retailers), rather than between different consumers. As discussed earlier (§2.4.5), the concern was that incumbent line businesses could cross subsidise their associated retailer, and thus make viable entry difficult for competing retailers. The access pricing literature is therefore most relevant to the fifth and sixth type of cross subsidy in Heald's taxonomy (§4.3.1), and relates to—as the title of a seminal paper by Baumol and Sidak, (1994b) indicates—“the pricing of inputs sold to competitors”. Such a cross subsidy problem did not relate to the calculation of the line charge faced by the external or incumbent retailer, since this was the same in either case and was simply passed on to the consumer by the winning retailer. Instead, it impacted the bundled electricity price offered by the incumbent and external retailers to consumers, since the cross subsidy was from the regulated line business to the competitive energy business. The network cost-related portion of the bundled price (i.e., the line charge) was required to be the same in either case.

Now that line and energy businesses have been legally separated (§2.4.5), such concerns about cross subsidy have been set aside, and the focus of both the Government and industry commentators has gone back to the more traditional concern about the *absolute* level of electricity line business (ELB) revenues (§2.4.6). But, although the access pricing problem for electricity distribution mainly focused on fairness issues for competitors, it implicitly involves a *relative* price level issue with respect to consumers as well; the line charge faced by the incumbent or external retailers may itself embody a cross subsidy between different consumers on the network—Heald's third category of cross subsidy.⁶ This problem

⁶ With respect to the telecommunication sector outside the United States, Curien (1991) notes that, similarly, more priority has been given to ruling out cross subsidies from monopoly to competitive services, than to the reduction of cross subsidies *internal* to monopolistic activities. Parsons (1998) comes to the same conclusion with regard to the US. Much of the debate concerning access pricing relates to the relative merits of the efficient component pricing rule (ECPR), attributed to Willig (1979b) and Baumol (1979), and elaborated on by Baumol and Sidak (1994a-b). The ECPR was primarily developed to address this issue of cross subsidies between monopoly and competitive services, and was designed to be implemented in tandem with price cap regulation for the final product market (§6.4.1). As such, this thesis does not discuss the ECPR in any

has not been resolved by the separation of line and energy businesses, which has effectively made all retailers “external”. Moreover, for subnetwork competition (§3.3.1-§3.3.2), the incumbent and an external ELB face the same line charge at the point of the subnetwork’s connection to the rest of the distribution network.⁷ In this case, the equitable relative price level can also be considered the equitable access price for a competing downstream line business, hence line charges also relate to Heald’s sixth category of cross-subsidy after all.

5.1.3 Applicability of SAC and IC Bounds

Apart from the theoretical complications, and the problem of combinatorially numerous bounds on subsidy-free prices (§4.3.4), there are two other practical complications that arise when attempting to apply SAC and IC bounds as regulatory guidelines for equitable (and efficient) prices. Berg and Tschirhart (1995), Miller (1995), as well as Rees and Vickers (1995), are among a number of authors who bring attention to the first of these problems—the potentially large gap between IC and SAC values. Some regulators, notably the Federal Communication Commission in the United States, have rejected the application of SAC tests simply because, in practice, the tests can result in very high ceiling prices.⁸ The implication is that, if the flexibility provided to the industry by price ceilings and price floors is large, regulation will have little effect, and the cost of regulation is likely to outweigh its benefits. In discussing access charges for telecommunication distribution networks, Heald (1997) also draws attention to the possibility that a wide gap may exist between the subsidy-free bounds.

When the regulator insists on access, there arises the question of how much the entrant must pay for access to the incumbent’s distribution network. Given the evident clash of interests, the regulator may have either to enunciate the principles upon which such access is available, or to approve (perhaps even to determine) a tariff structure for access. The complexity of the issue is obvious. The IC might be very low (provided that there is some *spare capacity* in the incumbent’s network), whereas SAC might be extremely high (the very reason why the question of access arises); (Heald, 1997 [emphasis added]).

detail. It is interesting to note, however, that while Baumol and Sidak indicate that efficient prices based on incremental costs provide no contribution toward common or joint costs (§5.3.1 and §5.3.3), the efficient access price based on ECPR does include a positive contribution toward common costs. The ECPR price comprises both the incremental costs of a service, plus the incremental opportunity costs of that service, although the incremental opportunity costs are only positive where capacity is fixed or constrained (Baumol and Sidak, 1994a, Ch. 7).

⁷ Jagger (1996) and Spong (1998) both discussed the possible relevance of the ECPR rule for setting line business network access prices for energy retailers in a New Zealand context, during the time prior to the legal separation of the two types of businesses.

⁸ Parsons (1998) points out that, while the legislation governing the US Federal Communications Commission—the federal *Telecommunications Act of 1996*—contains the word “cost” over thirty times, and the terms “subsidy” or “cross-subsidy” eight times, “there is no mention of the terms ‘incremental’ or ‘stand-alone’”.

This may make application of the approach “politically” unpalatable since it provides firms with a relatively free hand to price discriminate within wide limits. Heald (1997) notes that under constrained market pricing, joint or common costs may be loaded onto residential consumers who have few effective alternatives. But provided that other consumers—such as industrial or commercial consumers, who may have better alternatives—are paying at least incremental cost, residential consumers, however antagonised they may feel by having to pay for all or most of the joint costs of shared inputs, only have an economic case if they are paying more than SAC, in which case they can exit from the market entirely. Baumol and his colleagues would view this as an example of the “weak invisible hand” at work (§4.3.6), since the residential consumers are likely to have more inelastic demands than other consumers.

However, with respect to electricity distribution, the possibility of a flat demand curve for network capacity (§3.1.2) has important implications for allocative efficiency. A key argument for the inefficiency of monopoly pricing is that it leaves customers who would be willing to purchase their output at cost reflexive prices with unmet demand (e.g., Williamson and Mumssen, 2000, p. 3). But if demand curves are flat for many consumers, then as long as prices are below consumer willingness-to-pay this will not result in any distortion to the demand for electrical capacity (although it may distort the consumption of other goods and services, including electrical *energy*). Moreover, a flat demand curve would mean that all such consumers are equally price inelastic.⁹ In a developed country such as New Zealand it seems unlikely that the majority of residential and commercial consumers will respond to increased power bills by disconnecting from the network, or even by reducing their demand for capacity. Intuitively, consumers might be more likely to respond by reducing their electrical energy consumption or by reducing their consumption of (luxury) goods and services.¹⁰

⁹ Similarly, the demand for telephone *connection* is generally perceived as being highly price inelastic (e.g., Baumol and Sidak, 1994a, p. 38), although demand for individual calls may be much less so. However, Rees and Vickers (1995) suggest that the willingness-to-pay for bundled electricity provision is much higher than for telephone service. Rees and Vickers imply that the demand for network connection might be flat, because they indicate that, where the variable component of a two-part tariff covers the marginal cost of electrical energy, the fixed component of the bundled electricity price (i.e., the component related to network capacity) can be raised to cover remaining costs without causing any consumers to disconnect: “Under some circumstances, each consumer’s benefit from consumption is so large at marginal cost prices that a lump-sum charge (for connection, line rental or standing charge) can be set that covers the revenue shortfall from marginal cost pricing without causing any consumer to stop using the product or service. While this may be approximately true for water and electricity, it is probably not so for gas and telephones” (Rees and Vickers, 1995).

¹⁰ The intuitive case for price inelasticity is not as strong for (large) industrial consumers that may either “vote with their feet” and choose to take supply at a different geographical location (i.e., from a different, less discriminatory ELB), or reduce their demand for capacity by switching part of their production process to an alternative energy source. Although the logic of a flat demand curve for residential and commercial consumers of network connection capacity seems reasonable, there is no research available to support such a supposition. Apart from standard market surveys, there has been minimal research in New Zealand into questions of price elasticity, even for energy (e.g., Hunter and Matheson, 1994) let alone network

The second key problem is that stand alone costs are difficult to measure (or estimate), and this is complicated—for the purposes of calculating incremental costs—by the fact that even regulated firms will rarely make exactly a normal economic profit.¹¹ In his discussion of cross subsidies, Heald (1997) ignores the complementarity between SAC and IC (§4.3.2). Possibly this is because such complementarity is dependent on the firm in question earning a zero profit. Where this does not hold, then total revenues (which are readily measurable) will not provide an accurate proxy for total costs (which are more difficult to determine—at least for a regulator).

Such a view is held by Curien (1991), who assesses “subsidy-free” prices for firms which do *not* make a normal profit, and demonstrates how the excess profit or loss can be redistributed (i.e., allocated) to consumers. Where normal profits are not made, Curien points out that the game will not be “zero sum”, and the cross subsidies will not add up to zero. Yet regulating natural monopolies through constrained market pricing is not just about achieving equitable *relative* price levels, but *absolute* price levels as well. However, Curien is primarily concerned with the problem of cost measurability, and points out that, because of this problem—at least for the telecommunications industry—regulators have moved to other methods which are not strictly related to the theoretical notion of cross subsidy. Curien investigates the use of “revenue trade-offs” as a proxy to cross subsidies, since these require no direct estimation of the cost function, nor an explicit assessment of internal economies of scope or scale, but the approach is more akin to traditional cost allocation methods (§4.2.3) than to Faulhaber’s cross subsidisation methodology. Curien concedes that the choice of the cost allocation rule is more or less arbitrary.

connection as a distinct product. Furthermore, since consumers typically face a bundled price, the cost of network connection is not necessarily passed directly through to the consumer, which would complicate attempts to measure price elasticities relating to the demand for capacity. This is particularly the case if the consumer chooses a retail tariff option with a low fixed cost component (i.e., 10% of the average bill), which the Government now intends ELBs to offer residential consumers under its most recent Policy Statement (§2.3.3). On the other hand, there is some circularity in the argument which relates price inelasticity to “captive” consumer groups. If effective competition can be introduced, then clearly any price inelasticity will only be valid up to the price level of the alternative supplier. Competition can itself improve the price responsiveness of demand (e.g., Lube, 1995; Lowry and Kaufmann, 1998, p. 4), and this view is also implied by Hay and Morris (1993, p. 605). Miller and Tye (1985) point out that endogenous R&D investments will also alter both long and short run elasticities.

¹¹ Faulhaber and Levinson (1981) themselves pointed out this problem, indicating that they did not “assert that that achieving least cost production with undominated zero-economic-profit prices is easy in regulated enterprises. However, it is not the problem to which this paper is addressed”.

5.2 Incremental and Stand Alone Costs in a Static World

5.2.1 Common Costs and Incremental Costs

One of the issues noted in the previous subsection which apparently makes constrained market pricing difficult to implement in practice, is the supposedly large gap between SAC and IC. The former is seen as being potentially very high, whereas the latter might be low, or even zero. This thesis submits that this concern may arise because stand alone costs and, in particular, incremental costs, can be evaluated incorrectly, without a full assessment of all the positive (and negative) factors impinging on those costs. This issue is addressed in more depth throughout the next Section, as well as in Chapters VI, VIII and IX.

Even though Baumol and Sidak (1994a, p. 51) acknowledge that the costs to society of miscalculating the price floor (IC) and price ceiling (SAC) are high, they suggest that this risk of regulatory error besets every form of price regulation. Nevertheless, it does appear that a number of authors (and as will be shown, even Baumol and Sidak themselves) misinterpret how bounds on subsidy-free prices—ones consistent with the benchmark of perfect contestability—should be calculated. For example, while Heald (1997) defines stand alone cost as the hypothetical cost of producing a set of products in isolation from the other products—an approach consistent with that of Faulhaber and the contestability theorists—he defines incremental cost as “the increase in cost associated with producing a ‘second’ output in addition to a ‘first’ output”. Based on this definition, Heald concludes that “it may matter crucially which output is defined as ‘first’ and which as ‘second’ because *the first must carry all the common costs*”. Similarly, Jamison (1996) states that, “in the traditional analyses, subsidy-free prices range from incremental costs to stand-alone costs, and *contain no assignment of common costs*”.¹²

But, as discussed earlier (§4.3.2), the incremental cost associated with a group of products (or consumers) should be derived by subtracting the stand alone cost of the complementary set of consumers from the total cost of supplying all products. Heald’s statement appears to suggest that incremental costs can be determined in their own right, and (as already mentioned) he does not draw attention to the complementarity of the SAC tests and the IC tests. Moreover, the supposition that incremental costs never include any contribution to common or joint costs is incorrect, if common costs are considered to be the physically non-separable costs of indivisible capacity. A simple justification of this can be seen from the reworking of the conventional peak load pricing problem presented earlier (§4.3.4), but in the presence of *declining average costs*. Ignoring operational costs, the incremental cost of the capacity that is physically “common” to both peak and off-peak products includes a positive contribution to the

¹² Emphasis added in these quotes. “Common costs” are used by Heald and Jamison in the sense of both “common” and “joint” costs (§3.4.2). Jamison (1996) only considered that incremental costs would include some contribution to common costs in the presence of “multilateral rivalry” (§4.3.6).

incremental cost bound for the peak product. Conversely, were the off-peak product to be defined as the “first” product, it would not pay *all* these physically common costs, since the stand alone cost bound is less than the total cost of the asset common to both. Heald’s and Jamison’s suggestion that incremental costs contain no contribution to the physically common costs of indivisible assets is only correct where the costs of common capacity are subject to constant returns to scale (§5.3.3).¹³ They appear to have fallen into a similar trap to those who maintain that, in general, off-peak consumers never contribute to capacity costs (§4.1.2).

5.2.2 *Bounds on Subsidy-Free Revenues in a Power Distribution Network*

Another example of the fallacy that incremental costs do not contribute to physically common or joint costs is useful at this stage. Consider a firm providing two goods or services, the demand for which is q_1 and q_2 , respectively. Capacity is a shared input for the production of both products, and there are no variable costs associated with production. Declining average incremental costs (§4.3.4, fn. 33) are assumed for the costs of capacity $K(q_1, q_2)$. Although the products are distinct, their quantities are measured in the same units, hence declining average incremental costs can be determined without running into Baumol’s “apples and oranges” problem (§3.4.4). While the costs of capacity can be considered “fixed” in the static sense, they are not “constant”, since the costs of capacity depend on the demand for both products. But because of the declining average incremental costs, the costs of capacity are not physically “separable” between the two products and therefore the capacity is “joint” to both. Using the same approach as earlier (§4.3.2-§4.3.4), it can be shown that the set of expressions defining the set of subsidy-free prices are as shown in (5.1)-(5.3), subject to (5.4), the declining average incremental costs assumption.

$$K(q_1, q_2) - K(q_2) \leq P_1 q_1 \leq K(q_1) \tag{5.1}$$

$$K(q_1, q_2) - K(q_1) \leq P_2 q_2 \leq K(q_2) \tag{5.2}$$

¹³ Jamison (1996) also misinterprets Faulhaber’s definition of subsidy-free prices. He states that Faulhaber’s proposition is that “prices are subsidy-free so long as customers are no worse off with these prices than they would be with the best alternative which can be offered”. This implies that the core of the game is itself Pareto optimal (in a partial equilibrium context), which is not correct. For example, in the conventional peak load pricing example subject to declining average costs (§4.3.4), one possible subsidy free set of prices would be for the monopoly supplier to offer the peak product consumers a price based on the stand alone cost of supplying peak demand. To satisfy the total revenue cost equality, this would mean charging off-peak consumers the incremental cost of their demand. But another monopolist could offer the reverse set of prices: peak consumers pay their IC, while off-peak consumers pay their SAC. Although both sets of prices are subsidy-free, the “best alternative” from the point of view of off-peak consumers is the first scenario, while the opposite is the case for peak consumers. Subsidy-free prices are therefore not in themselves an equilibrium position, and a Pareto optimal price set will need to lie somewhere within the core (e.g., the Aumann-Shapley value; §4.3.5).

$$P_1q_1 + P_2q_2 = K(q_1, q_2) \tag{5.3}$$

$$\frac{K(q_1, q_2)}{q_1 + q_2} < \frac{K(q_1)}{q_1} \quad \text{and} \quad \frac{K(q_1, q_2)}{q_1 + q_2} < \frac{K(q_2)}{q_2} \tag{5.4}$$

If each product has different price elasticities, within the price limits defined by (5.1) and (5.2), then a Ramsey efficient solution would require both products to be priced at a different level. This price would be non-zero for both products, and would satisfy all the expressions (5.1)-(5.3). A problem of course arises if the products cannot be “distinguished”—in other words, the firm is unable to determine which of the two products is being consumed, or whether arbitrage might be occurring between consumers of the two products. Price discrimination would therefore not be possible, and the firm would have no alternative but to charge the average price of the two products for both products (i.e., $P_1=P_2=K(q_1, q_2)/(q_1 + q_2)$). However, this average price, although not second best efficient—since it does not equal either of the efficient Ramsey prices—is still third best efficient, since it satisfies the subsidy-free price bounds.

But consider if the demands in question are for MV connection capacity at two different locations in a power distribution network. The demands are then both distinguishable and, for all practical purposes, consumers are unable to arbitrage their “capacity option” (§3.1.1). Price can be discriminated (in a Ramsey manner if desired) and demand can potentially be constrained by the firm. In addition to the costs of joint capacity, there will also be capacity costs associated with the dedicated assets required to link both locations with the joint capacity. However, these dedicated, rather than joint, costs of capacity are separable and can be directly attributed to the consumer(s) at each location. The joint costs of capacity are the costs the “semi-individual network” (§3.1.1)—namely, all distribution network assets upstream of both consumers, such as the local zone substation, and the subtransmission network linking that zone substation to the transmission grid. The dedicated costs of capacity relate to the “individual connections”—the lines or cables which link the consumers at the two locations to the zone substation.

The situation is simplified if only a single consumer demands capacity at each location. However, there may be a number of electricity consumers at a particular location, who although they could each be considered to consume a different product, are not able to be distinguished for some reason. Alternatively, because consumers are all at the same site, it might not be possible to prevent arbitrage or to ensure that each consumer’s demand for capacity is constrained.¹⁴ In this case, the costs

¹⁴ Although such practices may seem unrealistic to a reader from a developed country, arbitrage of power, and the possibility that demand can exceed measured installed capacity, is commonplace in many developing countries. For example, Davies (1987) describes a failed electricity tariff in South Africa. As a proxy for actual demand for capacity, residential consumers

of the individual network would be “common” rather than “joint” to all of the “products” at that location, and the firm would have no choice but to allocate the costs of both individual and upstream semi-individual capacity on an equal basis. Hence, distinguishability of products or consumers, and measurability of demand, makes a difference to the prices which can be charged in practice.

In any event, regardless which of the two location-distinguishable products is defined as “first” or “second”, the incremental cost of either product does make a positive contribution to the joint costs of capacity. Heald and Jamison’s statements (§5.1.4) suggest that if, say the demand q_1 was considered to be associated with the “first product”, then the price charged for that product (P_1) would recover *all* non-separable costs (i.e., $K(q_1, q_2)$). The simple example above demonstrates that this is not correct.

5.2.3 *Average Incremental Costs in Constrained Market Pricing*

Baumol and Sidak provide their own “generic” definition of incremental cost, presumably to differentiate this concept from the test they consider appropriate to use in determining regulatory price floors—namely, an ‘*average incremental cost*’ test. Rather than defining incremental cost with respect to some already-specified state of the market (both present and future), generic incremental cost is associated with the direct increase in cost incurred upon expansion of the market, when compared to current conditions. This definition implies a *sequencing* of events—current market state followed by expansion—in a manner rather similar to Heald’s notion of a “first” product (i.e., the current market state) followed by a “second” product (i.e., the expanded market state).

Incremental cost is a generic concept referring to the addition, per unit of the additional output in question, to the firm’s total cost when the output of X expands by some preselected increment. Thus, marginal cost can be approximated by incremental cost if the increment in question is small. But if the increment is large, marginal cost and incremental cost can differ substantially, because the ranges of the outputs examined in the two calculations are not the same (Baumol and Sidak, 1994a, p. 57).

In describing the appropriate concept that regulators should use to assess price floors, Baumol and Sidak present and define the term “average incremental cost” (AIC). Their definition of AIC resembles Faulhaber’s definition of incremental costs as the *complement* of stand alone costs under conditions of zero profit (§4.3.2), and in fact, Baumol and Sidak (1994a, pp. 59 and 82-85) do reaffirm this complementarity.

were charged a fixed amount per billing period on the basis of the number of power outlets installed throughout their residence. This provided incentives for consumers to rip out most power outlets and to run extension cords around their house from only one or two outlets.

Average-incremental cost, along with marginal cost, is the concept most frequently cited in recent discussions of public-interest floors on prices. The average-incremental cost of the entire service is defined as the difference in the firm's total costs with and without service X supplied, divided by the output of X . Formally, if we let x,y,z,\dots represent the outputs of the firm's various products, and $TC(x,y,z,\dots)$ is the total amount the firm must expend in producing that combination of outputs, then we have $AIC_X = [TC(x,y,z,\dots) - TC(0,y,z,\dots)]/x$ (Baumol and Sidak, 1994a, p. 57).

There is an important difference, however. Faulhaber's incremental cost for a *set* of products X , is equal to Baumol and Sidak's *average* incremental cost for the products in that set, multiplied by the total number of units of the products in that set (i.e., $IC_X = x.AIC_X$). Harkening back to BPW's "apples and oranges problem" (§3.4.4), Baumol and Sidak (1994a, p. 65) make it clear that the outputs over which the averaging occurs must be homogeneous. They provide an example where the costs of providing two products differ because the distance from the supplier to the distinct consumer of each product is not the same. Clearly it is not legitimate to average the two costs and claim that the price to the nearer consumer must exceed the average incremental cost.

But Faulhaber's method does *not* require or even imply averaging, hence it can be applied to products which are not homogeneous. Incremental cost, unlike AIC and marginal cost, is not measured on a *per unit* basis. (Rather oddly, Baumol and Sidak do not define an average SAC value on a per unit basis. SAC is directly comparable to gross revenues rather than revenue per unit). For Faulhaber, the incremental cost of providing the two products in Baumol and Sidak's example, must not exceed the total revenue earned from both products. Therefore, significantly, that single incremental cost value provides no information about the individual *price* charged to *each* product, it solely provides a lower bound on the *revenue* to be earned from *both* products (§4.3.2). Faulhaber's game theoretic approach provides a core solution which is defined by a set of constraints relating to every possible combination of products in the market. The bound on the price for a specific product is never defined by just a single equation, but is evaluated simultaneously with the solution of all the other equations describing the multidimensional core region. Specifying the price of one product immediately reduces the choice of the prices for every other product by one degree of freedom. Consequently, it may appear that Baumol and Sidak's definition of average incremental cost is potentially open to some misinterpretation (although as is explained in the next sub-section, this should not be the case).¹⁵ Even if the two products in their example were homogeneous after all, this still does not mean that the lowest price that can be charged for each product is one half of the incremental cost of providing those two products. Again, the prices of

¹⁵ For instance, see the definitions of SAC and AIC provided by the New Zealand Ministry of Commerce and The Treasury (1995) in fns 24 and 29.

those two products depend on all the other constraint equations, and not just that single incremental cost equation.

Baumol and Sidak (1994a, pp. 58 and 67) indicate that *both* average incremental costs and marginal costs are the pertinent figures for regulatory floor ceilings—the higher of the two values is the applicable one, and is termed the ‘effective price floor’. In the presence of declining average incremental costs, AIC will be the effective price floor. The relationship between AIC and marginal cost is outlined by Baumol and Sidak in their additional definition of AIC as the average of: the marginal cost of a homogeneous product consumed by the first customer; plus, the marginal cost of the same product when consumed by the second customer; and so forth (Baumol and Sidak, 1994a, p. 65). AIC is thus the average of the marginal costs of supplying a *single* homogeneous product to a number of consumers.

5.2.4 Fixed Costs in Constrained Market Pricing

Baumol and Sidak (1994a, p. 69) state that “by definition, marginal cost makes no contribution toward recovery of any fixed costs of the firm”. Yet, if marginal cost provides no contribution to “fixed costs”, then under this definition, it might appear that neither does AIC. However, Baumol and Sidak do make an exception. Where the fixed costs in question are directly attributable to a single product—for instance, they are associated with *dedicated* assets (§3.1.1)—then those “product-specific fixed costs” are included in that product’s AIC. On the other hand, Baumol and Sidak clearly point out that, where fixed costs are *not* directly attributable to a particular product, but are *common* to two (or more) products instead, then the AIC of each individual product does not include any contribution to those common costs.

[A]verage-incremental cost ... does include any fixed cost incurred exclusively for the service in question. But the average-incremental cost of service *X* does not include any contribution toward any fixed costs incurred *in common* for *X* and some other service or services supplied by the same firm. ... [The fixed cost outlay for an asset common to both products *X* and *Y*] must be made if the company supplies only *X*, only *Y*, or both *X* and *Y*. Consequently, the [common cost] is not incremental to either *X* or *Y* alone. If either service were discontinued, the company could not avoid the cost of replacing the [common asset] when the time for that arrived (Baumol and Sidak, 1994a, p. 69).¹⁶

On first examination, it may seem as though Baumol and Sidak are drawing the same conclusion as Heald and Jamison—namely, that incremental costs do not contribute to common costs. However, there is a subtle, but important, difference. In the above example, Baumol and Sidak are considering the

¹⁶ Emphasis in original. Baumol and Sidak’s (1994a) example of a common asset with fixed costs is based on an air-conditioning unit required to prevent contamination of the single space used to house equipment that produces both service *X* and service *Y*. The exact same unit is considered to be required regardless of whether either or both services are produced.

costs associated with common assets to be *fixed*, not in the sense that the costs of capacity are “fixed” in just the short run, but “fixed” in *both* the short run and the *long* run.¹⁷ Effectively, it appears that Baumol and Sidak are only considering common costs that are unaffected by any changes in demand, present or future. Capacity costs would only be fixed in this sense if they are independent of demand over the long run as well as the short run. Conversely, fixed costs are not necessarily associated with durable capacity. As Baumol and his colleagues have noted elsewhere (BPW, p. 408), fixed costs can arise from the costs incurred for the establishment of a firm. These may include “outlays on the legal process, the initial acquisition of managerial and technical skills, and the collection of required marketing and technical information”. Apart from fixed capacity, such as a bridge or a railroad track, Baumol and Sidak (1994a, p. 72) cite the cost of senior management in a large, diversified corporation, as being an example of a fixed cost.

Although fixed costs—if associated with only two products—are not incremental to either product, Baumol and Sidak (1994a, p. 70) explain that they are nevertheless incremental to *both* products *in combination*. Hence AIC_{XY} does include the total fixed costs common to both products, whereas neither AIC_X nor AIC_Y includes any contribution to the common fixed costs. Baumol and Sidak provide the reminder that a proper evaluation of subsidy-free bounds on prices—one performed in a manner consistent with Faulhaber’s approach—involves a *combinatorial* incremental cost test. *All* the incremental cost tests associated with products individually and in every possible combination must be satisfied. Hence, Baumol and Sidak’s definition of AIC, if applied correctly (i.e., in the same way as Faulhaber), should not be open to misinterpretation after all. The combinatorial nature of the problem might imply that determining subsidy-free prices might become irresolvable if there are a large number of products. However, Baumol and Sidak (1994a, p. 72) suggest that, in practice, “because it is often possible to verify that many combinations of a firm’s products entail no common fixed costs, the task is considerably simplified”.¹⁸

¹⁷ Baumol and Sidak’s presentation of constrained market pricing is only a precursor to their key purpose, the application of the efficient component pricing rule (ECPR; fn. 6). The ECPR is particularly designed to compensate incumbent monopolists for the opportunity costs associated with losing business that would otherwise contribute to their common fixed costs. Baumol and Sidak (1994a, p. 113) argue that, unless opportunity costs are included in the access price charged to rivals, if capacity is fixed, then losing an existing (or potential) consumer to a rival means a loss of revenue that should be contributing to the cost of capacity. Since the focus of their tract is on telecommunications—an industry characterised by fixed capacity—the examples which Baumol and Sidak provide focus on cases where capacity is fixed in both the short run and the long run. However, Baumol and Sidak (1994a, p. 114) do briefly discuss the case where capacity is easily expanded, and note that in such cases, while the marginal opportunity cost is zero, it does not follow that the incremental opportunity cost is also zero, since the opportunity cost of the *inframarginal* units of capacity are likely to be positive. (Nevertheless, there is some ambiguity regarding how Baumol and Sidak have defined fixed cost, as is discussed below; §5.3.3).

¹⁸ Baumol and Sidak (1994a, p. 79) go on to say that: “the procedure advocated here does not require calculation in advance of *every* pertinent incremental and stand-alone cost. Only those few cost figures at issue in a particular hearing need be

To demonstrate how fixed costs impact incremental costs, fixed costs can be introduced into the earlier example of distribution network supplying just two products (§5.2.2). The fixed costs (F)—for instance, the cost of the land upon which the zone substation is sited (§3.6.2)—are considered to be joint to both products. The constraints which simultaneously describe the set of subsidy-free revenues—since IC is calculated rather than AIC—are shown in (5.5)-(5.7). Once again, declining average incremental costs are assumed, hence (5.4) also holds.

$$K(q_1, q_2) - K(q_2) \leq P_1 q_1 \leq F + K(q_1) \quad (5.5)$$

$$K(q_1, q_2) - K(q_1) \leq P_2 q_2 \leq F + K(q_2) \quad (5.6)$$

$$P_1 q_1 + P_2 q_2 = F + K(q_1, q_2) \quad (5.7)$$

The total revenue/cost equality can be considered to be the intersection of both the SAC condition and the IC condition for both products in combination (§4.3.2). Therefore, it can be seen from this simple two good model that the IC of *both* products includes the common fixed costs (F), whereas the IC of *each* product does not. The SAC of each product does include the fixed costs, since those products could not be served on a stand alone basis without the firm incurring the entire fixed costs. However, the SAC of supplying the market only requires the outlay on fixed costs to be made a single time.

Heald and Jamison did not indicate whether they consider common and/or joint costs to necessarily be fixed. The example above appears to indicate that, had they defined common costs to be fixed, then their conclusion that (individual product) incremental costs provide no contribution to physically common costs would have been correct after all. The issue seems to be simply one of appropriately defining the characteristics of the common costs. Unfortunately, this simple example, although not incorrect, is misleading. Although the costs involved can be considered to be relevant over the long run, *time* is not a distinct dimension in the example above. It is the premise of this thesis that as soon as *sequencing* of demands is introduced into the example, the incremental costs of individual products may contribute to physically common costs after all, even where those costs can be considered to be fixed in both the short run and the long run.

determined, and those are legitimately determined *ex post*. One can examine *in retrospect* whether a particular price fell short of the corresponding incremental cost or exceeded the corresponding stand-alone cost. That is why it has proved feasible in practice to carry out the requisite calculations, at least to a reasonable level of approximation”.

5.3 The Time Dimension in Constrained Market Pricing

5.3.1 Sequencing of Demands in Constrained Market Pricing

Baumol and Sidak, however, would not necessarily agree with the statement concluding the previous subsection. As in Baumol (1986, p. 116), they explicitly consider the impact of sequencing on incremental costs as follows, and discount the possibility that incremental costs will contain any contribution to the fixed common costs.

Nor can one argue that the [common] cost is the responsibility of the service that happened to be provided first. That the company started to supply *X* in 1980, while *Y* was not introduced until 1987, is an irrelevant piece of history. Today, neither service can be provided without the [common asset], and once the firm has decided to continue either one of the services, provision of the other adds zero to total [common] costs. Thus the [common cost] is not part of the incremental cost of either *X* or *Y*; and since the cost is fixed, it is also not part of either service's marginal cost (Baumol and Sidak, 1994a, p. 69).

Baumol and Sidak here seem to be invoking the “generic” incremental cost concept involving a *sequencing* of events (§5.2.3), where demand for a “first” service emerges *before* the demand for a “second” service. Heald's concern regarding the labelling of products as “first” or “second” seems to be because it is arbitrary; either one of the products could be considered as being provided “*in addition*” to the other product (§5.2.1). However, where there is a clear sequencing of events it might not seem unreasonable to label the product for which demand commences first, the “first” product. The “second” product is not just supplied “in addition” to the “first” product, but *after* the “first” product.¹⁹ However, Heald's conclusion that the first product must carry all the common costs is not consistent with Baumol and Sidak's view. Baumol and Sidak clearly state that *either* service could carry all the common costs; the fact that the demands for each service begin at different times is, to them, “irrelevant”. Because either service could legitimately cover all common costs without causing a cross subsidy, the complementarity of SAC and IC ensures that the IC of each product individually does not contribute to common costs.

Even though, in theory, Baumol and Sidak's approach could result in revenues from either the “first” or the “second” product covering all the common costs, Miller (1995) suggests that, in practice, it is usually the “captive” consumers of “first” product, who end up paying for the common costs. Miller views constrained market pricing as unfairly allowing consumers of new and more-advanced services to be charged only the incremental cost of upgrade required for provision of that service, while the original consumers are allocated all the residual costs. Hence, she critiques Baumol and Sidak's (1994a)

¹⁹ Note that the demand for the first product is not considered to cease once demand for the second product begins. Demand for both products continues in parallel.

approach as being “a political rather than an economic exercise, designed more as a tactical maneuver to achieve strategic objectives than as a means to achieve so-called economic goals such as Pareto optimality”.

The average incremental cost of a service is identified as covering only the direct cost of the service, including *service specific* fixed costs, but not any contribution toward fixed costs incurred in common or jointly for the specific service or for other services supplied by the firm. ... All other costs are to be absorbed by the native load base. The stand-alone cost is further identified as the highest price it is possible to charge in contestable markets. ... Needless to say, if the precepts and assumptions of contestability theory are incorrect, that price turns out to have few economic limits and the process becomes a matter of *saufe qui peut*—that is to say, pricing limits are exclusively political. The proposal is advanced under the banner of economic efficiency (Miller, 1995).

5.3.2 *Fallacies of Forward-Looking and Backward-Looking Costs*

Taking a rather less extreme position than Miller, Heald attributes the wide gap between SAC and IC to low incremental costs due to *spare capacity* (§5.1.3). Had a shared input been constructed to be ready to serve demand for the first product, but sized at a sufficient capacity to meet the demand for both products—either due to equipment indivisibilities, or *in anticipation* of the future demand for the second product (§3.4.2 and §6.1.1)—then perhaps consumers of the first product should pay the total costs of capacity after all. Since the later demand for the second product causes no *additional* cost once it commences—in other words, the marginal cost of capacity is zero—perhaps the incremental cost of capacity at that time is also zero.

Sidak, in his later work with Spulber (i.e., Sidak and Spulber, 1997), points to a number of fallacies associated with the way costs are incorporated into both pricing and investment decisions. (It is interesting to note that—unlike in the previous collaborative work with Baumol—in this text Sidak makes no claims to be working within the perfectly contestable paradigm).²⁰ In their text, Sidak and Spulber identify such fallacies in light of the deregulation of US utilities, which has led to what they term breaches in the implicit ‘regulatory contract’ between regulators and regulated firms (§6.3.4). In this context, Sidak and Spulber (1997, pp. 419-426) cite a “*fallacy of forward-looking costs*”, a “*fallacy of sunk costs*”, and a fallacy associated with “*ignoring investment-backed expectations*”.

²⁰ Contestability theory is of course discussed at times throughout their text (especially Sidak and Spulber, 1997, pp. 26-28, and 292-295), but the contestable market model is only presented as one of a *number* of possible equilibrium models of competition, including a capacity or Cournot-Nash competition model (e.g., Kreps and Scheinkman, 1983), and a product differentiation model (e.g., Perloff and Salop, 1985). Sidak and Spulber’s divergence from strict contestability theory can also be seen in the limited role which they see sunk costs playing with respect to barriers to entry (§3.5.4).

Sidak and Spulber explain that before a firm makes an irreversible investment, it examines its expected future ‘economic rent’—namely, revenues net of operating costs and investment costs. But once the firm has made the investment, it will continue to produce the good associated with that investment as long as ‘quasi rent’—in other words, revenue less operating costs—is still positive. A *fallacy of sunk costs* occurs when a decision takes into account already-incurred irreversible expenditures, and in this case would occur if a firm decided to stop producing the good if it was unable to meet its entire expected economic rent. Quasi-rent provides the incentive for a firm to stay in the industry *after* entry costs have been “sunk”. This is why “sunk” costs are typically viewed as a barrier to entry; incumbents base their production decisions on quasi-rent, while entrants have to consider economic rent in its entirety (§3.5.1). Sidak and Spulber state that past investment expenditures “do not affect the benefits and costs associated with later decisions; thus such expenditures should not enter into one’s decision-making process”. Although past expenditures in durable assets do implicitly limit the available potentially-efficient choices weighed up in future decisions—due to the intertemporal interdependence of costs caused by the non-fungibility of assets (§3.5.1)—any past *costs* should be not explicitly included in the cost-benefit analysis of future investment or production decisions.

On the other hand, Sidak and Spulber consider that a *fallacy of forward-looking costs* occurs if investment decisions are evaluated either on the basis of costs which are irrelevant to the decision, or ignoring costs which *are* relevant to the decision, particularly opportunity costs; the value of a resource in its best alternative use (§4.1.5). A fallacy of forward-looking costs will occur if a firm (or a regulator requires that a firm) makes its investment decisions on the basis of quasi-rents alone, ignoring the magnitude of investment costs. *Before* the firm has sunk the investment, it is economic rents that count, not quasi-rents. Back within the perfectly competitive paradigm, this difference in decision-making *ex ante* and *ex post* of the commitment to a sunk investment, appears to underlie the earlier SRMC versus LRMC pricing debate (§4.1.3-§4.1.4). Weisman (1991) is one author who clearly articulates the view that *price*, and not just investment and production decisions, is also dependent on this *ex ante* or *ex post* perspective. Weisman suggests that, if consumers have not committed to a particular level of capacity in advance—and some excess capacity is uncommitted from the firm’s point of view—then those consumers are allowed a “free ride” (e.g., in the manner in which Weintraub, 1970, uses the term); the cost of capacity to them is zero, and they need only pay the SRMC of the service. Whereas if they were to let the firm know their level of demand for capacity in advance, then they are duty-bound to pay LRMC.

The critical distinction in this analysis rests on whether the consumers’ commitment to purchase is secured *ex ante* or *ex post* capacity costs are incurred. In the latter case it is sub-optimal to import sunk costs into the marginal cost measure upon which the price for usage is set. On this basis, we conclude that whether SRMC or LRMC is the first-best marginal cost measure depends explicitly on the nature of the sales transaction (and, in turn, the attendant risk allocation) between

buyer and seller. Time is once again paramount, but in this model, the time dimension that is critical is the point in time that the buyer commits to purchase relative to the utility's commitment to a level of capacity (Weisman, 1991).

Notwithstanding such a viewpoint, Sidak and Spulber make it clear that, simply because it is cost-effective for a firm to continue to produce a good if its quasi-rents are positive, does not mean that the price for that good—particularly if that price is regulated—should *not* include any contribution to investment costs. To do so would be to *ignore the investment-backed expectations* of the firm. If the investment has been externally financed, the firm is still committed to paying its lenders a return on, and return of, the loaned principal, implying that prices have to look backward, at least to some extent. Expected economic rent, and not quasi-rent, provides the incentives for a firm to “sink” investment at all, and buyers and sellers enter into contracts on the basis of economic rents, not quasi-rents.

If deprived of a return to capital facilities after capital has been sunk in irreversible investments, or if faced with reduced returns to such investments already made, any economically rational company will have the incentive to eliminate or reduce future capital investments of a like nature (Sidak and Spulber, 1997, p. 502).

As Blaug (1990) highlights in the quote at the beginning of this Chapter, financing cannot simply be ignored, otherwise lenders would not have provided the firm with the capital in the first place. Historic costs are not simply “bygones”. Therefore, Blaug concludes that “long-run marginal costs remain the reference point to pronouncements on optimal resource allocation”. Note, however, that Blaug is not proposing that long run marginal costs provide the basis for pronouncements on *prices*, but for *resource allocation* as a whole.

Yet the fallacy of ignoring investment-backed expectations is rife among those advocates of SRMC pricing who see historic expenditures on capacity as “sunk” and thus “irrecoverable” (§4.1.4). For instance, Andersson and Bohman's (1985) SRMC-pricing prescription explicitly prohibits electricity suppliers from recovering the costs of capacity, even though their investment rule triggers new investment on the basis of consumer willingness-to-pay for that capacity. “Resources invested in capacity expansion cannot be recovered” is their view, as is stated in full in the quote at the beginning of Chapter IV. Andersson and Bohman were silent, though, on why suppliers might be willing to invest under such circumstances, although the possibility of private sector provision of power supply appears to have been entirely ruled out in their suggestion that the existence of indivisibilities was the key reason for electricity provision to be undertaken by *public* utilities.²¹ However, the issue of public versus private

²¹ As Blaug (1990) observes, the “heart of the matter” has traditionally been the question: should *public* enterprises be expected to pay their own way? Blaug phrases the debate on marginal cost pricing (MCP) in the following terms: “Those who advocate MCP, even with many ‘ifs’ and ‘buts’, deny any presumption that public enterprises ought always to make a profit

ownership of utilities is a different question, from whether a regulated firm—already in the private sector—deserves to make a “fair” return on its investment.

Critics of LRMC electricity pricing often point to the fact that LRMC and SRMC are only equivalent at long run equilibrium, and since electricity markets never attain a true equilibrium—due to indivisibilities and other factors—LRMC is not the appropriate equilibrium price. But such critics fail to recognise that the converse is just as true. SRMC pricing suffers from the same problem as LRMC pricing; it too implicitly assumes that capacity is optimally adjusted to demand, in which case the costs of capacity *would* be implicitly reflected in both the short run and the long run marginal costs, since both would be equal. However, because capacity is never optimally adjusted, SRMC takes no account of the opportunity cost of current capacity; historic capacity costs are considered “sunk”, and thus ignored. Yet capacity costs are only “historic” due to indivisibilities—as well as other factors resulting in *optimal spare capacity*, as is discussed in the next Chapter (§6.1.1)—and SRMC pricing provides no “signals” at all to consumers relating to the opportunity costs of using current levels of capacity (§4.1.5). Consequently, dynamic efficiency loses out to a statically efficient price which ignores that capacity is not costless.

The prevalence of the irrecoverable “sunk” costs view is partly attributed by Spulber to Alfred Kahn (1970), who in his highly influential tome on regulation reaffirmed the principle that SRMC is the basis for the efficient pricing of public utilities. Any fixed costs of capacity—including depreciation (i.e., return *of* capital), return *on* investment, property, income taxes—should not be reflected in the price. Commitment to new investments should be undertaken if consumer surplus exceeds capacity costs. *After* capacity costs have been incurred, the static optimum requires that only short run marginal costs are included in the price.²² But as Spulber (1989, p. 235) points out, the strict application of such a rule will

or even to break even ... On the other hand, those who reject marginal cost pricing in any and all of its varieties, maintaining that only average cost pricing provides an accounting check on management and denying that efficiency and equity can ever be separated, end up insisting that every public enterprise must be expected to pay its own way, which paradoxically undermines the very case for public ownership that gave rise to the debate on public utility pricing in the first place”.

²² Yet this view of Kahn’s work, provided by Spulber (1989, p. 235), is overly simplistic. Kahn (2001) himself has stated that SRMC pricing “is a counsel of perfection”. Kahn goes on to say that: “When pricing at short-run marginal congestion costs is infeasible, or is by common consent rejected, the principle of charging on the basis of marginal responsibility requires incorporating in price the *long-run* incremental cost of expanding capacity sufficiently to hold congestion within economically efficient limits. The short-run marginal and long-run incremental costs are effectively substitutes for one another. ... In theoretical equilibrium, the two will be equal. If (short-run) marginal congestion costs were lower than the (long-run) cost of incrementally reducing consumption, it would be inefficient to construct new capacity and more efficient to use the existing facilities more intensively. If or when the short-run marginal congestion costs exceed the cost of relieving congestion by building capacity, the economically efficient expansion path would be construction of additional facilities, and economic efficiency would require users to face a price reflecting the lower long-run incremental cost (including the

affect the regulated firm's financial incentives to invest in that capacity. Consequently, regulators cannot expect to pursue static efficiency without allowing the regulated firm to be compensated for its investments, or firms will underinvest, leading to dynamic efficiency losses.

However, the fallacies which Sidak and Spulber highlight are just as relevant to debates of efficient and equitable prices in countries like New Zealand, which has not had such a history of "regulatory contract". In a report to the New Zealand Treasury on regulating network industries, Williamson and Mumssen (2000, p. 16) point out that firms will only invest if they have a reasonable assurance of cost recovery, in other words, that the cost of committing funds to irreversible investments will be recouped. Nevertheless, even outside the framework of the SRMC versus LRMC pricing debate, the notion that prices should be low where sufficient capacity already exists to accommodate demand growth, is widely held. Perhaps this viewpoint owes some debt to the "common sense" views of still ubiquitous texts on regulation such as Kahn's. With respect to New Zealand's power distribution sector reforms, Kahn's viewpoint has at times been stated explicitly, as is evidenced by the following commentary.

The expenditure on the distribution network is sunk. Therefore, for marginal increases in usage that do not trigger a capacity increase, the correct price signal would seem to be zero—i.e. an extra unit of electricity does not require any marginal increase in the distribution system (until such time as the next subdivision or infill housing scheme is introduced); (Bertram and Terry, 2000, p. 21, [emphasis added]).

Another of Sidak and Spulber's forward-looking cost fallacies is to include *irrelevant* future costs in investment and production decisions. They consider "irrelevant" costs to be any future costs which are unaffected by today's decisions. Just as there is a controversy of the role of historic costs in prices in the SRMC versus LRMC debate, there is also a controversy over the role of future costs. Although future costs are clearly relevant to current decisions, the question remains whether future costs are relevant or irrelevant to current prices, and if so, how.

For instance, Williamson (1966) suggested that, because of asset indivisibilities, when demand is constrained by capacity, price should be greater than LRMC in order to pay for the *next* indivisible item of plant, *in advance* (§4.1.1). Williamson, however, defined LRMC as simply the annuitised cost of existing capacity. In contrast, Turvey and Anderson (1977) embed the cost of future capacity in the LRMC price itself. For them, when demand is not constrained by capacity, LRMC-based prices are directed at *tomorrow's* consumers rather than today's, and thus *signal* the cost of future investments (§4.1.3). And, as seen in the quote opening this Chapter, Schramm's (1991) prescription for pricing

marginal cost of that future addition of capacity)". This explanation is similar to Williamson's (1966) explanation of marginal cost pricing subject to capacity constraints (§4.1.1, §4.1.4, and later in this sub-section).

based on long run incremental cost goes one step further. He requires present day consumers to bear the future *replacement* cost of existing assets as well, even though the period of a particular consumer's demand might be considerably shorter than the lifetime of the durable assets involved (§4.1.5). Which historic or future costs should be included in today's prices are discussed further in the next Chapter (§6.3).

5.3.3 *The “Incremental Cost Fallacy” in Constrained Market Pricing*

In their text, Baumol and Sidak are not exclusively discussing the costs of (“sunk”) *capacity*, but rather *fixed costs*. Over time, the costs of capacity are not necessarily fixed. Unfortunately, Baumol and Sidak do not provide an example where capacity actually is dependent on demand, so it is difficult to surmise how they would view the example above of allocating fixed costs (§5.2.4), where capacity is “joint” to two products, but not fixed. Nevertheless, an examination of Sidak's later work with Spulber suggests that the term “fixed” is applied fairly loosely (at least by Sidak), and is used to describe capacity in a network industry which in its strictest sense is not fixed at all.

Fixed costs are costs that do not vary with fluctuations in output, unlike operating or “variable” costs. The fixed costs of establishing a network system are the costs of facilities such as transmission lines, which are not sensitive to the level of transmission on the lines (Sidak and Spulber, 1997, p. 22).

This statement regarding the fixed cost nature of transmission facilities might suggest that Sidak and Spulber view fixed costs as a long run concept, in addition to being a short run one. In the long run, where demand is increasing, the costs of a power transmission (or distribution) line cannot be considered fixed. Even though part of the total cost can sometimes be considered fixed (e.g., the transmission towers, power poles, and cable trenching; §3.6.2), conductors can be—and often are—changed to increase the capacity of the line. However, Sidak and Spulber go on to “clarify” the distinction between common costs and fixed costs as follows.

Costs that are not attributable to any particular good or service are called *common costs*. ... The greater the proportion of common costs relative to total costs, the greater are the economies of joint production. ... Common costs can be fixed costs, but they need not be. For example, in a simple network system with a single trunk line and multiple distribution lines, the cost of the trunk line is a common cost because it is not attributable to any specific transmission service between any two origination and distribution points. To some extent we can view the trunk line as a fixed cost because it is not sensitive to the usage of the transmission system. In the long run, however, adjusting the system's size to reflect transmission capacity requirements would necessarily entail changing the size of the trunk line. Thus, common costs generally are not fixed costs, because those costs depend on the levels of the outputs of the multiproduct firm. ... The greater the proportion of common costs relative to total costs, the greater are the economies of joint production (Sidak and Spulber, 1997, p. 23).

Sidak and Spulber seem here to imply that common (or joint) costs are generally *not* fixed in the long run, although they confuse the issue by suggesting that the trunk line “to some extent” can be viewed as a fixed cost. Clearly, whether capacity is considered to be fixed or not makes a significant difference to the calculation of SAC and IC. Even in a static world, if common or joint costs are considered to be independent of demand, then as has been demonstrated above (§5.2.4), contributions to those costs do not appear in the incremental cost equations of the *individual* products served by that common capacity. In estimating stand alone costs in their own right, it makes a big difference whether the trunk line can only be constructed with a single capacity, regardless of the number of downstream distribution lines, or whether a number of standard trunk line capacities are available. Simply because network assets may involve indivisibilities does not imply that only a single size (i.e., capacity) exists for a particular equipment type.²³

Common costs are often viewed as physically “non-separable” or “non-attributable” costs. A single indivisible asset which is joint to two products is not separable *in its entirety*, as in Sidak and Spulber’s trunk/distribution line example above. Thus the common costs in their example would appear to be the cost of the trunk line as a whole. The power distribution network example above (§5.2.2) can be used to investigate their trunk/distribution example. The joint capacity can be considered to be a trunk line which serves only two downstream distribution lines with demands q_1 and q_2 . (The cost of each distribution line is not included in the example, but since they are separable costs relating to dedicated assets, including them would not affect the result). Since the cost of the trunk line is subject to increasing returns to scale, the presence of more than one downstream distribution line results in economies of scope. The physically common cost in this example, as Sidak and Spulber themselves state, would therefore be the non-separable cost of the trunk line in its entirety, that is $K(q_1, q_2)$.

Nevertheless, unlike Baumol and Sidak, Sidak and Spulber make it clear that it is *common* costs (rather than fixed costs) which do *not* appear in the incremental cost expressions of *individual* products.²⁴

²³ Although in this case, the trunk line—which is a capital cost—is considered a common cost, in a later example, relating to *local* telephone service, Sidak and Spulber (1997, p. 313) describe the common costs as being the “general and administrative costs (for example, accounting and finance, external relations, and human resources) and support costs, such as general purpose computers”. The incremental (or attributable) costs are cited as including “the costs of central office switching and cable and wire facilities”. Hence, the costs of capacity (i.e., the durable assets) are in this case considered to be incremental (i.e., dedicated), rather than common. On the same page, Sidak and Spulber provide an example of a food stand selling both hot dogs and hamburgers. In this case, the durable assets—the stand and the grill used to cook the products—are considered to be the common costs.

²⁴ In discussing access prices into an incumbent monopolist’s telecommunications network, the New Zealand Ministry of Commerce and The Treasury (1995, para. D11 [emphasis added]) likewise point out that: “An access price set at LRAIC [i.e., long-run average incremental cost] for a *particular product* prevents the monopolist from earning any contribution towards the fixed or common costs on that product”. Additionally, the avoidable (incremental) cost allocation methodology

They state that “we use the term *common costs* to refer collectively to all costs that are not incremental costs” (Sidak and Spulber, 1997, p. 313). This approach is more explicitly outlined in another definition.

The difference between a firm’s total costs and the sum of its incremental costs equals the firm’s shared costs and common costs. ... The firm’s shared costs and common costs are precisely its economies of scope, which means that they are the firm’s efficiency gains from jointly producing multiple services (Sidak and Spulber, 1997, p. 406).

Under this definition, the common costs in the trunk/distribution line example would be determined by subtracting the sum of the individual product incremental costs—the sum of the left hand sides of expressions (5.1) and (5.2)—from the *total* costs (which are also the *physically* “non-separable” common costs, if the costs of the dedicated distribution lines are ignored) in the RHS of (5.3). Under this definition, the *economically* non-separable common costs are $K(q_1) + K(q_2) - K(q_1, q_2)$. This value also represents the economies of joint production (i.e., economies of scope). (Taking the same approach with the peak load pricing problem in (4.15)-(4.17) demonstrates that the economies of scope are $K(q_2)$). The complication is that any economies of scope associated with a particular asset will be smaller than the overall cost of that asset where it is physically non-separable in its entirety. As Parsons (1998) observes, incremental costs are often assessed by considering that they are any costs which are *not* “common”. This can be an unreliable approach since disputes can arise over which costs are in fact considered to be common. If common costs are incorrectly identified, particularly if they are associated with the cost of physically non-separable and indivisible assets, then so too will incremental costs.

In any event, a key proposition which Baumol and Sidak make is that the *time dimension* is *irrelevant* to the calculation of incremental costs in the presence of fixed common costs (§5.3.1). They seem to imply that the subsidy-free revenue bounds derived for the example above would not change if the demands q_1 and q_2 commenced at different times. Fixed costs would not appear in the incremental costs for either product in the event of demand sequencing. This result is highly significant, because Baumol and Sidak (1994a, pp. 82-83) make it clear that the complementary nature of SAC and IC works both ways. Like Heald (§5.2.1), Baumol and Sidak specifically state that incremental costs can be estimated in their own right, and they consider such an approach to be particularly useful if stand alone costs are difficult to calculate; a situation they consider likely to occur frequently. For instance, in discussing the regulation of the US telecommunications industry, Parsons (1998) notes that: “stand-alone cost estimates are rarely attempted and generally considered to be impractical or hypothetical in nature”.

Consequently, Baumol and Sidak (1994a, pp. 81-82) suggest that “stand alone costs can be calculated indirectly with the aid of the much more familiar concept of long-run incremental cost”, a

described in the Handbook for New Zealand’s Optimised Deprival Valuation (ODV), specifically excludes the joint costs of

concept which they point out has been used frequently in rate hearing submissions before the US Federal Communications Commission, and is often called ‘*total service long-run incremental cost*’ (TSLRIC).²⁵ Where long run incremental costs can be determined in their own right, then stand alone costs can, at least in theory, be determined through complementarity (§4.3.2). But if incremental costs are *incorrectly* evaluated in their own right, then stand alone costs will also be wrong, and it is SAC which provides the price ceiling that serves to constrain monopolistic pricing behaviour.

Bradley *et al.* (1999) observe that the measurement of incremental costs “has been discussed infrequently, and it presents some special problems”. They point out that the “fundamental problem in attempting to measure” incremental costs, is that they exist “only as answers to hypothetical questions”. For instance, Loube (1995) affirms that biases can be introduced into constrained market pricing tests depending on the choice made between the two possible approaches for determining incremental cost. He recognises that incremental costs can be estimated using a “*difference method*”—Faulhaber’s approach of subtracting stand alone costs from total costs—or a “*capacity cost method*”, where incremental costs are evaluated directly. Loube suggests that bias might arise under the difference method if stand alone costs are not estimated on the basis of the least cost technology. However, this simply underlines Baumol and Sidak’s requirement that the benchmark for SAC should be a hypothetical efficient entrant. On the other hand, Loube suggests that determining incremental costs directly, by definition, leaves out significant costs, such as the costs of sunk investment. Consequently, Loube provides real world examples where calculating incremental costs from the capacity cost method has resulted in a substantially lower value than using the difference method, because the latter method will require incremental costs to contribute to sunk costs, whereas the former method will not.²⁶ Loube

upstream capacity from the incremental costs of an ELB’s network segment (§7.4.3).

²⁵ The total service long-run incremental cost (TSLRIC) of a service sold to end users is defined by the US Federal Communications Commission as the difference in the firm’s total costs with and without the provision of that service (e.g., Sidak and Spulber, 1997, p. 312). On the face of it, this definition basically sounds like a long run version of Faulhaber’s (1975) incremental cost concept (§4.3.2), as well as being similar to the concept of avoided cost (§5.3.3). However, since—based on this definition—the calculation of TSLRIC would itself appear to require the SAC of the complementary service to be calculated (i.e., the firm’s total costs without the service), it is not clear how Baumol and Sidak would suggest calculating TSLRIC in order to be able to derive complementary SAC. At some point one of either SAC or IC has to be estimated “in its own right”, otherwise the process is chasing its tail. In any event, as is discussed later (§6.4.1), Sidak—in his later work with Spulber—appears to have lost his faith in the ability of regulators to correctly calculate TSLRIC.

²⁶ For instance, Loube (1995) points to regulatory hearings relating to the US telecommunications industry where two estimates of incremental cost for a particular service differed by a factor of 35 to one.

suggests that, if competitive markets are the benchmark, then prices should generally include some contribution to common costs (whether fixed or not).²⁷

The problem appears to be partially one of terminology, as well as familiarity with pre-existing concepts of incremental cost which differ from that associated with cross-subsidisation. “Incremental cost” appears to be a very poor term for the regulated price floor. The very notion of “incremental”—like Baumol and Sidak’s “generic” definition (§5.2.3)—implies that something is being *added* to, rather than taken away from, something else. For instance, Rees and Vickers (1995) define the incremental cost of a set of products as “the *extra* cost of producing that set of products *in addition* to the other products”.²⁸ And as Sidak and Spulber (1997, p. 33) explain, unlike the stand alone cost test, the notion of an “incremental cost test” has been “widely applied for more than a century”, and requires that revenues generated by each service must “at least cover the additional cost of producing that service”. However, this definition of incremental cost was most commonly associated with debates over the presence of ‘predatory pricing’ behaviour (e.g., Areeda and Turner, 1975).

By contrast, Faulhaber’s (1975) definition of incremental cost was unrelated to predatory pricing, and was derived in the context of cross-subsidisation. As is discussed above (§4.3.2), Faulhaber defined the incremental cost of producing a particular set of products in the *absence* of the other products. Moreover, Faulhaber himself explicitly acknowledged the earlier concept of the “incremental cost test”—of which he said “its lineage is ancient”—had been “generally used to examine a *single* product or service, rather than *groups* of products” and as such “can be misleading”. The original concept of the incremental cost test is traced by Faulhaber back to the 1880s, when the idea was applied to pricing the use of “common” railroad tracks. In such an example, common costs clearly are fixed, and incremental costs (at least in the *static* case) would include no contribution to common costs. Familiarity with the railroad example, or others involving fixed costs in both the short and long run, and the older generic concept of incremental cost, may be responsible for some of the confusion.

This confusion is particular evident in Heald’s (1994 and 1997) discussion of cross-subsidies. Even though explicitly acknowledging Faulhaber’s work, and the roots of the SAC test in contestability theory, Heald (1994, pp. 8-11) does not acknowledge the complementarity of SAC and IC. Heald states that: “With only a slight exaggeration, it is possible to characterise FDC [i.e., fully distributed cost; §4.2.3] as an accountant’s method; SAC as an economist’s method; and IC as shared ground”. Heald

²⁷ Similarly, as shown above, evaluating “non-separable” common costs in their own right could end up with a very different value from that found by subtracting all the individual product incremental costs from total costs.

²⁸ For instance, in its entry for “increment”, the Webster Comprehensive Dictionary-International Edition (1998; Ferguson Publishing, Chicago) includes the following definitions: (i) “the act of *increasing*”; (ii) “*that which is added*; increase”; and (iii) “*the amount by which a varying quantity increases between two of its stages*” [emphasis added; also in main text].

sees incremental cost as “shared ground” because “IC systems can allocate *either* accounting costs *or* economic costs”, and he distinguishes incremental cost, which “refers to expansion (producing one more output)” from ‘*avoided cost*’ (§4.2.6), which “refers to contraction (producing one less output)”. Heald points out that “although the two concepts are obviously closely related, the existence of long-lived assets which may constitute sunk costs can lead to marked differences between the two in public utility sectors”. While avoided cost may appear to come closer to Faulhaber’s concept of incremental cost—since it involves subtracting something from something else—in Heald’s description, avoided cost relates to the absence of just as *single* product and, using Faulhaber’s expression, such an approach would be “misleading”. As Baumol and Sidak emphasise, the appropriate SAC and IC tests are combinatorial in nature (§5.2.4).²⁹ Nevertheless, Parsons (1998) observes that, while incremental cost calculations are “relatively common”, “they are almost always limited to the estimation of the costs of providing a single product or service rather than the costs of providing some combination or subset of the total services of the firm”.

Sidak and Spulber (1997, p. 23) themselves reaffirm Faulhaber’s approach by stating that: “The incremental cost of some service *A* is calculated by taking the difference between the total cost of producing a set of services including service *A* and the total cost of producing a set of services excluding service *A*”. Notwithstanding this apparent endorsement by Sidak and Spulber of Faulhaber’s approach, in their critique of the regulatory imposition of TSLRIC pricing in the US telecommunications industry, they raise no objections to the calculation of individual product incremental costs in their own right, seeming to imply that this is acceptable standard practice.³⁰ Nevertheless, among other reasons, Sidak and Spulber (1997, pp. 411-412) eventually end up rejecting TSLRIC pricing on the same basis that Faulhaber highlights traditional estimates of incremental cost as being misleading. “TSLRIC pricing turns out to be a misnomer: It should more appropriately be termed ‘individual-service LRIC,’ for it ignores the incremental costs of *combinations* of services”.

Yet Sidak and Spulber also object to the application—rather than the calculation—of TSLRIC pricing when applied to *individual* services. Although discussing equitable access prices—rather than equitable absolute or relative price levels (§5.1.2)—Sidak and Spulber (1997, pp. 403-407) reject setting the network access price to TSLRIC. Interestingly, Sidak’s earlier endorsement of constrained market

²⁹ Notwithstanding the fact that the New Zealand Government’s discussion paper on access pricing for natural monopolies directly references Baumol and Sidak’s (1994) text, stand alone cost is defined as “the cost of producing *one* product line *on its own*, using best, forward looking technology”. Similarly, incremental costs are considered to “arise as a result of the provision of the ‘increment’. In this document, the increment refers to *a* service” (Ministry of Commerce and The Treasury, 1995, Appendix IX, [emphasis added]). Also note the emphasis in the quote in fn. 24.

pricing with Baumol has been set aside. (However, this is implicitly justified by the assertion that pricing an incumbent's *outputs* sold to competitors, differs from pricing the incumbent's *inputs* sold to competitors—in other words, network access).³¹ While constrained market pricing would accept any price between (and/or equal to) IC and SAC as equitable, Sidak and Spulber feel that setting the network access price for a potential entrant on the basis of TSLRIC is unfair to the incumbent firm.³² TSLRIC pricing “does not equal economic costs”, simply because TSLRIC—by their definition—does *not* include any contribution to common costs. The circularity inherent in this objection would vanish if it were demonstrated that individual service incremental costs in many cases *do* contribute to common costs after all, as Loube (1995) suggests is the case.

Sidak and Spulber begin to identify the problem by stating that: “The definition of attributable cost is partly at issue”, although they consider the problem of cost attribution to arise due to the inappropriate use of “accounting cost” instead of “attributable economic cost” (which includes the opportunity cost of the firm's inputs). On the other hand—at least with respect to the identification of cross subsidies in the provision of postal services—Bradley *et al.* (1999) suggest that the traditional “attributable cost measure is inferior for checking cross subsidy and propose instead the use of incremental cost”. They observe that the concept of “attributable cost assumes that marginal cost is constant over the entire range of a product's volume”. Hence, for services which “are subject to increasing returns to scale, this means that attributable cost understates the total cost caused by a product”.

Since Sidak and Spulber associate common costs with non-attributable costs—and common costs are seen as being the complementary set to the sum of all individual product incremental costs—this would imply that, in their view, incremental costs and attributable costs are also equivalent. By contrast, Bradley and his colleagues are suggesting that, in the presence of increasing returns to scale, some of the “non-attributable costs” should be “attributed” to incremental costs after all. Nevertheless, while making

³⁰ Sidak and Spulber (1997, p. 404): “Our critique of TSLRIC pricing of mandatory network access presupposes that incremental costs are calculated correctly. Incorrect calculation of incremental costs exacerbates the problems with TSLRIC pricing that we identify”.

³¹ It also relates to the case where constrained market pricing is applied to the incumbent, but not to potential entrants. “It is well known that in price regulation, ‘a ceiling can become a floor and a floor can become a ceiling’. With regulatory price floors set on the incumbent, entrants could price just below the floor, tacitly coordinate their prices, and capture consumers from the incumbent while avoiding competition themselves. Eliminating the incumbent's price floor injects the incumbent into the market as a credible rival and allows the market price to fall freely, as true competition requires” (Sidak and Spulber, 1997, p. 506).

³² Sidak and Spulber (1997, p. 406): “The firm's shared costs and common costs are precisely its economies of scope, which means that they are the firm's efficiency gains from jointly producing multiple services. To price without regard to those costs is to penalize a firm for its efficiencies”.

it clear that “product-specific costs [i.e., attributable costs] constitute a trivial portion of the incremental costs of most products”, Bradley and his colleagues also reaffirm the position that “common costs are not part of the incremental costs of any” product. Economies of *scale* are the reason why incremental costs are greater than attributable costs based on solely product-specific fixed and variable costs. It appears that Bradley and his colleagues do not seem to feel that economies of *scope* would have a similar impact on incremental costs.³³

5.3.4 Toward Intertemporal Subsidy-Free Prices: Option Value and Opportunity Cost in Investment

Attempting to evaluate incremental costs in their own right ignores the ‘*internalities*’ between producing a particular set of products and the other products due to economies of scale and/or scope (§3.4.2). Spulber (1989, p. 54)—although not applying his definition to the issue of cross-subsidies—defines an internality as the “costs and benefits of a transaction experienced by the parties of the transaction that are not accounted for in terms of the exchange”. In the presence of economies of scale or scope, incremental consumers can provide a benefit to other consumers, because they reduce the average cost of supplying *all* consumers.³⁴ If these benefits to other consumers are not accounted for in the terms of exchange—in other words, the incremental costs do not include these benefits—then an internality arises. Sorenson *et al.* (1976) described a similar sort of internality—although in very different language—which arises between peak and off-peak consumers in their game theoretic model of the peak load pricing problem, in the presence of economies of scale or scope (§4.3.5). Unlike the peak load pricing problem, consumers of an incremental product (as opposed to an off-peak product) might add both a benefit *and* a cost to the game, not just a benefit. Nevertheless, where declining average incremental costs are present, then this may make the consumers of the incremental products potentially valuable players, in the same way that off-peak consumers are valuable in the peak load pricing game.

However, where demand of *future* consumers requires that spare capacity be *optimally* built *in anticipation* of that demand (§6.1.1), then future consumers raise the costs of current consumers, even though they may lower the costs of those consumers who are still around when the future consumers begin to demand the product associated with that spare capacity. This would imply that even if spare capacity exists, the incremental cost of allowing new consumers access to that capacity is not zero, because the *opportunity cost* of building surplus capacity in anticipation of later demand is positive

³³ It is not clear whether Baumol and Sidak’s (1994a) affirmation that incremental costs include no contribution to fixed common costs might be due to an implicit assumption of *constant* returns to scale. In a different context, Economides (1996) suggests that Baumol and Sidak’s work is only “correct under a strict set of assumptions”, including “that there are no economies of scale in either one of the complements”. (By “complements”, Economides is referring to an upstream monopolistic service being a complement to a downstream competitive service).

(§5.3.2).³⁵ There are alternative uses for the capital committed to investments in spare and unutilised capacity. Those costs are incurred at the present time, but they are attributable to “incremental consumers” at some later date. On the other hand, if present day consumers were to delay their consumption until that later date, then there would also be no need for spare capacity (or in fact any capacity) at the present time. Consequently, there is likely to be a “value of waiting to invest” (McDonald and Siegel, 1986). This value is often termed the ‘*option value*’ inherent in the ability of the incumbent firm to delay investment (particularly if, by delaying investment, the firm is able to base its decision on more reliable forecasts about the future).³⁶ Today’s consumers have to be prepared to contribute to the opportunity cost of spare capacity, otherwise the option value might be sufficiently high for the firm to actually decide to delay its investment. Hence, the costs of spare capacity are also attributable to some extent to *existing* consumers, and not just to future consumers.³⁷ The decision variables include the initial consumption dates of *all* consumers, not just future consumers, since every consumer’s consumption affects the firm’s attempt to meet demand over time at least cost.

Sidak and Spulber’s (1997, p. 23) own definition of “incremental cost” is of a “cost attributable to an individual product or group of products”, to distinguish it from “common costs”, which are unattributable. Although this makes sense in theory, the problem is that, in practice, costs may be difficult to attribute to products, particularly where intertemporal effects are present, as the notional example of spare capacity just provided suggests. Since the benefits and costs associated with incremental demand for capacity are difficult to estimate, or even conceptualise, in their own right, the best approach is to apply Faulhaber’s original approach, but incorporate intertemporal factors. This

³⁴ In the literature relating to networks, benefits which arise from demand growth (or from additional consumers) in the presence of economies of scale (or scope) are often termed an “indirect network externality” of a “one-way network” (e.g., Economides, 1996).

³⁵ This is not dissimilar to Baumol and Sidak’s logic as to why access prices based on the ECPR should include an opportunity cost contribution to the cost of common assets (fn. 6).

³⁶ The “real options theory” of “investment under uncertainty” has its clearest articulation in Dixit and Pindyck’s (1994) major text, but elements of the concept in the electricity pricing debate can be seen in the original definitions of long run marginal cost by authors such as Turvey in the late 1960s (§4.1.3). Small and Ergas (1999) attribute the ideas, if not the terminology, to Kenneth Arrow (1968). As Small and Ergas point out, the firm will not invest at all if the firm expects to profit from delaying the investment. Turvey (1969), however, attributes the concept that “the cost of producing any given flow of output can be reduced by postponing the period in which delivery is made” to Oi (1967), the instigator of more recent work on optimal two-part tariffs (§4.2.4).

³⁷ Salinger (1998) addresses this issue in the context of access pricing (§5.1.2). Where regulators mandate access prices for an incumbent firm that do not include forward-looking costs such as the cost of (optimal) spare capacity, then *competitors* are gaining an option value—related to that excess capacity—for which they are not paying. If entrants do not contribute to the costs of that excess capacity, then the incumbent will have little incentive to invest in that capacity even if it is optimal. The same applies to existing *consumers* rather than competitors. In either case, ignoring this option value of spare capacity would be a fallacy of forward-looking costs.

would mean deriving a set of bounds on ‘*intertemporal* subsidy-free prices’. The ‘*intertemporal standalone cost*’ of a particular set of products would be evaluated directly—taking into account the demand of current and future consumers—as would the zero profit constraint over the long run (i.e., the intertemporal total cost). Then the intertemporal incremental cost of the complementary set of products could be found by subtracting the intertemporal stand alone cost from the intertemporal total cost of supply.

The premise of this thesis is that both Heald’s, as well as Baumol and Sidak’s, position that incremental costs do not contribute to common costs—a view which can even be problematic in a static sense, given the difficulty in determining common costs—is particularly problematic in an intertemporal world, for two reasons. Firstly, Heald’s conclusion (§5.3.2) does not explicitly recognise the role which *spare capacity* plays in dynamically efficient (i.e., optimal) network investment. Baumol and Sidak do not mention the role of spare capacity at all. Secondly, neither position adequately takes into account the impact of the *time dimension* on the subsidy-free prices themselves. The result of the neglect of the full impact of the time dimension is that—as Loube suggests for the static case—significant costs can become left out of the estimation of incremental cost. Given the apparent misapplication of the time dimension in constrained market pricing, the remainder of this thesis mostly focuses on the question of the *relative* efficiency and fairness of price levels in a *temporal*—rather than a locational—sense. This is somewhat similar to examining *absolute* (but *static*) price levels over time, although in this case, the influence of the intertemporal interdependence of costs are explicitly taken into account. Hence, the next Chapter tackles the issue of intertemporal subsidy-free prices in the context of dynamic efficiency and optimal investment.

CHAPTER VI

SUBSIDY-FREE PRICES AND OPTIMAL INVESTMENT: TOWARD STATIC OR INTERTEMPORAL EQUITY?

Much of the discussion has stressed the properties of static economic efficiency of the regulatory rules proposed. The constrained market pricing approach to regulation is intended to be a workable program that seeks to prevent interference with the achievement of either static efficiency or growth performance: William Baumol and Gregory Sidak (1994a, p. 142)

The principle of sale at marginal cost is applicable to existing plant but cannot alone govern investment policy. It is investment policy which must decide plant expansion and reconcile the apparently incompatible needs of short-term and long-term demand. Long-term pricing must guide the customer's decisions according to what it would cost if plant were constantly maintained at optimum capacity and not according to momentary conditions of overequipment or underequipment: French pioneer of marginal cost pricing, Marcel Boiteux (1949)

The primary purpose of the price ceiling, aside from its role in eliciting economic efficiency, is to protect consumers from monopolistic exploitation: William Baumol and Gregory Sidak (1994a, pp. 51-52)¹

As in the perfectly competitive paradigm, contestability theory looks upon time more as a *static* rather than a *intertemporal* concept. But as some of the commentators on New Zealand's power sector reforms caution, too much attention on efficient prices can draw attention away from efficient investment (§2.5.2). This Chapter examines optimal investment in distribution networks, and highlights that spare capacity can be part of a dynamically efficient construction configuration, particularly given the intertemporal interdependence of costs stemming from the non-fungibility of distribution assets (§3.5.1). The dynamic efficiency aspects of constrained market pricing are then briefly evaluated in the presence of changing technology and, more importantly, changing *demand*. As a consequence of this brief analysis, it is suggested that ignoring intertemporal equity considerations is likely to result in SAC and IC values which can allow considerable pricing freedom for an incumbent monopolist.

6.1 Dynamic Efficiency: Optimal Investment and Intertemporal Subsidy-Free Prices

6.1.1 Dynamically Efficient Spare Capacity and Asset Duplication in Distribution Networks

“Unused”, “spare”, “idle”, “excess” or “surplus” capacity in any industry—including network industries—is often regarded in a pejorative sense. As has been noted earlier (§3.5.1), excess capacity is considered to act as a potential barrier to entry which can distort markets away from the perfectly contestable ideal. Even the peak load pricing problem arises because durable capacity common to two or

¹ All quotes have been abridged for clarity, and acronyms spelt out in full.

more periods is optimally sized to the peak period, but exceeds demand in off-peak periods. Consequently, although the allocatively efficient solution to the problem is to have lower—although *not* zero—charges for capacity in off-peak periods (§4.1.2), the result also appears to make some intuitive sense, in that these lower prices will make increased demand in the off-peak period attractive, presumably leading to more efficient utilisation of the idle capacity. This has sometimes led to prescriptions that greater utilisation of excess capacity should be “encouraged”, even if this requires deviating from the strictly efficient price. In other words, encouraging the use of spare capacity—perhaps by manipulating price—might be beneficial in its own right.

For instance, Schramm (1991)—although an advocate of long run incremental cost pricing for electricity (§4.1.5)—suggests that, “if excess capacity were to exist in the system at any time (say, because of erroneous forecasts or simply because of the lumpiness of new investment components), it would make perfectly good sense to try to sell this surplus capacity at or above short-run marginal cost to whoever is able to absorb this surplus”.² Similarly, the architects of New Zealand’s power sector reforms—in an early discussion paper on the appropriate philosophy for deriving distribution line charges—proposed that price “incentives” could be given “to encourage use of under utilised capacity” (Ministry of Commerce, 1991, p. 58).

In these cases, spare capacity makes increased consumption attractive because the surplus causes economies of scale or economies of fill (§3.4.2). However, these economies are *short run* in nature—because capacity is temporarily assumed fixed—and arise because capacity is not optimally adjusted to demand (during off-peak periods). Yet, apparently “excess” distribution network capacity can itself be a “good thing” for a number of reasons—especially its contribution to dynamic efficiency. Banks (1994), for one, considers that the need for spare capacity in a power system has a somewhat greater significance than it does in most other markets. Likewise, Sidak and Spulber (1997, pp. 126-127) acknowledge that “excess” capacity can be beneficial in electricity supply, particular due to its reduction of the risk of power outages.

In particular, the cost of ‘unserved energy demand’ can be substantial (e.g., Munasinghe and Sanghvi, 1988). Since the level of future electricity demand is subject to uncertainties, some excess capacity and/or redundancy (i.e., asset duplication) is likely to be optimal.³ Recognising the stochastic nature of demand for electricity, Carlton (1977) suggested that price could optimally be above LRMC in

² Likewise, Anderson and Bohman (1985) draw attention to Munasinghe and Warford’s (1982) prescription that, when excess capacity occurs, demand charges could be reduced below the long run marginal cost level until the demand grows sufficiently.

³ Turvey (1969) also explicitly pointed out the impact of uncertainty in demand on capacity and capacity cost, and attributed this notion to Stigler (1939).

order to recoup the costs of having had under-utilised capacity up to that point. Depending on the location of any “bottlenecks” throughout the power system, should price or quantity rationing not successfully keep demand to a level below installed capacity, then not only will some demand be left unserved, but a blackout could be triggered, affecting many more consumers than just those associated with the marginal demand. This will result in a significant “disruption cost” or “outage cost” (e.g., Kleindorfer and Fernando, 1993) over and above the cost of unserved energy. Concerns about such costs may cause some consumers to be directly willing to pay for high levels of reliability, particularly with respect to the design of their dedicated assets (§3.1.2).

In discussing power distribution network planning (§3.1.3), Lesser and Feinstein (1999) point out that an optimal distribution network expansion plan must be optimal with respect to the *uncertain* demand, not to the *expected* demand. Echoing Boiteux’s concept of the greater “individuality” of the distribution network (§3.1.1), Lesser and Feinstein note that demand uncertainty will tend to be exacerbated at the local distribution network level. Consequently, estimating the optimal level of spare capacity for reliability purposes in the collective and semi-individual distribution networks is somewhat problematic. Empirical evidence strongly suggests that the majority of consumers tend to attach a much higher cost to a marginal *decrement* of service reliability, than they are willing to pay for a marginal *improvement* in reliability (e.g., Hartman *et al.*, 1991). But as Crew *et al.* (1995) warn, ignoring the costs of disruption entirely will lead to an under-investment in capacity.

Nevertheless, such problems of demand uncertainty are shared to varying degrees by all aspects of any power system, and particularly relate to widely-analysed problems of ‘loss of load probability’ for generation plant (e.g., David and Li, 1993a). But in distribution networks, there are a number of additional justifications for spare capacity apart from improved security due to uncertainty.⁴ Firstly, efficiencies in distribution equipment are obtained by manufacturing them in standard discrete sizes, and distributors can themselves reduce capital and maintenance costs by standardising on equipment (Wilson, 2000). This is the ubiquitous issue of asset indivisibilities. Moreover, because energy losses in network equipment cause decreasing returns to scale, the cost of losses makes operating at full capacity economically inefficient (§4.2.6). Finally, economies of scale and scope can make ‘*anticipatory construction*’ optimal. All these factors may lead to apparently low levels of utilisation for distribution network assets, which nevertheless, could be optimal.

⁴ Increased reliability through asset duplication—particularly by paralleling equipment—is more applicable to transmission and distribution than to generation. Given the combined influences of the cost of losses and reliability, duplication of assets may be a more efficient solution than a single larger asset—with some built-in excess capacity—performing the same function. Thus, an optimal distribution system may include both spare *and* duplicated capacity. Conversely, if the system is a sub-optimal system anyway, then duplication of assets may not necessarily be “wasteful”, since it may improve the characteristics of the existing network configuration.

Like many issues relating to the economics of electricity supply, the significance of excess capacity was also recognised by Boiteux (1949). He considered that uncertainty in demand would make some excess capacity desirable, but acknowledged that cases of “unintentional overequipment” could sometimes arise due to erroneous demand forecasting (as could unintentional underequipment). However, Boiteux also highlighted cases of “*deliberate over-equipment in anticipation of a subsequent development of demand*”.⁵ Rather than being associated with short or medium term disparities between actual and forecasted demand, Boiteux stated that in some circumstances it could be far less costly to bear the financial charges and amortisation of substantial surplus capacity for many years, than to expand capacity at some much later date. Hence, Boiteux emphasised that spare capacity “*has its own income*”. Boiteux said that, “in the language of economists, such deliberate overequipment is an ‘arbitrage activity’ and not a ‘production activity’”, seemingly an identification of the intertemporal interdependence (i.e., “arbitrage”) caused by investments in durable non-fungible assets (§3.5.1). Consequently, even though Boiteux proposed that capacity should be priced as though it were optimally matched to demand (§4.2.6)—as also noted in the quote at the beginning of this Chapter—this by no means ruled out the possibility that this “shadow” (or hypothetical) optimal capacity would contain some (possibly substantial) level of optimal “excess” capacity.

6.1.2 Constrained Market Pricing, Contestability Theory and Dynamic Efficiency

So far, the discussion relating to cross subsidies has not explicitly addressed the *time dimension*, only a rather vague notion of “demand sequencing”. Jamison (1996), for one, highlights that Faulhaber’s formulation of subsidy-free prices ignores concerns which may be important to policymakers, in particular, “dynamic efficiency” (although he does not propose a solution to this problem). Meyer and Tye (1985) suggest that where prices far exceed costs as a result of Ramsey or other pricing rules, then this can induce investments which are not economically efficient in the long run. And in a broader sense, Witteloostuijn (1990), for one, observes that “the contestable market framework serves well as a *static* benchmark case of competition”. Nevertheless, unlike Faulhaber’s definition of incremental cost, Baumol and Sidak’s definition of AIC (§5.1.4) does include some allusion to the time dimension, since they indicate that AIC is the “long-run figure obtained after plant and equipment are adjusted so as to minimise the average of the pertinent output”. Similarly, the time dimension creeps into their stand alone cost concept.

[T]he firm is never permitted to adopt a price so high that it could not prevail in a perfectly contestable market, but it is allowed to set a price at any level that could prevail *in the long run* in such a market (Baumol and Sidak, 1994a, p. 51, [emphasis added]).

⁵ Emphasis added. Miller (1995) also cites excess capacity as being a common characteristic of network utilities, “if only because facilities typically are constructed in advance of demand”.

However, such definitions appear equivalent to the *static* concept of the “long run” associated with the perfectly competitive market model (§2.1.4), since Baumol and Sidak’s total cost function is $TC(x,y,z,\dots)$, rather than $TC(x,y,z,\dots, t)$. And as Parsons (1998) observes: “With regard to the ‘long-run nature of the cost calculation, the economic literature on cross-subsidization is largely silent”.⁶ In fact, notwithstanding their claims that their approach to constrained market pricing fosters innovation and enhanced productivity (§6.4.1), as the quote opening this Chapter suggests, Baumol and Sidak—like Faulhaber (1975) before them (§6.3.2)—are frank that the focus of their regulatory prescriptions is on encouraging *static* efficiency.

Much of the discussion has stressed the properties of static economic efficiency of the regulatory rules proposed. But the flexible earnings permitted under the [constrained market pricing] approach are designed to supply the incentives required to elicit investment in innovation and enhanced productivity. Thus the program also addresses itself to the encouragement of improved intertemporal performance. In sum, the constrained market pricing approach to regulation is intended to be a workable program that seeks to prevent interference with the achievement of either static efficiency or growth performance (Baumol and Sidak, 1994a, p. 142).

In describing the appropriate price ceilings for regulatory purposes, Baumol and Sidak (1994a, p. 78) remind their reader that, “like price floors, the price-ceiling standard applies not only to the products of the firm considered one at a time, but to every combination of the services of the firm”. Yet, they do not suggest that the combinatorial SAC test should apply to any but the *existing* products of the firm. However, where investments are non-fungible, future costs will be intertemporally interdependent on current capital outlays, and both will depend as much on *future* demand as on existing demand. As Boiteux recognised (§6.1.1), non-fungible capacity will not necessarily be optimally sized if it is designed solely to serve present day demands.

Notwithstanding the criticisms of traditional pricing methodologies such as fully distributed cost, on the basis that they ignore *demand* considerations (§4.2.3), the contestability theorists consider that the “weak invisible hand” will ensure that demand side issues will be implicitly taken into account under constrained market pricing through the regulated firm’s own decisions (§4.3.6). Like their admission that they focus on static efficiency, Baumol and Sidak are also quite frank—as were BPW (p. 508) before them (§5.1.1)—that they make no claims regarding the impact of demand on the subsidy-free price bounds, and it is not intended that the floor-ceiling calculations described in their text include any data relating to demand. Only *cost* information is needed, and none of that information entails any cost

⁶ Parsons (1998) also observes that “the economic literature on cross-subsidization is silent on the specific application of forward looking (rather than accounting or embedded) costs”.

allocation (Baumol and Sidak, 1994a, p. 79). As described earlier (§4.3.1), ignoring demand is a key principle of the methodology they prescribe for regulating the prices of natural monopolies.

Such an approach might be acceptable in a static world, as long as cross-elasticities of demand are small (§4.3.6). But in an intertemporal world, dynamic efficiency requires that investment decisions consider future demand, since it is demand—both current and future—which drives investment and thus costs (§3.1.2). Where assets are predominantly non-fungible, even if *demand* is intertemporally independent (i.e., cross-elasticities of demand are actually zero), *costs* will still be intertemporally interdependent (§3.5.1). Since present day decisions relating to investments in physically common or joint non-fungible capacity do affect future costs, which themselves are also dependent on future demand, it would be a fallacy of forward-looking costs to exclude future demand or future costs from the decision-making process (§5.3.2). And as Baumol and his colleagues admit (BPW, p. 429), when faced with “the underlying dynamic of the process of evolution of industry structure and ownership”, they “cannot be comfortable with a standard equilibrium analysis of industry structure, or feel confident that the invisible hand has matters here firmly under control”.

Admittedly, neglecting demand will still produce upper and lower *bounds* to subsidy-free prices. However, as is shown in Chapters VIII and IX, such bounds will only be necessary, but not sufficient, constraints to the true set of intertemporal subsidy-free prices. Hence, the practical concern that constrained market pricing can result in a very wide range of potentially efficient prices is valid (§5.1.3), since SAC and IC tests which ignore intertemporal effects will usually allow the incumbent firm great pricing freedom. However, this thesis demonstrates that introducing intertemporal considerations reduces the gap between SAC (i.e., price ceiling) and IC (i.e., price floor) values considerably (Chapters VIII and IX). As such, SAC and IC calculations which ignore demand are really only useful for *rejecting* prices which fall outside the bounds, rather than accepting prices which fall within them.

Unfortunately, Baumol and Sidak recommend stand alone costs to be calculated from estimates of incremental costs over the long run (§5.3.3). While such long run incremental costs (LRIC) will take account of some future costs, only those future costs which are unaffected by future demand will be included (§6.1.5). Where future demand is ignored, incremental cost estimates can be much lower than they would be were demand included. Conversely, the allowable price ceiling, and the level of supposedly “subsidy-free” prices, will be much higher. Ironically—depending on how the zero profit constraint is evaluated—a productively-*inefficient* firm may have even greater pricing freedom, since Baumol and Sidak (1994a, p. 83) indicate that, if a firm is not an efficient supplier, then the incremental cost figure should be adjusted *downward* by an appropriate amount.

For dynamic efficiency, and intertemporal anonymous equity, stand alone costs (and by complementarity, incremental costs) need to be evaluated on the basis of both current and future demand,

and the costs associated with those demands. As this thesis demonstrates (i.e., Chapters VIII and IX), assessing such incremental costs is *not* intuitive in an intertemporal world. Furthermore, since the zero profit constraint is equivalent to the SAC constraint associated with the natural monopolist's entire market, this would imply that the *absolute* price level of the firm should be determined on the basis of an asset configuration which optimally serves the current and future market demand. Finally, due to the complementarity of SAC and IC tests, the timing of future demands will also impact incremental costs; incremental costs will need to be determined by subtracting the stand alone costs associated with the current and future market state, from a zero profit constraint also applicable over the long run.

Note that the “vanilla” or “true” stand alone cost of providing a service can be considered the cost incurred if there were no demand in the future. The terminology used here does not imply that the stand alone cost itself necessarily changes as a result of future demand, but SAC remains a useful shorthand for recognising the effect intertemporal interdependence has on the complementary intertemporal SAC (and IC) tests which constrain intertemporal subsidy-free prices via price ceilings (and price floors). Baumol (1986, pp. 115-120) used similar terminology for describing incremental cost (i.e., “net incremental cost”) to recognise the relationship between “vanilla” IC and the effect of cross-elasticities of demand on subsidy-free prices (§4.3.6). From complementarity, there must be an associated “net” stand alone cost.

6.1.3 The Zero Profit Constraint over the Long Run

Although they do not require the cost-based price ceiling and floor values to account for demand, Baumol and Sidak do make the time dimension explicit in their description of the zero profit constraint. They consider that there are three important factors involved in determining the total revenue/cost equality (Baumol and Sidak, 1994a, p. 55). Firstly, although the cost of capital must be included in all costs, they suggest that no single figure can be taken to constitute the “proper” universal value for the current competitive rate of return. Risk must be taken into account when determining such a figure for a particular industry.

Secondly, even in contestable markets, firms with superior innovation and productivity growth can expect as their reward earnings temporarily *exceeding* the cost of capital. Appearing to step entirely outside the neoclassical paradigm and borrow concepts from the neo-Austrian school of economics—ideas that do not appear in the standard contestability text, since perfect contestability requires that firms make zero economic profits (e.g., Hay and Morris, 1993, p. 576)—Baumol and Sidak consider that positive economic profit “is the incentive for the firm to undertake the effort and risk entailed in the innovation process”.⁷ The idea that profits are a short-term phenomenon arising from successful

⁷ This notion of firms as a “purposeful agents taking advantage of the profit opportunities arising from market disequilibrium” is a marked change from standard contestability theory (i.e., BPW) which, like the model of perfect competition, is a static

entrepreneurial activity—rather than persistent monopoly—have been suffusing into the mainstream literature on regulation through the works of economists such as Stephen Littlechild (1981). Littlechild has had a major influence on the theory and the practice of regulating monopolies (e.g., Berg and Tschirhart, 1995), and a particular area of focus has been the regulation of power distribution utilities, given his position as the head of the United Kingdom’s Office of Electricity Regulation (OFFER) for much of the 1990s (e.g., Kennedy School of Government Case Program, 1998). Littlechild is frank that he views the neo-Austrians’ research programme as providing promising insights for the regulation of monopolies.⁸

Finally, and most significantly for intertemporal considerations, contestable industries can expect to earn a return equal to their cost of capital (i.e., zero economic profit), but only “on average” and “in the long run”. This harkens back to definitions of long run marginal cost as requiring an assessment of future costs in *present value* terms (§4.1.3). Baumol and Sidak affirm that it does not violate the rules of even perfect competition that a firm which loses money during a recession recoups those losses through earnings that temporarily exceed the cost of capital during an ensuing period of prosperity.

model of equilibrating markets. While standard contestability theory holds the similar view to the neo-Austrians that the number of firms has no direct bearing on the existence of monopoly, it includes no such notion of “market discovery”. Baumol and Sidak appear to be invoking the neo-Austrian dynamic vision of competition, rather than a static one. Although not abandoning the notion of equilibrium, which many of the neo-Austrian school do—for example, Joseph Schumpeter (see Ekelund and Hébert, 1990, pp. 567-570), and O’Driscoll and Rizzo (1985)—Baumol and Sidak appear here to view the firm itself as the equilibrating mechanism, a position taken by the “middle-of-the-road” neo-Austrian, Israel Kirzner (e.g., Kirzner, 1973, 1985 and 1994). As Ekelund and Hébert (1990, p. 576) recall, this concept goes back to Cantillon: “Somewhere along the way in the evolution of economic theory, neoclassical economists had forgotten or ignored Cantillon’s original vision of the market as an arena in which market participants (i.e., entrepreneurs) nudge prices in the direction of equilibrium by exploiting profit opportunities offered by disequilibrium prices. This vision has been more consistently grasped and maintained by Austrian economists than by any other group”.

In his later work with Spulber, Sidak moves away from this earlier middle-of-the-road neo-Austrian perspective, toward that of Joseph Schumpeter (1950). Sidak and Spulber (1997, p. 45) refer to “the Schumpeterian process by which superior production technologies continuously vie to displace inferior ones”, and warn regulators against interfering with this process. Sidak and Spulber (1997, pp. 518-519) also show deference to Schumpeter’s market process of “creative destruction” where “the pursuit of market power is a creative, dynamic force that incessantly revolutionizes the economic structure from within”. Sidak also cites one of the fathers of the neo-Austrian paradigm, Friedrich Hayek, in rebutting the view that regulators can discern price levels more accurately than markets can (§6.3.1). (Some of the comments in this footnote are taken from the present author’s brief examination of contestability theory in the context of the neo-Austrians, in Gunn, 1995b).

⁸ For instance, in examining the experience during the 1980s of regulating privatised monopolies in the UK, Littlechild cites the neo-Austrians directly: “Future research might usefully reflect the Austrian insistence on profit as the engine of capitalism and, in particular, on the hitherto unforeseen profit opportunities as central to the continuing market process (Schumpeter, 1950; Kirzner, 1973, 1985)”; (Beesley and Littlechild, 1989). Rees and Vickers (1995) point out that Littlechild had felt that regulation in the UK should only be a “stop-gap” until sufficient competition arrived.

Consequently, they prescribe that any “regulation that undertakes to follow the contestable market model and prevent excess profits, must not adopt rules that prohibit any of these three forms of deviation from zero economic profits, lest the result damage the public interest”.

Acknowledging this long run nature of the zero economic profit constraint, Teplitz-Sembitzky (1990, p. 38) notes that costs under either the stand alone cost criterion, or the incremental cost test can be determined on a present value basis. However, for his two product example of subsidy-free prices it is assumed that demand for both products begins at the same time—there is no sequencing of demand. Similarly, Heald (1997) also acknowledges that judgements about cross subsidy can only be satisfactorily answered by means of discounted cash flow appraisals “which look beyond annual accounting data and which explicitly acknowledge the time dimension of cash flows”. Curien (1991) also points out that assessing cross subsidies for new investments on a year-by-year basis can be misleading, because its profitability may need to be evaluated over a long time period, and possible subsidies from existing services need to be examined over that time span. Curien asserts that not to allow the firm to practice temporary—or yearly cross-subsidies—would in fact amount to forbidding any restructuring of its activities”.

6.1.4 The Optimal Investment Rule versus “Greenfields” Optimality

Likewise, although not addressing the issue of subsidy-free pricing, Lesser and Feinstein (1999) imply that pricing for dynamic efficiency is unlikely to result in a “smooth returns” path. Specifically discussing distribution network planning, they maintain that, simply because something is “*cost-effective*”—for example, the net present value of a proposed investment program is positive—does not mean that it is *optimal*. This is important, because in evaluating intertemporal subsidy-free prices, the total revenue/cost equality only ensures dynamic efficiency if the investment program to which it relates is optimal.

It is well-known that a dynamically optimal policy is not a sequence of short-run optimal decisions, much less a sequence of cost-effective (non-optimal) decisions. Dynamic optimality requires making intertemporal tradeoffs; such tradeoffs may result in short-term sacrifices in order to yield long-term benefits. Yet the avoided cost methodology suggests that it is best to invest in assets that are cost-effective each year. This ‘greedy’ (second-best) approach to decision making is almost never optimal. Nor is it necessarily anywhere near optimal. There is simply no logical relationship between the short-term greedy deferral solution and the long-term optimal policy: Lesser and Feinstein (1999).

Interestingly, Lesser and Feinstein’s paper has no intention of contributing to the standard SRMC versus LRMC debate—they are attempting to educate US regulators regarding the complexities involved in optimal investment planning for distribution networks—and they reference none of the key literature relating to the debate (§4.1.3-§4.1.5). Moreover, their allusion to “short-term” and to “second-best” are

not made in the context of either SRMC pricing or Ramsey pricing. Nevertheless, they indirectly provide weight to criticisms of LRMC-based pricing, since they identify the “avoided cost” basis to calculating the marginal price of distribution capacity (§4.2.6) as being simply the “average cost” associated with the deferral of an “investment of arbitrary capacity”.⁹ As has already been seen (§4.1.4), this view—that the standard definition of long run marginal capacity cost is not marginal after all, but average instead—forms the basis for one of Andersson and Bohman’s (1985) key criticisms of LRMC pricing.

Nevertheless, Andersson and Bohman (1985) advocated the same ‘*optimal investment rule*’ as both LRMC pricing advocates such as Turvey on the one hand, and spot pricing theorists on the other, namely that marginal capacity costs should be equated to marginal benefits attributable to that investment (§4.1.4). Effectively, this optimal investment rule has some similarities to the avoided cost methodology criticised by Lesser and Feinstein as potentially being sub-optimal. Lesser and Feinstein’s criticism is mainly directed at the use of optimal investment rule over too short a time frame. The successive use of the rule on an annual basis will result in a very different series of investments than examining the entire set of costs and benefits over the full lifetime of an investment or, when investment decisions are interdependent, over the lifetimes of all investments. Besides, the optimal investment rule can be viewed as simply a statement of zero economic profit over the long run, and there may be many investment options for which this rule holds. Furthermore, just because an investment viewed from the perspective of the firm at the *current* moment in time is cost-effective (i.e., it ensures zero, or even positive, economic profits), does not mean that it is necessarily part of the optimal investment plan had that investment been evaluated at some earlier point in time, considering any trade-offs with other investment possibilities. In other words, the optimal investment rule does not say anything about the optimality of past investment decisions. Nevertheless, where decisions are intertemporally interdependent—because of asset non-fungibility (§3.5.1)—sub-optimal investments made in the past cannot be reversed, hence the optimal investment rule is the only practical rule that can be applied to make the best of the existing situation.¹⁰

⁹ As Berg and Tschirhart (1995) observe, “avoided cost is another disguised version of marginal cost”. By contrast, Teplitz-Sembitzky (1992, p. 49) equates the concept with incremental cost, rather like Heald (§5.3.3). In discussing methods for estimating the costs of providing a social obligation, the New Zealand Ministry of Commerce and The Treasury (1995, para. 224, and Appendix IX) talk about “avoidable incremental cost”, which is defined as “the difference between the forward-looking cost of providing the obligation in the least cost manner and the cost the firm would willingly incur in order to provide the service in the absence of the obligation” (§4.4.1). Furthermore, it is noted that “avoidable cost ... may be seen as a further approach to estimating incremental cost”. In defining “average incremental cost”, Baumol and Sidak’s (1994) definition is directly referenced (§5.2.3).

¹⁰ Turvey (1969) himself drew attention to this point: “the cost structure of the industry in any year depends upon the past evolution of its gross investment, its technology and relative factor prices. This brings out the irrelevance of the traditional long-run average-cost curve for a whole industry. Such a curve shows what costs would be at various alternative levels of output if the industry were built from scratch using to-day’s technology and minimising costs at to-day’s relative factor

Notwithstanding this possibility of global sub-optimality, Lesser and Feinstein’s critique of the avoided cost methodology serves firstly as a potent reminder that, for the optimal investment rule to be applied correctly—by an existing firm—it should cast a wide net over *all* the costs and benefits that are attributable to the proposed investment over the entire relevant time frame. This is similar to Sidak and Spulber’s admonition not to fall into a fallacy of forward-looking costs (§5.3.2). Secondly, what may be optimal *now*, for an *existing* firm subject to the intertemporal constraints on its decisions caused by its past investments, may not be optimal now for a *new* entrant desiring to serve all or part of the market on a *greenfields* basis (i.e., as if the market were currently unserved by any capacity).

6.2 Optimal Construction Configurations in a Power Distribution Network

6.2.1 Optimal Construction Configurations: the Sequencing of Demands Example Revisited

The importance of the time dimension, and the difference between greenfields optimality and the optimal investment rule, can be seen by returning to Baumol and Sidak’s example discussed earlier (§5.2.4-§5.3.1) in which the initial demand for two services occurs in sequence. Baumol and Sidak actually present two distinctly different scenarios in their example, although they do not distinguish between them. (Possibly this is because they consider that the outcome is the same in both cases). The first scenario—their “irrelevant piece of history” scenario—is that the demand for services *X* and *Y* commenced at different times, but this occurred for *both in the past* (i.e., one in 1980 and one in 1987). The second scenario is that the demand for one service began in the past, but demand for the second service is yet to commence. Baumol and Sidak allude to this latter state of events by stating: “once the firm has decided to continue either one of the services, provision of the other adds zero to total costs”. In fact, the incremental costs and the optimal asset configurations associated with the two scenarios are very different.

For both scenarios, it is assumed that the services *X* and *Y* are the only services to ever have been, or that ever will be, provided in this particular perfectly contestable market. (Perfect information is therefore automatically assumed). This contestable market is deemed to be a simple network where the two services are distinguishable by location (e.g., §5.2.2). Therefore, the incumbent monopolist is able to price discriminate between services (but not between consumers). Further, it is assumed that there are declining average incremental costs due to economies of scope from a shared capacity input. Therefore, the set of subsidy-free product revenue equations ensures anonymous equity, as was shown by Faulhaber and Levinson (§4.3.3).¹¹ In addition, the total demand for each service remains constant, and extends

prices. This is clearly irrelevant in most cases. Where an industry already exists, therefore, the traditional analysis either applies only to the costs of new marginal capacity or must treat all plant inherited from the past as a fixed factor”.

¹¹ Note that these are the set of subsidy-free *revenues*, and *not* the set of subsidy-free *prices*. As is discussed later (§9.1), to determine the set of subsidy-free prices requires viewing each service as a distinct product at every instant of time. Hence, both *X* and *Y* actually describe an infinite set of products. Although this may sound intractable, given that this results in an

perpetually into the future. No specific assumptions regarding consumers are made however—they are kept “anonymous”. Consumers may consume a bundle of one or both services, and they may enter and leave the market over time at any time. Hence, it is not assumed that the same set of consumers is associated with any part of the demand for a particular service at any particular time. This ensures *intertemporal* anonymous equity. Finally, although this can be relaxed, it is assumed that the assets involved are perfectly non-fungible (§3.5.1).

Under the first scenario, both services are *already* being provided. From the perspective of current and future consumers—as well as a potential entrant—efficiently serving the entire market demand for both products requires optimally matching capacity to *existing* demand only, since the existing demand for both services continues into the indefinite future (even though the consumers associated with that demand may change). Baumol and Sidak are therefore correct. The fact that such demands began in the past is irrelevant, and the constraints on the subsidy-free revenue set—is basically the same as the *static* two good problem provided above where q_1 relates to the demand for X , and q_2 to Y (§5.2.2). Nevertheless, in applying the optimal investment rule, current consumers might perceive the optimal asset design to be different from the already constructed design, even where there have been no changes in technology, or changes in capacity costs (such as from inflation). This is because the incumbent monopolist made its investment decision on the basis of sequenced demand, whereas the current (and future) consumers would not.

Given that the demand for X began in 1980, and the incumbent firm knew then that the demand for Y would begin in 1987, the incumbent had *three* possible intertemporal investment choices. By contrast, there is only a single hypothetical greenfields asset configuration that is least cost; optimally matching capacity to the total market demand, here termed ‘*static construction*’. The incumbent’s first option—a construction program of ‘*anticipatory construction*’—would have been to build sufficient capacity to provide both services in 1980, in anticipation of the total market demand in 1987. The second—‘*capacity expansion*’—would have been to construct only sufficient capacity for service X in 1980, and then construct additional “incremental” capacity in 1987 for service Y . In this case, no capacity can be considered “joint”, and no economies of scope would be realised. Finally, the incumbent could have engaged in a program of ‘*capacity replacement*’. In this case, like the capacity expansion configuration, the incumbent would have initially built only sufficient capacity for service X in 1980. However, rather than building just the incremental capacity required to supply service Y in 1987, under this construction configuration the incumbent *scraps* the original capacity, and constructs an entirely new network in 1987 sufficient to provide *both* services.

infinite set of constraints, it is in fact no more problematic than solving the anonymously equitable solution to the conventional peak load pricing problem, which similarly involves infinite constraint sets (§4.3.3).

Of course, if there are more than two services, there will be many more than three potential least cost construction configurations, since the number of possible construction configurations—rather like the static SAC and IC constraints (§5.1.3 and §5.2.4)—is combinatorially related to the number of products. But the general programs of investment configurations will be similar, involving (a) trade-offs between building some capacity in anticipation of future demand, (b) building incremental capacity to meet current demand growth, and (c) scrapping some older capacity to make way for new capacity which can take better advantage of the prevailing economies of scale and/or scope.

Which of these three construction configurations is least cost depends on: the cost of capacity required to provide X and Y separately; the economies of scope embodied in joint capacity that can supply both services; the useful lifetime of the durable assets providing the required capacity; and the intertemporal opportunity cost of capital (i.e., the ‘*discount rate*’). In general, for any case involving only two products, the optimal construction configuration also depends on the time interval between the initial provision of the first-supplied service and the initial provision of the second-supplied service. In this example, it is seven years. Also if the assumption of perfect non-fungibility is relaxed, this will also alter the relative attractiveness of the three options. For instance, if the assets are perfectly fungible, then this is the same as suggesting that capacity is perfectly adjustable from the incumbent’s perspective (and not just a hypothetical entrant’s). This means that capacity replacement would always be optimal, since there would be no loss involved in scrapping the original asset before the end of its useful lifetime. On the other hand, if the declining average incremental cost assumption were changed to one of *constant* returns to scale and scope, then capacity expansion would always be optimal, since there would be no cost advantage either in constructing capacity in anticipation of later demand, or in scrapping capacity early to take advantage of non-existent economies.

If either anticipatory construction or capacity replacement had been the incumbent’s least cost option, then at the present point in time, a *single joint asset* would exist supplying both services, although the remaining lifetime of those assets would be different under either configuration. However, if capacity expansion had been the least cost construction program back in 1980, then there will now be *two dedicated assets*, each associated with a particular service. Yet, the least cost hypothetical greenfields network will always be the static construction configuration involving the joint capacity. Hence, while the optimal investment rule applied in 1980 may require there to be two assets supplying the entire market at the present day, this is sub-optimal compared to the application of the optimal investment rule today based on constructing a greenfields network. This global sub-optimality arises from the trade-off between the *intertemporal opportunity cost of capital*, and the *static economies of scope*. However, because of the perfect non-fungibility of capital it would be even less efficient to scrap the existing network which historically evolved under a construction program of capacity expansion. Dynamic efficiency requires that the incumbent make any future investment decisions subject to the constraints imposed on it by its past (optimal) decisions, while allocative efficiency (and equity) requires that the

incumbent firm price its assets on the basis of recovering the costs of a currently optimal greenfields network.

Under Baumol and Sidak's second scenario, only one service is already being provided, and demand for the other is still to ensue at some known later date. The existing construction configuration will be the result of the incumbent monopolist having already assessed which of the three construction configurations was least cost at the time of the initial capital outlay. However, under this second scenario, rather than having no other choice but to engage in a program of static construction, a hypothetical entrant intending to supply current and future consumers would have three possible construction configurations for the optimal greenfields network it could use to force out the incumbent. An entrant would need to evaluate which of these three options is least cost.

Interestingly, whereas the first scenario can result in a globally *sub*-optimal existing network design, the second scenario can result in a *supra*-optimal asset configuration. Consider if both the incumbent's initially least cost network design, and the optimal greenfields network, were *both* capacity replacement. If the optimal greenfields configuration is considered to be a strategy taken by a hypothetical entrant, it is clear that this entry strategy could never force out the incumbent firm. Under capacity replacement, both the incumbent and a later entrant would need to scrap capacity at the time when demand for the second service commences. The incumbent has a significant advantage over the later entrant however, since the incumbent has a longer time period over which to recover its capital outlay on the original non-fungible asset. This might suggest that the incumbent is free to exploit its monopoly power. However, in deriving the set of subsidy-free revenues, there is still one overriding constraint, the total revenue/cost equality (i.e., the zero profit constraint).

The set of intertemporal subsidy-free constraints on revenues associated with Baumol and Sidak's second scenario is derived in Chapter VIII, although the solution is derived by using the slightly earlier model which BPW used to investigate intertemporal unsustainability (§3.3.3). Use of this model allows both issues, intertemporal subsidy-free pricing and unsustainability, to be examined. Unlike the first scenario, which has already been discussed above (§5.2.2), the second scenario is considerably more complicated. In Chapter IX the result is extended to intertemporal subsidy-free *prices*.

6.2.2 Optimal Construction Configurations using New Zealand Distribution Cost Data

The relative attractiveness of the various construction configurations—as well as the difference between the optimal investment rule and global optimality—can be demonstrated by using the real distribution cost data from New Zealand presented earlier (§3.6.2). In this earlier discussion, it was noted that average cost curves of the static, or greenfields, zone substation design—shown in Figure 3.4—take no account of the optimal expansion path.

Although grossly simplifying the problem of optimal distribution network investment, a simple example is provided here. The zone substation cost function and zone substation design constraints discussed earlier are applied (based on data underlying the discussion in Mercury Energy Lines Business, 1999, and Vector, 1999). However, it should be pointed out that these design parameters and costs were specific to one electricity line business (ELB) at a particular point in time, and hence no generalisation can be made on the basis of this example, which solely is provided to show how optimal asset configurations can differ over time in a distribution network. Moreover, in all cases, only capital costs are considered, whereas in a real network, trade-offs involving the cost of operating costs (e.g., maintenance and losses) should also be taken into account (§3.1.3).

To simplify the analysis, time is split into two distinct periods, and initially it is assumed that a single zone substation is constructed to meet current and future demand for a particular geographic area. Because of the indivisibilities in transformer sizes, this does not necessarily require that demand for capacity within either of the two periods is assumed constant, simply that demand is restricted within a certain range over that period, particularly with respect to an upper limit. Given the possible impact of different contingency criteria on design, there is something of a disconnect between the demand for capacity and the optimal expansion path over time, as will be demonstrated shortly.

Initially, it is assumed that a new zone substation is constructed to an $n-1$ contingency standard (§3.1.3 and §3.6.2), and 10MVA transfer capacity is available from neighbouring substations. Initial demand, and demand for the entire first period, is considered to range between 15MVA and 22.5MVA.¹² If demand were to remain indefinitely within this range, then—from the first two columns of Table 3.1—the least cost transformer configuration would be 2x15MVA. However, if demand for capacity were expected to *exceed* 22.5MVA within the lifetime of the substation, then the ELB has a number of options, depending on when and by how much demand will exceed the firm capacity of the original design. For instance, if initial demand was 15MVA, and demand for capacity was expected to incrementally grow at 2% per annum, then after 21 years the firm capacity of the substation would be reached. Although this may seem a long time into the future, the lifetime of most assets associated with a zone substation, and particularly the transformers, have asset lives ranging from 45-60 years (Energy Markets Regulation Unit, 2000c, sB.41). Hence, if the original configuration was 2x15MVA, then load has to be

¹² Such initially high demand could be due to the substation being constructed to initially meet some large greenfields load such as a major industry or the stage I development of a large commercial or residential subdivision (§3.1.2). Alternatively, it could also arise from system augmentation. For example, where infill housing is causing two existing substations to near their firm capacities, a new substation could be constructed “between” them. Load from the two existing substations would then be rebalanced between all three once the new substation is constructed. Consequently, the initial demand could be high. For example, if the two neighbouring substations had demands close to firm capacities of 30MVA, then rebalancing could cause an initial demand at the new substation of around 20MVA.

(i) permanently off-loaded to an existing neighbouring substation with some spare capacity, although this may reduce transfer capacity and thus overall firm capacity; (ii) served by constructing a new zone substation; or (iii) served by expanding the capacity of the existing substation.

However, if the ELB expects the demand to increase, then the ELB has three choices when making its decision concerning the original substation design. Assuming that demand is not projected to eventually exceed 30MVA, the first option is to construct a 2x15MVA substation, and add a third 15MVA transformer to the substation once demand exceeds 22.5MVA. This option can be labelled “capacity expansion” (§6.2.1). However, this option clearly involves excess capacity in the second period (i.e., after expansion). From Table 3.1, the firm capacity of the expanded substation would now be 40MVA, even though future demand is never expected to exceed 30MVA, leaving 10MVA permanently “idle”. The second option is therefore to resize the substation capacity to more optimally match future demand. Hence, once demand exceeds 22.5MVA, the ELB could remove both existing 15MVA transformers and replace them with two 20MVA transformers. If there were a national resale market for second-hand transformers, or the ELB could utilise them at some other location within its own network, then the old 15MVA transformers might not need to be entirely scrapped (§3.5.5). This option can be labelled “capacity replacement”. The third option is to construct capacity in anticipation of future demand growth at the outset—a programme of “anticipatory construction”. This would require initially constructing the zone substation with 2x20MVA transformers.

An additional assumption—discussed in more detail later (§7.2.3)—is that demand continues perpetually into the future. This means that at the end of any asset’s lifetime, it must be replaced by an asset of identical characteristics. For an asset costing K and with a lifetime of N years, it can be readily shown that the present value cost of replacing that asset indefinitely into the future is $K/(dB_N)$, where d is the discount rate, and B_N is the ‘uniform series present worth factor’ (e.g., Bertram *et al.*, 1992, p. 145)—the general equation of which, for an arbitrary period of X years, is as shown in (6.1).

$$B_X \equiv \frac{1 - \left(\frac{1}{1+d}\right)^X}{d} \quad (6.1)$$

6.2.3 Total Cost Equations for a Single Zone Substation

The total cost equations for each of the three construction configuration options are shown in equations (6.2)-(6.5), with the simplifying assumption that all substation assets have the same asset lifetime. The basis for the total cost equations is the zone substation cost function presented in (3.4). The costs of land are ignored, since unlike all other costs, the value of capital “sunk” in land is of quite a different nature from all other costs involved in providing distribution network services. One-off fixed costs (such as administration) are combined as F_O , and other fixed costs subject to a finite lifetime (such as the cost of the substation building) are combined as F_N . Transformer-specific fixed costs (such as

Transpower modifications, the HV network, transformer bays, earthing and metering) are all combined in a single cost F_T . The simplest total cost (TC) equation is for anticipatory construction (6.2), which involves renewing the original substation design every N years. The term X_{AC1} refers to the individual transformer capacity under anticipatory construction installed in the first period, which in this specific example would be 20MVA. The term n_{AC1} refers to the number of transformers initially installed, which in this specific case is two.

$$TC_{AC} = F_O + \frac{1}{dB_N} (F_N + n_{AC1} [F_T + (a + bX_{AC1})]) (1 + f_p) \quad (6.2)$$

Under capacity expansion, the initial capital outlay is similar to that of anticipatory construction, although the initial period individual transformer capacity will be different (X_{CE1}), 15MVA. However, once demand exceeds 22.5MVA at the end of T years, an additional transformer will need to be installed. This will incur not just the purchase and installation costs of the new transformer, but the costs associated with any additional transformer, such as the costs of the transformer bay and earthing. (It is assumed that the original building is of sufficient size to house the additional transformer).¹³ The cost of expansion is discounted back to the present through the use of a discount factor ρ (e.g., BPW, p. 409) for any period of X years, as defined in (6.3). This is then added to the initial period cost to give the total present value cost of a capacity expansion configuration (6.4).

$$\rho_X \equiv (1 + d)^X - 1 \Rightarrow X = \log_{(1+d)}(1 + \rho_X) \quad (6.3)$$

$$TC_{CE} = F_O + \frac{1}{dB_N} (F_N + n_{CE1} [F_T + (a + bX_{CE1})]) (1 + f_p) \\ + \frac{1}{dB_N} \left(\frac{(n_{CE2} - n_{CE1}) [F_T + (a + bX_{CE2})]}{1 + \rho_T} \right) (1 + f_p) \quad (6.4)$$

Under capacity replacement, the initial capital outlay again appears similar, however, the costs of the original transformers do not need to be replaced indefinitely. Only the fixed costs which are not one-off expenditures need to be replaced in perpetuity. Moreover, this time the second period costs are more complex. Costs specific to the number of transformers only increase if the number of transformer increases. Also, there may be some *reduction* of costs due to the receipts from reselling the original transformers (S). On the other hand, additional costs are incurred in replacing transformers. Total costs are given in (6.5).

¹³ Zone substations are not always enclosed structures, but may also be outdoors.

$$\begin{aligned}
TC_{CR} = & F_O + \left(\frac{F_N}{dB_N} + n_{CR1} \left[\frac{F_T}{dB_N} + (a + bX_{CR1}) \right] \right) (1 + f_p) - \frac{S}{1 + \rho_T} \\
& + \frac{1}{dB_N} \left(\frac{\text{Max}(n_{CR2} - n_{CR1}, 0) F_T + n_{CR2} (a + bX_{CR2})}{1 + \rho_T} \right) (1 + f_p)
\end{aligned} \tag{6.5}$$

6.2.4 Optimal Construction Configurations for a Single Zone Substation

The optimal construction configuration is thus the one with the least total cost from (6.2), (6.4) and (6.5). For the specific case introduced in the previous subsection, the crucial variable is the *time* when demand for capacity first exceeds 22.5MVA. The transition year T where the optimal construction configuration changes can be found by equating the RHS of (6.2) with the RHS of (6.4), and then the RHS of (6.4) with the RHS of (6.5). Using the indicative cost data from before (§3.6.2)—where the total per transformer costs are \$107,000; the transformer cost function is \$168,333 + \$16,333 X_T ; N is 50 years; and d is 8%—the transition year from anticipatory construction optimality to capacity expansion optimality occurs at around 15 years. Similarly, the transition year from capacity expansion being optimal to capacity replacement being optimal is close to 43 years (assuming that there is no resale market for the original transformers, and thus $S=0$). This means that if demand exceeds 22.5MVA within 15 years, then anticipatory construction is optimal, and a 2x20MVA substation should be constructed from the outset. If demand does not exceed 22.5MVA until after 15 years, but not as late as 43 years, then a program of capacity expansion is optimal. Finally, if demand will not exceed 22.5MVA until after 43 years, then capacity replacement will be the optimal construction configuration.

The fact that anticipatory construction might be optimal does not imply rapid demand growth. The indivisibilities in distribution assets mean that small or large changes in demand growth could equally require a program of anticipatory construction. For example, if the zone substation were sized to meet a single industrial load of 21MVA known to expand its process by only 2MVA in 10 years time (i.e., to 23MVA) then anticipatory construction would be optimal. Alternatively, if initial demand was 15MVA, and demand were to grow by 5% per annum for the next 10 years and then level off, this would also require a program of anticipatory construction. Hence, the optimal construction configuration can to some extent be treated independently from knowledge of the actual path of demand or demand growth. If the transfer capacity were different, say 5MVA, then from Table 3.1 it can be seen that a 2x15MVA design has a firm capacity of 20MVA rather than 22.5MVA, and a 2x20MVA design has a firm capacity of 25MVA instead 30MVA. Clearly this would alter which design is optimal given a particular set of current and future demands. However, the transition year at which it becomes optimal to initially construct a 2x15MVA substation rather than a 2x20MVA substation does not change.

Table 6.1 presents a broader range of optimal construction configurations for different combinations of first and second period demands, given the same cost assumptions as above. In a number of cases it is clear that *only* capacity expansion can be optimal. On the other hand, in one case—

where demand increases from 30-40MVA to 40-50MVA—only anticipatory construction or capacity replacement can be least cost. Again, changing the overload factor or transfer capacity will not change the transition points between optimal construction configurations, only the demand ranges to which those configurations apply. For instance, if the transfer capacity is only 5MVA rather than 10MVA, only the first and second period demand ranges would change; for example, the 15-22.5MVA demand range would narrow to 15-20MVA, and the 22.5-30MVA range would fall to 20MVA-25MVA. However, the transition point, where an anticipatory construction configuration of 2x20MVA would cease to provide a lower cost supply than a program of capacity expansion of 2x15MVA to 3x15MVA, would still be 15 years.

First Period Demand	Second Period Demand	Anticipatory Construction Configuration	2nd Period Transition Year	Capacity Expansion Configuration	2nd Period Transition Year	Capacity Replacement Configuration
15-22.5MVA	22.5-30MVA	2x20MVA	15 years	2x15MVA → 3x15MVA	43 years (25 years) ¹⁴	2x15MVA → 2x20MVA
	22.5-40MVA	<i>na</i>	-	2x15MVA → 3x15MVA	-	<i>na</i>
	22.5-50MVA	<i>na</i>	-	2x20MVA → 3x20MVA	22 years (17 years)	2x15MVA → 3x20MVA
22.5-30MVA	30-40MVA	3x15MVA	7 years	2x20MVA → 3x20MVA	46 years (38 years)	2x20MVA → 3x15MVA
	30-50MVA	<i>na</i>	-	2x20MVA → 3x20MVA	-	<i>na</i>
30-40MVA	40-50MVA	3x20MVA	22 years (19 years)	<i>na</i>	-	3x15MVA → 3x20MVA
40-50MVA	40-50MVA	3x20MVA	-	<i>na</i>	-	<i>na</i>

Table 6.1: Optimal Single Zone Substation Construction Configurations

These results can now be contrasted with the earlier discussion regarding the optimal investment rule and global optimality (§6.1.4 and §6.2.1). Consider a case where the incumbent ELB making the initial greenfields substation design correctly identifies that a program of capacity expansion is optimal. This would occur say if demand in the first period were constant at 20MVA, and increases to 30MVA suddenly after 20 years. Accordingly, the incumbent ELB would initially construct a substation of 2x15MVA, and subsequently expand it to 3x15MVA. However, a *hypothetical entrant* evaluating the market at the time that demand increases would recognise that its own optimal greenfields design—a program of static construction sized to 30MVA—should be 2x20MVA. The average cost of this design is clearly lower than the average cost of the incumbent’s (as can be inferred from Figure 3.4). Figure 6.1 shows that where there is demand growth, average costs rise substantially under a program of capacity

¹⁴ Assuming a perfect resale market for second-hand transformers, and that transformers depreciate in value according to economic depreciation (§7.2)—hence transformer resale value (R) satisfies (7.8)—results in the transition points in brackets throughout Table 6.1.

expansion or replacement. Simply because average costs rise neither implies that the incumbent has not made the optimal investment decision, nor that the incumbent is not a natural monopoly.

The optimality conditions clearly change as cost conditions change and as other constraints are imposed. For instance, changing the transformer-specific fixed costs changes the transition points. Where initial period demand ranges from 15-22.5MVA, and second period demand ranges from 22.5-30MVA, the transition points in Table 6.1 from anticipatory construction to capacity expansion, and from capacity expansion to capacity replacement are 15 years and 43 years respectively. This result is applicable to transformer-specific costs of \$107,000 per transformer. Figure 6.2 presents the transition points as a function of transformer-specific costs, for d equal to 8% (and d equal to 10%, for comparative purposes; §9.4.5). It can be seen that as transformer-specific costs increase, the likelihood of capacity expansion being the optimal configuration declines. In fact, where these costs exceed \$480,000 per transformer, capacity expansion can no longer be a feasible construction configuration.

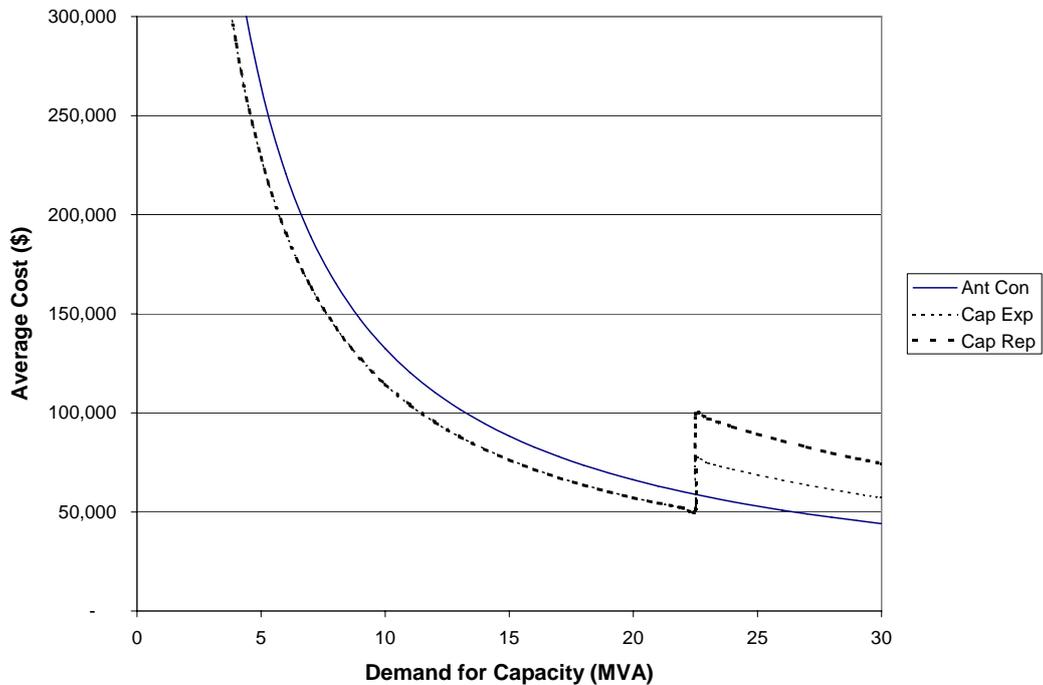


Figure 6.1: Single Zone Substation Average Costs under Different Construction Configurations

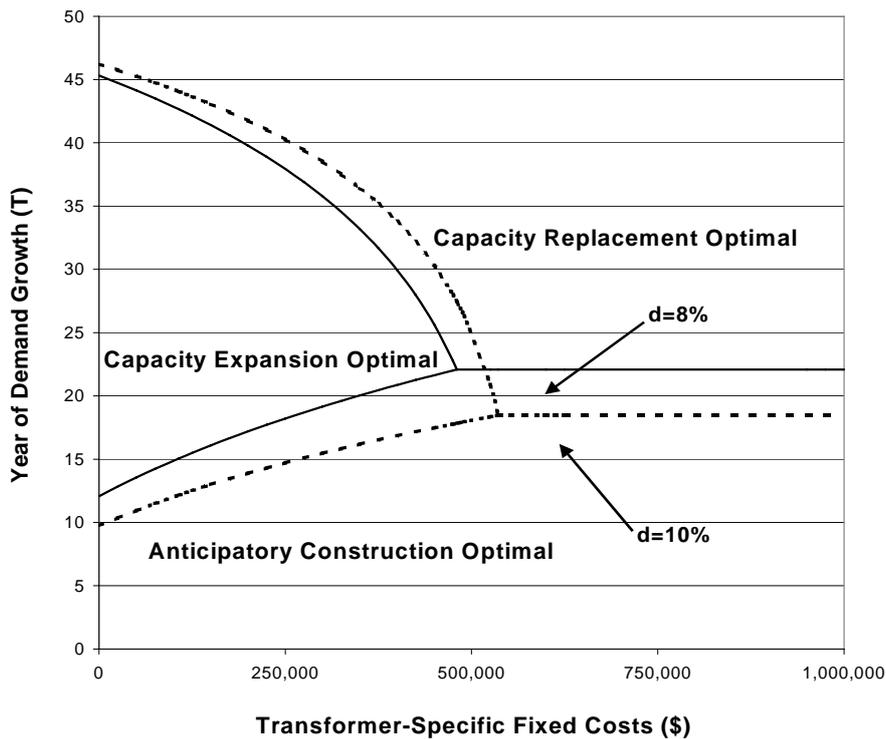


Figure 6.2: Single Zone Substation Optimal Construction Configurations

6.2.5 Optimal Construction Configurations for One or Two Zone Substations

Consider a different scenario where there are reasons why an already-constructed zone substation cannot subsequently be altered or expanded, and would instead have to be relocated or rebuilt. For instance, this would occur if the geographical area to be served had a total future demand higher than 50MVA; the limit on zone substation size given in the earlier assumptions (§3.6.2). Capacity expansion would thus require constructing a *second* zone substation rather than expanding an existing one. But this could also occur if local government planning restrictions change over time. In such case, capacity replacement might require decommissioning and demolishing an existing substation, and reconstructing a new one, with a larger firm capacity, at a different, or the same site. Although the total cost equation for anticipatory construction would remain the same as equation (6.2) under these circumstances, this would change the corresponding total cost equations for capacity expansion and capacity replacement, to those in (6.6) and (6.7) respectively. The corresponding optimal construction configurations are presented in

Table 6.2 (assuming in this case that F_N is non-zero, and equals \$250,000, the cost of a substation structure).¹⁵

$$\begin{aligned} \text{TC}_{\text{CE}} = & F_O + \frac{1}{dB_N} (F_N + n_{\text{CE1}} [F_T + (a + bX_{\text{CE1}})]) (1 + f_p) \\ & + \frac{F_O}{1 + \rho_T} + \frac{1}{dB_N} \left(\frac{F_N + n_{\text{CE2}} [F_T + (a + bX_{\text{CE2}})]}{1 + \rho_T} \right) (1 + f_p) \end{aligned} \quad (6.6)$$

$$\begin{aligned} \text{TC}_{\text{CR}} = & F_O + \left(\frac{F_N}{dB_N} + n_{\text{CR1}} \left[\frac{F_T}{dB_N} + (a + bX_{\text{CR1}}) \right] \right) (1 + f_p) - \frac{S}{1 + \rho_T} \\ & + \frac{F_O}{1 + \rho_T} + \frac{1}{dB_N} \left(\frac{F_N + n_{\text{CR2}} [F_T + (a + bX_{\text{CR2}})]}{1 + \rho_T} \right) (1 + f_p) \end{aligned} \quad (6.7)$$

Of course, optimising zone substation configurations in reality will often be much more complex than these indicative examples suggest. The process might require considering many zone substations over a large geographical area, and a substantial number of possible temporary and permanent load transfers between them over time (§3.1.3). The examples here are solely intended to reinforce the concept of optimal spare capacity, and that what might be optimal for an entrant, is not necessarily optimal for an incumbent firm.

First Period Demand	Second Period Demand	Anticipatory Construction Configuration (1 Substation)	2nd Period Transition Year	Capacity Expansion Configuration (2 Substations)	2nd Period Transition Year	Capacity Replacement Configuration (1 Substation)
15-22.5MVA	22.5-30MVA	2x20MVA	24 years	2x15MVA + 2x5MVA	38 years	2x15MVA → 2x20MVA
	22.5-37.5MVA	3x15MVA	11 years	2x15MVA + 2x10MVA	42 years	2x15MVA → 3x15MVA
	22.5-40MVA	3x15MVA	12 years	2x15MVA + 2x15MVA	39 years	2x15MVA → 3x15MVA
	22.5-45MVA	3x20MVA	7 years	2x15MVA + 2x15MVA	44 years	2x15MVA → 3x20MVA
	22.5-50MVA	3x20MVA	9 years	2x15MVA + 2x20MVA	41 years	2x15MVA → 3x20MVA
22.5-30MVA	30-37.5MVA	3x15MVA	11 years	2x15MVA + 2x10MVA	-	na
	30-40MVA	3x15MVA	12 years	2x15MVA + 2x15MVA	-	na
30-40MVA	40-45MVA	3x20MVA	7 years	2x15MVA + 2x15MVA	-	na
	40-50MVA	3x20MVA	9 years	2x15MVA + 2x20MVA	-	na
40-50MVA	40-50MVA	3x20MVA	-	na	-	na

Table 6.2: Optimal Construction Configurations where Initial Substation Cannot be Modified

¹⁵ In the single zone substation cases, F_N (and F_O) do not impact the transition points, because it (they) cancel out when equating the total costs of the single construction configuration with any other. However, in equations (6.6) and (6.7), F_N no longer cancels out (although F_O still does).

6.3 The Appropriate Costs to Include in Stand Alone Costs and Incremental Costs

6.3.1 Opportunity Costs and Replacement Costs in Subsidy-Free Bounds

Baumol and Sidak (1994a, p. 59) state that there are three sets of costs which must be included in any calculation of marginal cost, average incremental cost, or stand alone cost: (i) *the cost of capital*, as already mentioned above (§6.1.3); (ii) the *opportunity cost* that the firm forgoes by undertaking the action in question (§5.3.2); and (iii) the *replacement costs* of existing assets, also evaluated at their opportunity cost. This prescription is almost identical to Schramm's (1991) definition of *long run incremental cost* (§4.1.5), which is not surprising, since Baumol and Sidak suggested that incremental costs could be calculated in their own right, based on the concept of long run incremental cost already familiar to (US) regulators (§5.3.3).

In equilibrium in a perfectly competitive or contestable market, the firm's revenues will always cover opportunity cost. Moreover, this condition is required for economic efficiency. ... [W]here the firm uses assets purchased in the past, the cost of using them must be evaluated as the opportunity cost of that use—that is, the price that those assets could fetch if transferred to an alternative use. In a business where demand is sufficient to make it profitable to replace those assets at the appropriate time, the assets will be valued in a competitive market at the current cost of the most economical replacement of the remaining output capacity of those old assets. No buyer will pay any more than this replacement cost for those assets, and their current owner will be unwilling to sell such profitable assets for any less (Baumol and Sidak, 1994a, pp. 59-60).

The cost of capital is itself derived based on opportunity cost, and Baumol and Sidak remind regulators that the concept of opportunity cost requires that firms receive at least the return *on* capital which they could have received from employing that capital in its best alternative use. Evans *et al.* (2000) address this issue in the context of the dynamic efficiency of New Zealand's power distribution sector reforms.

The importance of the pricing and management of capital in creating dynamic efficiency is also illustrated by the electricity distribution industry. The ... [pre-reform] distribution entities had trust ownership or were departments of local government. Their prices were set to the breakeven revenue requirement on the basis of anticipated maintenance and capital works. Any surplus was used for cost stabilisation over time, including capital financing. In pricing no allowance was made for the opportunity cost of capital in the business. The approach essentially bundled the required equity return into the prices that were charged consumers. To the extent that the consumers served matched the "owners" of the distribution companies, this approach went a considerable way to improving allocation over that of a local investor-owned monopoly. But it meant that capital was not fully costed into prices. Since the *Electricity Act 1992* energy companies are required to be commercial enterprises and operate as successful businesses, with an implied imperative to make profits and a requirement to pay tax (unlike their pre-reform equivalents). In the transition, prices might reasonably be expected to rise to reflect incorporation

of the opportunity cost of capital. While pricing capital fully in all its uses should improve dynamic efficiency through its effect on the choice of investments, in the case of electricity distribution companies there is a trade-off between this effect and that of any local monopoly power in aspects of the distribution business (Evans *et al.*, 2000).

Opportunity costs are by definition “forward-looking”, hence Sidak and Spulber (1997, pp. 286 and 319) make it clear that to ignore opportunity costs would be a fallacy of forward-looking costs (§5.3.2). In a competitive (or contestable) market, a firm would not accept a price (*ex ante*) unless it included compensation for such opportunity costs. Sidak and Spulber state that when market alternatives are present, then it is straightforward to determine the opportunity cost, since the opportunity cost is simply the market price of the alternative use.

The common thread is the interplay between economic expectations and opportunity costs. Economic agents make choices by comparing the expected net benefits of available alternatives. All economic choices are ‘forward-looking’, with the referred choice yielding the greatest expected benefits. The opportunity cost of the preferred choice is the value of the best alternative. By definition, the expected benefits of the preferred choice exceed its opportunity cost (Sidak and Spulber, 1997, p. 393).

Economides (1997) observes that what is cost effective from the firm’s perspective, and what is optimal to society, can diverge where *private* opportunity costs—which the firm takes into account in making its investment decision—diverge from *social* opportunity costs. Sidak and Spulber (1997, p. 373) reject this viewpoint, maintaining that “the opportunity cost of an action should be determined from the point of view of the decision maker”. Moreover, “the market price of a good remains its best indicator of opportunity cost”.

Economic analysis defines the value of a product or service as its market price. If value were based on any other yardstick than market price, economic value would be cast adrift without the foundation of market exchange, which reflects what consumers are willing to pay to obtain the good and what producers are willing to accept in return for supplying the good. The profitability of an economic activity does not alter the fundamental identity between economic value and market price. Market prices cannot be said to reflect value only when there is zero economic profit. Moreover, even if it were desirable to exclude monopoly profit from valuation, how could it be determined whether or not a firm were earning a monopoly profit? The returns earned by a firm include not only the returns to invested capital, but also the returns to innovation, entrepreneurship, arbitrage, brand equity, goodwill, creativity, managerial effort, and other intangible factors. The economic value of those inputs is best determined with reference to market prices. Not only is it impractical to exclude monopoly profit from valuation, it is meaningless to do so. A firm evaluates a forgone opportunity in the same manner whether the firm is a monopolist or whether it faces many competitors. ... Therefore a distinction between

opportunity costs for a monopolist versus a competitive firm has no economic content (Sidak and Spulber, 1997, pp. 373-374).

Nevertheless, simply because a firm makes a cost-effective investment does not mean that the use of the resources in question is its *best* use to society. The opportunity cost to the investing firm is the value of the best alternative available to *that* firm. Where a different firm would be able to make *better* use of those resources, the opportunity cost to society is higher than the private opportunity cost. In a perfectly contestable market the resources in question would end up—via market exchange—with the firm that could maximise the value in use of those resources, and thus marginal social opportunity costs and marginal private opportunity costs would be equivalent. Therefore, the market price for those resources as held by the original sub-optimal firm would not necessarily be sold at their marginal social opportunity cost. Rather the original firm would be content if it recovered its marginal *private* opportunity costs. In a regulated industry, there may be barriers to free market exchange, and resources may remain with inefficient firms.¹⁶ Consequently, a regulated firm employing resources sub-optimally should not be entitled to set its prices on the basis of recovering the *social* opportunity cost.

Turning now to *replacement* costs, in describing how these should be included in SAC and AIC, Baumol and Sidak once again—like their allowance of supranormal profits (§6.1.3)—step outside the realm of strict contestability theory. The difference is subtle, and the impact will often be minor, but the implications are important. While Baumol and Sidak do not sanction the recovery of *historic* costs, they do appear to suggest that the incumbent firm should be allowed to recover the *replacement* costs of historic investments at that *future* time which those assets reach the end of their useful lifetime.

Although their prescription is somewhat ambiguous, Baumol and Sidak seem to be allowing for the fact that an incumbent firm will use the most economical solution at the time when existing assets actually need replacement. This is certainly the viewpoint of Schramm (1991), who states that future costs of replacement assets “that have to be built when current ones are worn out and must be replaced” should be “discounted back to the present at an appropriate opportunity discount rate” (§4.1.5). When a firm comes to replace assets, it will rationally replace the existing equipment taking into account advances in technology since the original equipment was installed, as well as changes in consumer preferences and demand. This might seem to be the incentive provided by the contestable market. Yet there is a difference between the current replacement cost of an existing asset, and the value which that asset might have in the optimal, but hypothetical, asset configuration of a *potential entrant* (§4.3.6). While Schramm appears to view price as a “cost-plus” concept from the incumbent’s perspective—where price is derived from the incumbent’s current and future (discounted) costs—Baumol and Sidak present

¹⁶ For example, where ownership structures preclude takeover—such as the “trust” ownership structure for some of New Zealand’s ELBs (§2.4.2)—inefficiently utilised assets may remain vested in inefficient firms.

price as being derived from the costs of a hypothetical entrant. The perfectly contestable benchmark for price ceilings is thus related to the *hypothetical entrant's* current and future costs and not to the hypothetical future replacement costs of the *actual incumbent* firm.

6.3.2 Cost Changes Due to Changes in Technology and Demand Patterns

As Heald (1997) points out, technological advances may mean that a cost structure that was previously subsidy-free (and sustainable) is now rife with cross subsidies; such as is likely to have occurred upon the initial introduction of mobile phones into telecommunications markets. Baumol and Sidak appear not to allow for the fact that any advances in technology could be used *now*, and not at some future date when superseded, stranded, or simply sub-optimal equipment reaches the end of its lifetime. Lesser and Feinstein's (1999) message is that what may have been optimal once, may not be so any longer (§6.1.4). Were the market a greenfields one, then a new firm would consider the optimal application of existing technology, required to serve both known current demand and uncertain future demand *now*, not at some later date.¹⁷

Even more significantly, in a network industry where economies of scope are significant, *changes in technology* over time are not the only key factor which cause past investments to become globally sub-optimal. As the network evolves to meet *changing demand patterns*, the optimal greenfields design of the network is likely to diverge from the actual design. In presenting his original approach to evaluating cross-subsidies (§4.3.2), Faulhaber (1975) made it clear that the static basis of the analysis did not allow for either of these two dynamic factors.

At least two caveats must be added to this discussion of competitive entry, both relating to the limitations of this static analysis when applied to actual pricing decisions. First, it may be the case that the opportunities for alternative supply arrangements may change drastically once a decision has been reached on a particular cooperative arrangement. ... The second problem arises when technological innovation changes the nature of the game. Dynamic changes of this nature may affect not only the level of costs, but also the cost structure of the industry ... Although its difficulty and importance to public policy are clear, the problem is not treated in this paper (Faulhaber, 1975).

Because of the non-fungibility of the durable network equipment, the historic design—comprising both the configuration and size of network assets (§3.1.3)—cannot be changed readily in

¹⁷ Sidak and Spulber (1997, p. 422) take issue with this viewpoint, arguing that there is a lag in introducing new products and technologies into any markets, and that the initially higher price of such eventually lower cost technologies provides the incentive for R&D into new technologies in the first place. Consequently, regulators are not entitled to “jump the gun” and immediately base price ceilings on newly-available technologies. Given that the topic of this thesis is power distribution, which unlike telecommunications is not—at least *yet*—currently subject to rapid technological advances (§6.4.1), this argument is set to one side.

response to such changes.¹⁸ Even when assets reach the end of their lifetime it will be difficult to alter the network layout unless all interconnected assets reach the end of their lifetime at roughly the same time. Although each investment made during the expansion of the network will have been judged on the basis of the optimal investment rule—and as such was an “economic” decision at the time, given the constraints imposed by past decisions—it is unlikely that those investments have resulted in a network configuration that would be optimal if the network were now to be redesigned on a greenfields basis. This is true whether the network evolved in a real world market or in a perfectly contestable one. Perfect contestability does not provide any guarantee that the network configuration which evolved as an outcome of the correct application of the optimal investment rule will be the same as the currently optimal greenfields network.

What impact does this have on dynamic efficiency? Actually, none. The fact that the existing network might be sub-optimal compared to a greenfields one is irrelevant from a dynamic efficiency perspective. The network evolved in a dynamically efficient manner as long as the optimal investment rule was applied correctly, and in a perfectly contestable market—where access to perfect information is assured—there is no reason for the optimal investment rule not to be used properly. However, this difference between the globally sub-optimal actual network configuration and the currently optimal, but *hypothetical* greenfields one, does impact the intertemporal subsidy-free *prices*. This is because prices should be based on the currently optimal network, not the historically optimal one, since that hypothetical network signals the opportunity cost of supply.

6.3.3 The Case for the Inclusion of Historic Costs

Two important arguments could be raised against this view. Firstly, as noted above, firms make investment decisions on the basis of assumptions regarding future revenues; they have “investment-backed expectations” (§5.3.2). Had an incumbent monopolist realised that it might not receive the expected revenues after all, then that would likely have altered its original investment decision, and perhaps that firm would not have “sunk” capital in the first place. For instance, Sidak and Spulber provide an example where new technology is introduced such that the replacement costs of an

¹⁸ The hypothetical optimal greenfields design will (mostly) be subject to the same exogenous constraints as the actual current design. Any local government regulations, such as zoning laws restricting substation sites, constraints on allowable conductor routes, or the requirement to use underground cables rather than overhead lines, all affect the hypothetical configuration. This is the case even where the constraints have changed, and as such seem to benefit the incumbent—for example, the requirement that any *new* conductors be installed underground, while existing lines are not affected. Were consumers to construct their own network, they would have to use underground cables, but the incumbent firm will itself have to use underground cables on current routes, once existing overhead lines reach the end of their useful lifetime. Hence, the optimal configuration, given changed circumstances, may be *more* costly than the existing network, in which case the existing network, at least for a short period after changes in regulations, might be “supra-optimal”, or lower cost than the currently-feasible optimal network (§6.2.1).

incumbent firm's existing assets are much lower than the original cost of those assets. They assert that, simply because new technology is available "would not mean that the contract price should be reduced to forward-looking costs. The purpose of the contract is to protect the expectation interests of the buyer and seller". Consequently, the price should remain at the original level even though a new technology is available. "Forward-looking economic costs are not simply the firm's avoidable costs after it has made investments. If that were the case, there would be no transaction-specific investments" (Sidak and Sidak, 1997, p. 424). Consequently, Sidak and Spulber argue that prices *should* include historic costs (which they term "embedded costs").

In the regulated context, the expected revenue of the incumbent [firm] happens to be based on embedded costs because, under cost-of-service regulation, the [firm's] capital costs are necessarily used to calculate revenue requirements. That calculation does not mean that embedded costs are part of the firm's economic cost. Nevertheless, because the regulated firm's expected revenues reflect those costs, the expected values should be used to compensate the firm. The fact that the firm's capital has a lower (or higher) replacement value in comparison with embedded cost is not relevant to the compensation decision. The embedded cost is part of cost recovery because it underlies the incumbent firm's investment-backed expectation (Sidak and Spulber, 1997, p. 425).

But the possibility that future entrants may face different costs from the incumbent is part of the risk of doing business in a competitive market. Simply because consumers *can* enter into contracts with firms in a competitive market does not mean that a regulated firm should be compared to a competitive benchmark which assumes that firms and consumers would have entered into a hypothetical contract which guarantees the firm a zero (or positive) economic profit. As such—in the context of reviewing Sidak and Spulber's (1997) book—Trebing makes the following observation.

While the utility must be given an opportunity to earn a fair return, it is not guaranteed a return. This is particularly true in the face of major changes in demand and the technology of supply (Trebing, 2000).

Where prices are restricted by regulators on the basis of the currently optimal asset configuration, given current technology, then there may be downward pressure on prices. But the firm could have allowed for this risk at the time the investment decision was made by forecasting its future revenues with such advances in mind, just as a firm in a competitive market has to do (or as regulators would have to do in order to assess SAC and IC). 'Regulatory risk' is seen as leading to inefficient under-investment, although the risk effects of regulation can cut both ways, depending on the type of regulatory regime and how it is implemented (e.g., Brennan and Schwartz, 1982; Taggart, 1985; Sidak and Spulber, 1997, Ch. 13; Small and Ergas, 1999). However, to counterbalance any risk which provides disincentives to investment, regulators could perhaps allow for some compensatory upward adjustment to the risk-based

component of the allowable cost of capital (up to the equivalent competitive level, unless the act of regulation itself introduces greater levels of risk).¹⁹

In any event, permitting changes in technology or optimal asset configuration to place downward pressure on the forecasts of future prices does not necessarily mean that the incumbent firm will not be able to make a return on, and return of, its capital outlay. In the period before the new technology is introduced, or until a different asset configuration becomes least cost due to changing demand, incumbent firms will have a cost advantage over hypothetical entrants (or self-producing consumers). The hypothetical entrant is not able to “hit and run” because of the need to “sink” capital in durable assets. A potential entrant has to be sure that it can recover its own capital outlay—in other words, the entrant’s current and future prices need to be as *sustainable* as the incumbent’s (§3.3.3). Before the new technology or asset configuration becomes viable, the hypothetical entrant has the same costs as the incumbent. Consequently, if the competitor did enter during that period, then compared to the incumbent, the entrant would have a shorter time period over which to recover its costs, before the less expensive network design becomes viable. Hence, the entrant’s *annualised* costs become higher than the incumbent, which acts to prevent entry. During this period, the incumbent has a “supra-optimal” asset design compared to a potential entrant. Therefore, it can raise its quasi-rent to the competitive price level.²⁰ This will enable the incumbent to more rapidly recover its capital costs, as long as the regulator is aware that the zero economic profit constraint should apply intertemporally, rather than on an annual basis, and does not restrict the incumbent’s temporarily high quasi-rent. (Demonstrating that incumbent

¹⁹ In discussing access pricing (§5.1.2), Guthrie *et al.* (2000) indicate that the regulatory risk inherent in forward-looking prices suggests that historically-based prices will be more dynamically efficient. They have highlighted that, given the uncertainty over the future *cost* of large projects involving irreversible investments, it is critical that firms face the right incentives to invest in the first place. They examine whether an incumbent regulated firm in a world of cost uncertainties will invest earlier if access prices are set under a backward or forward looking cost rule, and ask which rule leads to higher overall welfare. Guthrie and his colleagues conclude that setting the access charge based on historic costs allows the incumbent to shift some of the cost of “investing early” onto its competitors. Unlike forward-looking cost rules, such backward-looking access prices do not expose the firm to risks inherent in the uncertain future movements in costs, and this is particularly the case if future costs are subject to “downward drift”, such as would be the case due to technological advances or increases in economies of scope due to demand growth (§6.3.2). Hence, they conclude that except in special situations where costs are *climbing* rapidly, backward-looking cost rules should be adopted for determining dynamically efficient access prices. On the other hand, if a forward-looking cost rule does have to be applied, they recommend that “the implicit rental rate should be set at a level considerably higher than the risk-adjusted discount rate. A high rate is required to compensate the incumbent for the risk it bears when faced with forward looking access charges. Without such compensation, the incumbent will delay investment too long”.

²⁰ In other words, the opportunity cost of the network is currently higher than its historic cost, since in an alternative use by a competing firm, the network has the value of a more expensive greenfields network. The greenfields network is “more expensive” not in terms of total cost, which might be the same, but the economic rent would have to be higher to recover that total cost, given that the time period for recovery is shorter.

firms can achieve a zero economic profit while charging intertemporal subsidy-free prices in the face of lower-cost later entry is the purpose of Chapters VIII and IX of this thesis. The key is in permitting accelerated economic depreciation—discussed in Chapter VII).

In fact, Sidak and Spulber (1997, p. 276) themselves make it clear that a firm’s investment-backed earnings are *not* necessarily related to “the purchase costs of the regulatory assets, nor their resale value, nor their replacement costs. The utility placed the assets in service in expectation of the earnings that would be received”. Sidak and Spulber’s concern arises where firms, with explicit sanction from regulators, have based their *ex ante* expectations on a return on, and return of, their purchase cost of durable assets. *Ex post*, those investment-backed expectations will relate to, what *now*, are historic costs. Hence, the reason Sidak and Spulber are endorsing the recovery of historic costs, is because historic costs simply happen to coincide with investment-backed expectations under the particular case of traditional ‘*rate base regulation*’ of utilities in the US (§6.4.1). Similarly, just because regulators might not allow regulated firms an adequate risk-adjusted return on capital in practice—which would be a situation of “regulatory failure”—is not necessarily a justification for pricing based on historic costs in order to mitigate such risk. Rather, it is an argument for regulators to allow firms to achieve a return on capital that adequately compensates them for their risk, in a manner comparable to that which they could achieve in a competitive market. Neither argument provides a compelling *carte blanche* endorsement of the recovery of historic costs under all circumstances.

6.3.4 The Case for Compensation Due to Breaches of the Regulatory Contract

Sidak and Spulber’s concern regarding the protection of the investment-backed expectations of incumbent firms leads to the second possible argument against using the currently optimal network design as the pricing benchmark. Incumbent firms that have been subject to regulation may deserve to recover the costs of decisions which were made in good faith in the past, but have now proved to be sub-optimal, especially if those decisions were explicitly sanctioned by regulators. Sidak and Spulber (1997, p. 425) argue that changing the rules of the game on regulated firms breaks the “regulatory contract”, and “contract law does not require paying the incumbent the offer of the entrant”. Rules can be changed either by a change of the regulatory regime, or simply by deregulating the industry in question. They argue that firms impacted by such changes deserve to be compensated (§5.3.2), since they have ended up holding onto some stranded assets (§3.5.1). Sidak and Spulber imply that the implicit regulatory contract is equivalent to a freely-negotiated long term contract with consumers in a competitive market. Consequently, regulators are no more entitled to break their implicit contract with a regulated firm than consumers would be entitled to break-off a contract with a firm in a competitive market simply because they no longer are happy with the price which they are paying.

Similarly, Teplitz-Sembitzky (1990, p. 27) had earlier considered that “if costs which were prudently incurred are disallowed because unforeseen changes in parameters have reshaped the

regulator's perception about what type of power facilities are needed, the 'regulatory contract' punishes bad luck rather than bad decisions". Although Teplitz-Sembitzky concedes that such a guarantee of cost recovery does not occur in competitive markets—assets which the market does not bear will be “written off” the firm's books—competitive firms are not under the “obligation to serve” commonly imposed on regulated firms.

There are of course arguments against this notion of the regulatory contract (e.g., Trebing, 2000). However, this thesis does not address the question whether such compensation is warranted or not. Rather, the key point to note here is that the issue of compensation is an entirely different one from determining the perfectly contestable benchmark price. In a competitive market, consumers are free to enter into a long term contract with a supplier in the expectation that, during the period of that contract, a cheaper technology, or less expensive asset configuration will *not* emerge. Of course, they may be wrong, and simply because they were wrong does not entitle them to break their contract with an incumbent firm, in order that they can take advantage of the less expensive technology. As Sidak and Spulber (1997, p. 425) observe, to allow a breach of contract would simply result in a transfer of income from the seller to the buyer. On the other hand, this does not alter the fact that the new technology is now the perfectly competitive or contestable benchmark. In a competitive market the advent of new technology might require exiting firms to “write-down” the value of their now-obsolete assets. If compensation were considered to be warranted for a regulated firm, due to a real break in the “regulatory contract”, then to be efficiently consistent with a perfectly contestable benchmark, such compensation would have to be made through other means than via a reflection of that compensation in the price.

6.3.5 A Return to Boiteux's Pricing Prescription

Neither historic expenditure nor compensation costs are relevant to the perfectly contestable cost of serving entire current and future market demand, as evaluated at the current moment in time. While the optimality of an incumbent firm's investment decisions *are* dependent on its past actions, even were those past actions themselves optimal at the time, none of those decisions have any bearing on a potential entrant evaluating the optimal way of serving current and future market demand on a greenfields basis at the present moment in time. Prices that reflect historic costs or compensatory costs, and as a result are higher than the prices which could be offered by a potential entrant, will not survive in a perfectly contestable market over the long run. In a perfectly contestable market, an incumbent monopolist needs to forecast future revenues on the basis of the offerings of potential entrants, and not on the basis of making a return on, and return of, its actual costs. Consequently, the incumbent should not invest in the first place if the hypothetical (and sustainable) future prices offered by potential entrants do not allow it to at least make a zero economic profit.

The need to include future replacement costs of existing assets into any calculation of SAC or AIC focuses on the incumbent firm's existing and future cost structure, rather than the entrant's cost

structure. For one thing, the date of asset replacement will be different from both the incumbent's and potential entrant's perspective. This would be acceptable in theory, since incumbents and entrants are assumed to have symmetric cost structures at the moment of entry. But in reality, non-fungibility causes intertemporal interdependence of costs, consequently, incumbent firms may face different costs from a potential entrant in serving the current and future consumers, even if they have access to exactly the same technology. Simply because entrants can supply current and future market demand does not make the incumbent unsustainable however; all the incumbent needs to do is lower its prices to the level that would be offered by the potential entrant. Even if the potential entrant's costs are lower, there may be real world barriers to entry that make the replacement of the incumbent by a new firm entirely infeasible (§3.4.5).

However, the issue to hand is not the strategic interaction between potential entrants and the incumbent, but the question of the perfectly contestable benchmark for determining bounds on subsidy-free prices. Basing SAC and IC on the real-world *incumbent's* costs is not consistent with the perfectly contestable benchmark, even if the incumbent were productively efficient, and had been making dynamically efficient investment decisions using the optimal investment rule. This is because a potential entrant may be able to serve the market—comprising both current and future demand—with assets that are more productively efficient, due to changes in technology. But even if technology has not changed, and both the potential entrant and the incumbent have access to the same technology, the entrant may be able to construct a more dynamically efficient asset configuration—one better suited to the current and future state of the market, whereas the incumbent's has evolved to serve demand as it changed in the past. The entrant is able to apply the optimal investment rule as if the market were to be served by an entirely greenfields construction project.

If the entrant actually forced the incumbent out of the market by constructing its own network, this would of course, be dynamically inefficient, due to the imperfect fungibility of the incumbent's own now-duplicated assets. But again, the issue is simply to determine what the incumbent's prices would need to be in a perfectly contestable market to stop that from happening. This calculation cannot be made on the basis of the incumbent's costs; it has to be made on the basis of constructing an entirely new optimal network.

Remarkably, this highlights that Boiteux's (1956) prescription for pricing capacity was almost correct after all (although for somewhat different reasons, since Boiteux was more concerned about price *stability*). Setting aside any issues relating to the problems of providing regulatory incentives for firms to price at the efficient level, the benchmark toward which regulation should be directed, is to set prices based on what *would* be optimal if capacity were perfectly adjustable (§4.2.6). Capacity is almost perfectly adjustable under constrained market pricing, because the benchmark is the costs of a hypothetical entrant ascertained from moment to moment. There is a slight difference between this

approach and Boiteux's, however. While Boiteux allowed for the optimal capacity to include spare capacity in order that it be sized optimally to meet current and future demand, accounting for the effect of losses, he appears to have assumed away spare capacity associated with asset *indivisibilities* (§6.1.1).²¹ Hence, under constrained market pricing, the hypothetical entrant's capacity can be optimally adjusted from moment to moment, subject to the realistic constraints of asset indivisibilities (in the same manner which the incumbent has to consider asset indivisibilities when successfully applying the optimal investment rule).

6.4 Regulating for Intertemporal Efficiency and Fairness

6.4.1 Incentive Regulation versus Rate of Return Regulation

Baumol and Sidak (1994a, p. 87) do make it clear that “the pertinent stand-alone cost is not the actual cost incurred by the regulated firm, but rather the cost that would be incurred *by the entry of a hypothetical efficient entrant*”. They cite this benchmark as being one of the key differences between using constrained market pricing as a regulatory tool, and the traditional ‘*rate base regulation*’ or ‘*rate-of-return regulation*’. Rate of return (ROR) regulation is widely considered as not placing sufficient *incentives* on firms to be productively or dynamically efficient. In fact, ROR regulation is often seen as doing the exact opposite, by providing the regulated firm with undesirable or “perverse” incentives. The best-known distortion related to traditional regulation is the Averch-Johnson effect (Averch and Johnson, 1962), which views ROR regulation as resulting in “cost-plus” pricing. If profit-maximising firms are regulated with respect to an allowable rate of return on capital, they have a clear incentive to expand their capital or ‘rate base’ beyond the optimally efficient level, in order that they can earn a higher absolute level of profit than would otherwise be the case. While ROR regulation can cause output to increase, and correspondingly raise consumer welfare, it can also distort input choice, thereby reducing efficiency and welfare. Firms have no incentive for cost saving, growth in efficiency, or cost-reducing innovation (e.g., Hay and Morris, 1993, pp. 626-627).

Although the much-cited Averch-Johnson effect has been subject to much scrutiny, criticism and extensions, Baumol and Sidak suggest that constrained market pricing does not suffer from the same efficiency problems as ROR regulation in general. Since the incumbent firm is compared to a hypothetical entrant, the regulated firm is not “condemned” to an automatic reduction in its price ceiling if it succeeds in reducing its own costs. Baumol and Sidak indicate that this is because such

²¹ By contrast, Turvey (1969) did not suggest that price be based on what *would* be optimal, he assumed that investment *was* optimal, taking implicit account of indivisibilities and other real world considerations in determining the optimal investment configuration (§4.1.4, fn. 11). What neither Boiteux nor Turvey addressed, was any explicit concept of *opportunity* cost. However, shortly afterward, Littlechild (1970b) explicitly incorporated opportunity cost into Turvey's (1969) formulation of marginal cost, and pointed out that Turvey “wishes to indicate his accordance with the analysis”.

improvements in efficiency do not reduce the costs of the hypothetical entrant upon which the ceiling is based.

On the other hand, Baumol and Sidak feel that the SAC ceiling still acts as a disincentive to investment in research and development, and a firm that achieves a revolutionary innovation will still be denied a corresponding increase in net revenues. To resolve this problem, Baumol and Sidak (1994a, pp. 88-90) advocate the application of ‘price cap regulation’ of final product prices—an approach they attribute to Baumol (1968) himself—in conjunction with stand alone cost ceilings (as well as access pricing for intermediate products; §5.1.2). Price cap regulation is seen as providing the required incentive for efficiency, innovation and productivity growth that is absent under a rigidly fixed profit ceiling, and it has cropped up in discussions regarding the further reform of New Zealand’s power distribution sector (§2.4.7).²² Furthermore, price cap regulation is seen to eliminate perverse cross subsidisation incentives that exist under cost-plus regulation. This is because prices are decoupled from costs, thereby providing no incentives for utilities to manipulative cost allocations in a manner which creates cross subsidies (Braeutigam and Panzar, 1989). However, as Lowry and Kaufmann (1998, pp. 35-37) explain, the cross subsidies in question are those between regulated and competitive activities, or between utility and affiliate operations, rather than between consumers (§4.3.1). And not all authors would even agree that price cap regulation necessarily provides incentives for the removal of the cross subsidies between regulated and competitive activities (e.g., Loubé, 1995).

According to Baumol and Sidak, price cap regulation builds on a “virtue” that derives from the phenomenon of ‘regulatory lag’. This lag is the regulator’s time lag in adjusting permitted prices in response to changes in cost or market conditions. The outcome of this lag is that it allows firms to enjoy superior profits as the reward for improved efficiencies until such time as regulators are able to reassess costs and impose new limits. Baumol and Sidak liken this outcome to that under competitive markets, where cost-cutting innovators temporarily enjoy higher profits until rivals are able to introduce their own cost-reducing measures. Regulatory lag thus provides the incentive required to elicit innovation and productivity growth. As Sidak and Spulber (1997, p. 375) explain: “monopoly profits provide economic incentives for competition over time, creating incentives for innovation, investment and market entry”.

Once again, Baumol and Sidak are stepping away from the original strict benchmark of perfect contestability, since they are using “real world” competitive markets here as the benchmark, rather than a

²² Although the fostering of innovation has not been an explicit goal of New Zealand’s power sector reforms, the broad economic statement on markets dominated by natural monopolies included an objective of rewarding innovation (§2.3.1, fn. 35). In discussing access (or “interconnection”) prices in the New Zealand telecommunications industry, the Ministry of Commerce and The Treasury (1995, para. 42) express the concern that, should the access price be too low, this may provide disincentives for the monopolist to innovate or invest.

perfectly competitive or contestable one.²³ They suggest that “a rigid profit constraint violates the precepts of the competitive-market model for regulation, because the competitive market *does* permit the successful innovator to earn especially generous profits” (Baumol and Sidak, 1994a, p. 87). However, in a perfectly contestable market, firms do not earn temporarily higher profits; they achieve a zero economic profit. The amount which an efficient incumbent can expend and recover on research and development will be equivalent to that of a hypothetical entrant using R&D to provide a net reduction in costs, and thus compete more effectively.²⁴

Price cap regulation works by firstly determining an initial price ceiling based on SAC, or “some defensible proxy”. This price ceiling is allowed to rise automatically each year by a percentage equal to the rise of some widely accepted index of inflation—such as the consumer price index (CPI) or retail price index (RPI)—*less* some percentage *X*. This ‘X-factor’ is typically greater than or equal to the industry’s rate of productivity growth in the past, or as an estimated target rate of productivity growth, with either value usually derived from econometric methods.²⁵ Such an approach allows efficient firms to earn superior profits, but also continually provides pressure on them to be more efficient. In addition, a Z-factor may be included in the regime that allows flexible, one-off changes to the price cap in response to exogenous conditions, such as changes in national or local government policies or tax rates. This factor can be used to reduce “regulatory risks” (e.g., Lowry and Kaufmann, 1998, p. 17).²⁶ Its

²³ Similarly, Sidak continues to use this “real world” competitive benchmark in his work with Spulber, for instance: “In its pricing recommendations and cost estimation methods, however, the [Federal Communications Commission] paints an incorrect portrait of how competitive pricing works. Technology and competitive entry occur with lags in competitive markets” (Sidak and Spulber, 1997, p. 421). Sidak goes even further though, by recommending that price caps should only be a *temporary* measure, and should be “phased out as soon as possible”, since “adjustments to price caps based on productivity and inflation indexes are unlikely to achieve the flexibility required for the regulated firm to keep pace with changing market conditions” (Sidak and Spulber, 1997, p. 529).

²⁴ Witteloostuijn (1990) attempts to endogenise technological progress into an otherwise perfectly contestable market. Under his concept of “investment contestability”, firms engage in R&D for the purpose of reducing their average costs, and he argues that even with positive sunk costs, contestability can be guaranteed. “The essential feature of investment contestability is that incumbent suppliers assume that there are potential entrants who have carried out *similar* sunk investments at the *same* time as they did themselves. In the literature it is generally assumed that incumbents undertake sunk expenditures *before* potential entrants do so. That is to say, potential competitors face the need to bear *incremental* sunk costs upon entry” (Witteloostuijn, 1990).

²⁵ For example, Lowry and Kaufmann (1998, pp. 23-26) state that the productivity growth trend of investor-owned power distribution utilities in the US has been around 0.9% per annum. Consequently, this would suggest that the industry-specific *X* for regulating such distribution companies should be at least 0.9%. Although this is three times the apparent trend in the US economy’s ‘total factor productivity’ trend, it is far below the growth trend of more high technology industries such as telecommunications (at around 3%). In the UK, different *X*-factors have been set for each regional electricity distribution company (Rees and Vickers, 1995).

²⁶ The imposition of a local government requirement that any new conductors in a power distribution network be undergrounded provides another example where such a Z-factor could be used (§3.4.5).

administrative simplicity has meant that price cap regulation is sometime referred to as “regulation with a light hand” (e.g., Weyman-Jones, 1990). However, the approach does differ from New Zealand’s “light-handed” regulatory regime (§2.1.1 and §2.4.3), in that it still requires an industry-specific regulator, rather than industry self-governance, and revenues are subject to direct controls, or at least the threat of controls.

As the result of an influential report by Stephen Littlechild in the early 1980s (§6.1.3), the price cap approach to regulating monopolies has had widespread application in the United Kingdom, particularly in telecommunications, water and power industries, including the electricity distribution companies (e.g., Rees and Vickers, 1995). In power distribution, price cap regulation has also been applied in Argentina, as well as parts of Australia, and to a limited extent, the US (e.g., Lowry and Kaufmann, 1998, p. i). The experience with implementing price cap regulation for the (bundled) power distribution utilities in the UK, demonstrates that regulators have similar difficulties in evaluating industry costs as they would under ROR regulation (e.g., Kennedy School of Government Case Program, 1998). The problem of asymmetric information between the regulators and the regulated still exists (e.g., Hay and Morris, 1993, pp. 628-629; §2.1.1). The initial price ceiling or “defensible proxy” has to be defensible, and as such, have removed any pre-existing inefficiencies from the base price level, and thus the base asset value. In the UK, regulators underestimated the potential efficiency gains in setting the X-factor (they were actually set to be negative), and the base prices reflected an accumulation of past inefficiencies. Consequently, upon the implementation of the new regulatory regime, the distribution industry achieved substantial financial windfall gains. As Lowry and Kaufmann (1998, p. 13) wryly observe, “ironically, this is another way in which price cap regulation mimics competitive markets”.

Baumol and Sidak do not really provide a convincing explanation as to why price cap regulation differs markedly from the application of standard constrained market pricing. Firstly, regulatory lag, and the corresponding allowance of temporarily higher profits for efficient incumbents, is a feature of any regulatory regime where the allowable or benchmark price is not determined frequently—including traditional ROR regulation (e.g., Lube, 1995; Small, 1999b, pp. 15-16). Secondly, because SAC pricing is, at least from Baumol and Sidak’s point of view, based on a hypothetical efficient *entrant*, the effects of inflation are automatically taken into account, since that entrant’s costs will be based on current input prices. Thirdly, the incentive for productivity growth is also taken into account, since the hypothetical entrant is assumed to be productively (and dynamically) efficient, and may use R&D expenditures to be so. Finally, Baumol and Sidak do not mention that additional efficiency incentives will be placed on incumbent firms in industries which are subject to economies of scale and scope, a fact pointed out by Rees and Vickers (1995).

Notwithstanding Baumol and Sidak’s prescriptions relating to price ceilings, in later work with Spulber, Sidak takes a somewhat dim view of the ability of regulators in practice to determine the

contestable benchmark at all, as well as a somewhat less favourable view of the contestable benchmark itself (§5.3.2 fn. 20). As Trebing (2000) observes: “the authors view the [S]tate (particularly the regulator) as a major factor that will denigrate the general welfare if given an opportunity”.

It is not realistic to presume that a government agency is better equipped than market participants to sort out those technological changes to determine which technology is the best available or most efficient. The process of price adjustment to technological change cannot be predetermined by government fiat; it can only be revealed through market competition (Sidak and Spulber, 1997, p. 423).²⁷

In any event, Baumol and Sidak—like Sidak and Spulber—focus on telecommunications, an industry which in the last decade has experienced remarkable technological innovation and change. By contrast, electricity distribution (distinct from other parts of the electricity supply industry) is not a sector associated with rapid technological advances. Although advances in small scale generation technology (including fuel cells) may eventually make parts of the existing distribution networks obsolete—by making the “distributed utility” concept a reality (e.g., Weinberg *et al.* 1993)—in the meantime, some might argue that electricity distribution is not an entrepreneurial business in any case.

If utility executives wish to be entrepreneurs or captains of industry, let them resign their utility post and seek jobs elsewhere. The managerial tasks of a stable public utility are far different than the visionary requirements of a technological innovator. Entrepreneurs exist to create markets; public utilities exist to satisfy customer requirements in markets that already exist (Copeland, 1989).

Nevertheless, even without changes in technology, the average costs of providing distribution network connection may reduce due to additional economies of scope associated with gradual demand growth. Moreover, the greenfields design developed by a hypothetical entrant for a new network to serve an incumbent’s entire market may be able to take advantage of economies of scope not available to the incumbent, since the incumbent is constrained by its past investment decisions (§3.4.2).

6.4.2 Intertemporal Anonymous Equity: Protecting Consumer Interests over the Long Term

Equity concerns are traditionally associated with *consumers* (§4.3.1), whereas Baumol and Sidak are more concerned with ensuring a level playing field between *firms*. For instance, the benefits accruing from the efficiency improvements due to the imposition of price cap regulation will usually go to the regulated firm, rather than to consumers (e.g., Hay and Morris, 1993, p. 628). Consequently, considering intertemporal cross subsidies from the consumers’ perspective might be more appropriate than Baumol

²⁷ Similarly: “The notion that some regulatory ‘social opportunity cost’ could provide a better indicator of market value than market prices not only is economically incorrect, it leads to interventionist policy recommendations” (Sidak and Spulber, 1997, p. 375).

and Sidak's focus on firms (§4.3.6). There would be no problem if the incumbent, entrants and consumers were all perfectly symmetric, in terms of their access to information and the costs which they face. Since in real world markets this is unlikely to be the case, the point of view used as a benchmark can make a difference to what might or might not be considered efficient and fair prices.

From a consumer perspective, the hypothetical greenfields network is built *by* the consumers, *for* the consumers (although the actual construction work could of course be contracted out, just as many ELBs in New Zealand are doing; §3.2.3). Consequently, consumers always have an alternative supplier to the incumbent firm—theirself. Hence, although consumers might appear highly price inelastic, in a perfectly contestable market prices can only be raised to a level at which self-production would be triggered. Baumol's "weak invisible hand" only allows incumbent firms to price to some ceiling at which the lowest cost alternative supply option becomes attractive, and self-production needs to be considered as one of the viable alternatives (§5.1.3). As noted earlier (§4.3.6), and in one of the quotes opening this Chapter, Baumol and Sidak do acknowledge the goal of regulation to be consumer protection, but this could be interpreted simply as the complementary result of protecting potential entrants from predatory pricing.

The primary purpose of the price ceiling, aside from its role in eliciting economic efficiency, is to protect consumers—both household and business purchasers of ... services—from monopolistic exploitation through the imposition of excessive prices by the regulated firm. Similarly, the primary purpose of price floors, economic efficiency aside, is to protect actual or prospective rivals of the regulated firm from predatory pricing and related practices that can seriously handicap these competitors in the competitive process or drive them from the field altogether: (Baumol and Sidak, 1994a, pp. 51-52, [emphasis added]).

Notwithstanding this consumer protection objective, for Baumol and Sidak, the appropriate price ceiling benchmark are the costs faced by a hypothetical efficient entrant (§4.3.6). Consequently, for Baumol and Sidak, subsidy-free prices are a *firm-oriented* concept based on *entry*, rather than a consumer-oriented concept based on *self-production*. Partly this seems to have arisen from the Baumol group's (at least initial) preoccupation with unsustainability (§2.1.8, §3.3.3, and §3.5.3). Sustainability is an *entry* and *equilibrium* concept from the point of view of the *supplier*; an undefined number of firms contest with each other to supply a given market (cognisant that some consumers may be willing and able to self-produce). It is a concept of more interest to strategic behaviour in real world markets when the conditions of perfect contestability do *not* hold. On the other hand, subsidy-free bounds on prices are a *self-production* concept from the point of view of the *consumer*, always derived within a perfectly contestable market benchmark, which by definition is at equilibrium.

By always taking the point of view of the firm, rather than the consumer, Baumol and Sidak subtly deviate from this benchmark. In a perfectly contestable market, the perspective of consumers

toward stand alone costs and incremental costs is equivalent to that of potential entrants, assuming that perfect contestability requires that consumers have access to perfect information (§5.1.1). But Baumol and Sidak (1994a, p. 87) explicitly state that stand alone costs should be based on the costs of a hypothetical entrant, even where that entrant does *not* have access to proprietary information held by the incumbent.²⁸ Baumol and Sidak are clear that this means a cost advantage will accrue to the incumbent for some period of time, and “under the regime of price ceilings based on stand-alone cost, the firm retains pricing freedom and some incentive for economy that was absent” under traditional ROR regulation. Consequently, the consumers are, by default, also considered to be excluded from such information.

Yet Sidak, with Spulber in their later work, goes even further by implying that, if in reality there are *no* potential entrants, then it does not make sense to consider *any* alternatives, even hypothetical ones. Sidak and Spulber’s (1997, p. 421) view that: “Setting prices on the basis of competitor’s costs is a good competitive strategy, but only when market alternatives are available”, appears to be a clear sanctioning of monopoly pricing, and price discrimination. This seems to be a circular argument. The very fact that competition is not viable is used to justify monopoly pricing. However, considering the consumers themselves to always be the “suppliers of last resort” places a cap on such monopoly power.

Small (1999b) observes that regulation typically concentrates on a *negative* objective—namely, restraining the abuse of market power by monopolists. However, he points to an alternative perspective; the “promotion of consumer interests”. Such an approach is cited as being used in the regulation of telecommunications in Australia, where the consumer perspective is embodied in the “long term interests of end-users” (LTIE) test. The regulator is required by law to have regard to long term consumer interests when making determinations in the sector. Small observes that one of the attractive features of the LTIE test is that it explicitly incorporates a long term planning horizon for investment. Hence, the approach “guards against the possibility that over-vigorous regulation will, by cutting prices excessively for current consumers, reduce investment in the assets that will be required to adequately serve future consumers”.

The LTIE concept is similar to that of achieving anonymous equity in a perfectly contestable intertemporal world. All the possible coalitions that present and *future* consumers will form with each other around an ‘*intertemporal level playing field*’ must be assessed. Future consumers and their demand must be considered due to the non-fungibility of greenfield investments. If future consumers are ignored then investments may be sub-optimal and there will be a loss of dynamic efficiency. Past consumers should be ignored however. Although, some consumers may have had demand in the past, have demand

²⁸ Baumol and Sidak (1994a, p. 87) point out that, at least in the US, it is well established in antitrust law that the innovating monopolist has no legal obligation to share its proprietary information with its rivals.

currently, and continue to have demand for the service into the future—and as such the intertemporal interdependence of investment decisions may affect them—the very “anonymity” of consumers means that any link between current and past consumers is broken. On the other hand, although similarly, any commonality between current and future consumers is not explicitly considered, the application of the optimal investment rule requires the intertemporal interdependence of current and future costs to be accounted for, and future costs are driven by future demand. But because current and future consumers are evaluating investments on a greenfields basis, historic costs are ignored—as are historic demands—by association.

Simply because historic costs are ignored, this by no means condemns the incumbent firm to bankruptcy. The optimal greenfields asset configuration from a consumer perspective may involve installing the same type of durable asset now, as that which the incumbent historically installed. Consequently, although historic costs are ignored in determining SAC and IC, a hypothetical entrant might be required to incur the same or very similar costs as the incumbent with respect to many assets. Depending on the time frame, in the absence of marked changes in technology, asset costs, demand, and the optimal asset configuration, the entrant’s costs might not differ substantially from that of the incumbent. Therefore, where the incumbent’s and entrant’s costs are similar, the incumbent should be able to make a return of, and return on, its capital outlays. Although the incumbent is not directly recovering its historic costs, it is making a return of, and return on, the notional asset which the consumers would install themselves at the present time in a greenfields site. In reality, this will often amount to the same level of revenue.

But as this thesis demonstrates in Chapters VIII and IX, the incumbent can still earn a zero economic profit where regulators restrict prices to within perfectly contestable bounds, even if the optimal asset configuration changes over time such that the total costs of the *new* optimal configuration are *lower* than the total costs of entirely replacing the incumbent’s existing—and now sub-optimal—configuration. This apparently paradoxical result occurs because of the opportunity costs associated with the time dimension. Baumol and his colleagues show concern that the time dimension causes unsustainability; a problem they attribute to “the opportunity cost of tying up resources by building spare capacity in anticipation of the demand for capacity in the future” (BPW, p. 473). Yet this thesis shows that—at least for a very simple two-good/two-period model—the set of subsidy-free prices associated with any optimal investment program will allow the incumbent firm to make a zero economic profit. Furthermore, it is shown that unsustainability is likely to be much less of a problem than the Baumol group claims. Much of the unsustainability arises because, in their model of intertemporal unsustainability, BPW assume that the incumbent is unable to price “discriminate” even where the costs contributed by different products are different (in which case charging distinct prices for the products would not be discriminatory after all).

6.4.3 *Intertemporal Cross Subsidies, Amortisation and Depreciation*

To conclude the discussion in this Section of intertemporal efficiency and equity, there is one other important difference between the firm and consumer perspective of prices. In evaluating investments, an unregulated firm does not necessarily have the same timeframe in mind as particular consumers or groups of consumers do. A potential entrant in a perfectly contestable market is interested in achieving zero economic profit over the finite lifetime of an investment in a durable asset. Variations in net revenue over time make little difference as long as the net present value of revenues equals the initial capital outlay (§7.1.3). But different consumers will come and go over time. For instance, in a distribution network, the connection capacity at a particular site may last up to 60 years or more (§6.2.2). The site itself, however, may change ownership many times during that lifetime. Consequently, to consumers, any intertemporal variations in required revenue over time make a substantial difference. A firm that decides to recover its capital cost rapidly, say through an accelerated depreciation schedule (§7.3.3), will place the burden of paying for the greater part of the durable asset on current rather than future consumers. However, if current consumers are no more price inelastic than future consumers, there is not necessarily any “second best” basis for such intertemporal price discrimination, and treating these consumers differently may result in an intertemporal cross subsidy.²⁹ As Baumol (1971) observes, the pattern of future prices should be “designed to minimize the distortion of consumer choice”.

For instance, although not addressing the issue of cross subsidy, Rogerson (1992) points out that, where consumer preferences do not change with time, the efficient Ramsey prices for a regulated monopolist subject to a zero economic profit constraint will be *constant* values in real terms, meaning adjusted for inflation. This also assumes that variable costs do not change with time. Prices net of variable costs are generally termed the ‘*payments to capital*’ or ‘*amortisation*’. Payments to capital comprise two components, a return *of* capital over a durable asset’s lifetime—known as ‘*depreciation*’—as well as a return *on* capital, in other words, economic rent. Since the Ramsey efficient prices do not treat consumers with the same cost and demand characteristics differently, these constant payments to capital could also be considered intertemporally equitable.³⁰

The time path of prices will affect consumers’ purchase decisions. This is essentially a multi-product monopoly pricing problem as first considered by Baumol and Bradford (1970)

²⁹ There is a substantial body of literature on ‘intertemporal price discrimination’. However, this work—beginning with Stokey (1979)—attempts to explain why it might be optimal for a monopolist to engage in such a strategy of intertemporal price discrimination in an attempt to exploit consumers with high reservation preferences first, rather than to address the issue of intertemporal cross subsidy.

³⁰ As Teplitz-Sembitzky (1990, p. 25) observes: “if demand is constant over time, strict economic reasoning suggests that the plant should depreciate at a rate equal to the rate of discount”, which would result in an annuitised capacity cost. Consequently, Teplitz-Sembitzky concludes that, unless there are changes in the elasticity of demand, no convincing economic case can be made for intertemporally discriminatory pricing.

where “electricity each year” is viewed as a separate good. One can solve for the price path which maximizes consumer surplus subject to the firm breaking even. This optimal price path implicitly determines the optimal amortization schedule. Under simple natural assumptions, the optimal price path involves constant real prices (Rogerson, 1992).³¹

Although neither the concept of intertemporal price discrimination nor intertemporal cross subsidy is explicitly present in the contestability literature, Sidak and Spulber (1997, p. 200) make a single allusion to a concept of “intertemporal subsidies”. They associate this concept with regulators requiring incumbent firms to extend their “depreciation schedules such that the payments by consumers have been substantially less than the benefits they have received”. Sidak and Spulber consider that: “The depreciation schedule required by the regulator meant that consumers received service at a price that paid for the retirement of a lesser amount of the utility’s invested capital than was realistic in light of the economic obsolescence of assets precipitated by newer, more efficient technologies or changes in regulation”. In other words, mismatches between asset lifetime and the depreciation schedule can favour consumers, over time, to the incumbent firm’s detriment, although—as Schramm (1991) observed—it is possible that the opposite situation could occur (§4.1.5). On the other hand, Sidak and Spulber (1997, p. 397) do not discuss the concept of economic depreciation at all, and while they acknowledge that the economic “value of a firm is the present discounted value of returns”, they observe that translating this “forward-looking” approach into “accounting data will inevitably require a number of compromises”.

Parsons (1998) draws attention to a similar depreciation mismatch problem, highlighting situations where “allowed depreciation rates in the past” were “too low to reflect market rates of depreciation of capital assets”. He notes that “while the literature on cross-subsidisation may ignore embedded [i.e., historical] costs”, the difference between forward-looking costs and accounting (or historical) costs due to depreciation, “has equity implications and implications for dynamic efficiency”. Parsons points out that “new investors consider the regulatory agencies’ records for establishing rules that allow for the recovery of investments” (i.e., acceptable approaches to depreciation). And Baumol and his colleagues (BPW, p. 476) indicate that regulatory rules on depreciation policy can be a source of unsustainability. Since unsustainability and subsidy-free prices are but two sides of the same coin, this also suggests that regulatory impositions on acceptable depreciation schedules may also be a source of (intertemporal) cross subsidy (§3.3.3). Consequently, given the significance of intertemporal pricing issues associated with the fact that investments in durable assets have a finite lifetime, the theory and practice of asset *depreciation* (and *valuation*) methodologies are examined in the next Chapter.

³¹ Rogerson (1992) is here drawing on Baumol’s (1971) early work on optimal depreciation (§7.1-§7.2).

CHAPTER VII

DEPRECIATION AND VALUATION OF DISTRIBUTION NETWORK ASSETS: HISTORIC COST OR REPLACEMENT COST BASED?

The optimised deprival valuation (ODV) methodology used in New Zealand is soundly underpinned theoretically and is capable of giving a valuation outcome consistent with contestable market outcomes. Use of the ODV methodology is therefore neutral or positive in terms of promoting behaviour consistent with the achievement of the Government's overall economic efficiency objectives: One of the principal authors of the New Zealand Ministry of Commerce's original ODV Handbook, Jeffrey Wilson (2000b, p. 4)

The primary problem facing consumers is that lines businesses are able to write-up their assets using the ODV methodology and then use much higher values as a justification for capturing monopoly rents. ... The ODV methodology relies on the deprival concept. Yet deprival was never suitable as the conceptual basis for rate base determination, was never proposed for that purpose by the original theorists who developed the concept, and has not established a successful track record in that role anywhere in the world: from "Lining Up the Charges – Electricity Line Charges and ODV" (Bertram and Terry, 2000, p. iii)

If the valuations are wrong, then the line charges are wrong: from Report of the Ministerial Inquiry into New Zealand's Electricity Industry (Ministry of Economic Development, 2000a, p. 42)

The last Chapter closed by introducing the link between intertemporal equity, capital cost recovery of investment, and depreciation schedules. In this Chapter, the issue of capital cost recovery through depreciation payments is investigated in more detail. In particular, the issue of the "efficient" and "fair" *absolute* price level for a regulated firm as a whole is addressed (§5.1.2), since allowable annual revenue has been—and at the time of writing, continues to be—a key area of controversy in New Zealand's reforms of the power distribution sector (§2.4.6). Although New Zealand's light-handed regulatory regime provides electricity line businesses (ELBs) with reasonable freedom in terms of the level of their annual receipts, like more traditional forms of rate-of-return regulation (§6.4.1), net revenues are still compared with—but not explicitly restricted by—an accepted rate-of-return applied to an asset base valued using a standard methodology (§2.4.3). The Chapter opens by examining the general issue of asset valuation in theory and in practice—as well as the inextricably-linked issue of asset depreciation—and then examines the optimised deprival valuation (ODV) methodology, the approach which New Zealand's ELBs are required to use in valuing the rate base of their network assets.

7.1 Capital Cost Recovery and Economic Asset Valuation

7.1.1 Capital Cost Recovery and Depreciation

The concept of “economic depreciation” goes back to Hotelling (1925) who outlined that depreciation of an asset comprises the change in two factors: the “service potential” of that asset; and the “value of that service potential”. The first factor—service potential—relates to an asset’s physical deterioration, for example, the decline of an asset’s effective capacity (and/or service quality) over time. At the end of an asset’s lifetime, by definition, that asset has zero service potential. Since even a partial decline in service potential may require the firm to engage in some costly remedial action—such as to make up for the lesser level of capacity—reduced service potential has implications for an asset’s value.¹ The second factor is that—assuming an asset provides some input into the production of a good or service for which demand still exists—the remaining service potential of the asset will itself have some value.

Baumol (1971) describes the “depreciation decision” as being “the choice of prices that permits recovery of a firm’s investment in an asset”. This definition involving a “choice of prices” is itself a recognition that real world markets are not some perfectly competitive ideal, since in a perfectly competitive market, firms are ‘price-takers’ rather than ‘price-setters’. The concept here is that firms have some freedom to set prices within regulatory or competitive—but not *perfectly* competitive—constraints. Baumol then discusses what is meant by “recovery”. As he points out, “in a changing world in which costs and prices are not stationary, it is not immediately obvious how large a sum after the passage of t years is equivalent to some given initial outlay”. Advances in technology are cited as reducing the cost of replacing an asset at the end of its lifetime, whereas inflation works in the contrary direction by “depreciating” the value of the currency received in repayment.

While the firm can keep its service potential intact in the face of constant demand by simply recovering the replacement cost of the asset by the end of its lifetime, Baumol states that the crucial factor is that the investor should recover the full “opportunity cost” of the investment (§4.1.5). Baumol considers that, “from the point of the investor, if no more than replacement cost is returned, the entire asset purchase can turn out to be a mistake. That is, the investment decision will have been worth his while only if at the end he receives back his initial purchasing power plus compensation for the use of funds”. Baumol justifies this result from the point of view of society, by noting that, if consumers of the services produced with the aid of that investment are unwilling to pay (in real terms) the opportunity cost of obtaining the asset in question, then construction of that asset represents a wasteful use of resources.

¹ Turvey (1969) draws attention to the fact that not all capital investment is made for *new* capacity. In reality, much capital expenditure goes into improving older plant (i.e., such as rehabilitating deteriorated capacity or extending useful asset lifetime).

Consequently, payments to capital should “return funds whose discounted value, after correction for changes of the price level, is equivalent to the cost of the investment. This may or may not be equal to the replacement cost of the asset”.²

Echoing Hotelling, Small and Ergas (1999) explain how “the economic depreciation of capital arises from the erosion of two stocks: total willingness to pay for the services of the asset, and the total service life of the asset”. They point out that when future economic depreciation is correctly estimated, there is no real *option value* of delaying investment (§5.3.4), because the depreciation refunds the firm for any loss in value that occurs as a consequence of investing too early.³

7.1.2 Valuation Under Regulation: The Circularity Problem

The question of “value” is one of the core questions of economic theory. With respect to a real world firm, the value of its assets is also a more practical question of cost accounting, particularly in relation to taxation, especially in the presence of inflation. As Hay and Morris (1993, p. 430) explain, there are many possible approaches to valuing a physical asset, for instance, using “*historic cost*”, “*value in purchase*” (i.e., replacement cost), “*value in use*” (i.e., present value of cash flows obtainable), “*value in sale*” (i.e., net realisable value), and possibly “*stock market valuation*”. Occasionally, but by no means always, some or even all of these values are the same. The *economic* asset value, however, is generally defined as being dependent on “forward-looking” revenues and costs, in other words, the present value of economic rent (§5.3.2) received from the “use” of that asset over its remaining life. Economic depreciation is therefore simply the change in an asset’s economic value over a particular period. Consequently, in a competitive market, future prices—which depend on market conditions—dictate economic asset value, which in turn determines depreciation.

For a regulated firm, however, revenues (and costs) are typically affected by the imposition of regulation itself. Therefore, there is an element of *circularity* in the idea of making a return on an asset’s

² Baumol (1971) made the assumption that taxation rules do not distort depreciation policies (and similarly, throughout this thesis, any impacts of taxation are ignored). Later, however, Baumol with his colleagues (BPW, p. 483) suggested that, because sunk costs inhibit real world markets from approaching the perfectly contestable ideal (§3.5.2), tax advantages could be provided in order to encourage accelerated rapid depreciation as one way of reducing sunk costs.

³ Ergas and Small (2000) outline the equivalence between economic depreciation and the option value of delaying investment, and Small and Ergas (1999) suggest that regulators can use standard empirical models for estimating real options to estimate economic depreciation on a forward-looking basis (although, strictly speaking, option value requires consideration of not just the costs of an investment, but its revenues as well). The similarity between option value and Turvey’s (1969) concept of marginal cost has been pointed out earlier (§5.3.4, fn. 36). And given that economic depreciation was linked to LRMC and SRMC by Baumol (§7.2.1), it is not surprisingly the link between option value and depreciation was also implied around the same time. For instance, Littlechild (1970b), building on Turvey’s paper, noted that “*amortisation on a new machine is the cost of bringing forward the purchase of machine from the next period—that is, the interest plus capital loss on its purchase cost, plus the discounted difference in running costs within the horizon, plus the discounted difference in terminal value*”.

“value in use”, since the asset value is itself based on the return achievable from that asset. The New Zealand Commerce Commission (2001, p. 9) expresses this difficulty as follows: “In competitive markets, prices are set independently of asset values, and the current value of a business or asset is able to be determined from the total present value of the cash flows it can generate—prices *determine* the value of assets”. However, where markets are not competitive, “prices may be *dependent on* the value of assets”. If regulators directly fix the levels of future prices, then this determines asset value. Traditionally, however, regulators—at least in the US power industry—have taken the opposite approach (§6.4.1). They have settled for approving a firm’s ‘rate base’ (i.e., its total asset valuation), imposed a depreciation schedule which determines how that rate base changes over time, and set an allowable rate-of-return (ROR) based on an assessment of the firm’s actual cost of capital.⁴

Such circularity, though, is not entirely limited to traditional ROR regulation. Vogelsang (1994) observes that ROR regulation, price-cap regulation, profit sharing and yardstick regulation are *all* various forms of price-level regulation. In the New Zealand context, Williamson and Mumssen (2000, p. 4) concur, indicating that alternative forms of regulation are not as distinct as they appear in theory when applied in practice, and Irwin (2000, p. 13) observes that, in reality, price cap regulation usually ends up focusing on the regulated firm’s profitability anyway. Wilson (2000b)—one of the architects of New Zealand’s ODV methodology—indicates that all of these regulatory approaches are to some extent based on valuation. Consequently, in responding to the question whether asset valuations should form part of the regulatory regime—posed in the Issues Paper for New Zealand’s Inquiry into the Electricity Industry (Ministry of Commerce, 2000)—Wilson responded that it is “inevitable” that “asset valuations will form part of the regulatory regime”. For instance, even price cap (or CPI-X) regulation requires an *initial* valuation base to be set (e.g., Simon Terry Associates, 2000) based on some “defensible proxy” for a firm’s overall stand alone cost (§6.4.1).

7.1.3 Economic Asset Valuation in Contestability Theory

Within the context of contestability theory, BPW (p. 385) define the “economic asset value” to be an asset’s “value in use”. Hence, they state that “the value of a capital asset can be assessed economically as the present value of the stream of future payments to capital, minus the present value of the stream of future costs of the net investments that, in part, make those payments possible”. Hence, the

⁴ Like almost every other issue, there is of course a controversy over how the cost of capital should be calculated. The generally accepted approach is that the weighted average cost of capital (WACC) should be derived from the Capital Asset Pricing Model (CAPM). As Grout (1995) suggests, the underlying formula for rate-of-return based on CAPM “is probably the most famous formula in finance”. Nevertheless, as Grout points out, there are other approaches, such as the “dividend-growth” model and the “sequential marginal costing” model (e.g., Paulo, 1992). Paulo observes that the cost of capital for a decision regarding a specific investment project (i.e., the application of the optimal investment rule) can be very different

economic asset value at the *beginning* of the i th year, over a period of X years, where $1 \leq i \leq X \leq \infty$, is defined in (7.1) below (BPW, 13G4).⁵

$$V_i = \sum_{j=i}^X PC_j \left(\frac{1}{1+d} \right)^{j-i+1} - \sum_{j=i+1}^{X+1} I_j \left(\frac{1}{1+d} \right)^{j-i} \quad (7.1)$$

where: PC_j are the “*payments to capital*” made at the *end* of each year; I_j are the (re)investment costs of capital made at the beginning of each year that are required to allow the asset to provide its service; V_i is the economic asset value at the beginning of the i th year, but assessed *after* any outlay of new capital (I) has been made for that year; and d is the post-tax discount rate, or post-tax weighted average cost of capital (WACC).⁶ If the discounted sum of all payments to capital equals the discounted sum of *all* investment costs (i.e., including I_1), then the initial investment cost must be such that $I_1 = V_1$. (However, this latter relation only holds under conditions of zero economic profit).

The annual revenue (TR_i) that needs to be obtained from investing in the asset is the annual payment to capital, plus any annual variable direct and indirect costs (VC_i) incurred in operating, maintaining and administering the asset (i.e., $PC_i = TR_i - VC_i$). Throughout the remainder of this thesis, the annual variable costs (per unit of capacity) are ignored, since these can be directly recovered by a corresponding increase in that year’s price (charged per unit of capacity), and have no effect on the required payments to capital or on the asset value.

If the asset is salvaged at the beginning of the $X+1$ th year (and therefore the service for which the asset is being utilised is discontinued), then any net asset ‘salvage value’ (S_{X+1}) is effectively a *disinvestment*; thus $I_{X+1} = -S_{X+1}$.⁷ Consequently, BPW (p. 386) define a “*set of payments to capital*” as being *any* stream of differences between total revenues and variable costs whose present values add up to the discounted cost of the capital in question (i.e., I_1), net of any returns from salvage (i.e., the asset salvage value discounted $X+1$ years into the future, for $i = 1$).

from the WACC for the firm as a whole. This issue of cost of capital is set to one side in this thesis, and WACC is assumed to be equivalent to the discount rate.

⁵ The reference 13G4, and other similar references in Chapters VII-IX, refer to the equation numbers in the standard contestability text (i.e., BPW, 1988).

⁶ The formulation for economic asset value in (7.1) is typical; BPW provide but one instance. For example, a similar definition, in both discrete and continuous time, is given in Hay and Morris (1993, p. 216).

⁷ The salvage value is effectively the economic value of the asset in (some alternative) use, but *net* of any sunk costs lost in the process of salvaging the asset.

7.2 Economic Depreciation

7.2.1 *Economic Depreciation, Marginal Costs and Spare Capacity*

In his early paper on optimal depreciation, Baumol (1971) used the term “economic depreciation payment” to refer to the sum of the return *on* capital in addition to the return of capital. Interestingly, Baumol (1971) equated the concept of an “economic depreciation payment” with the difference between long run marginal cost and short run marginal cost for a particular year. In fact, this is not surprising, since Baumol’s paper was an extension of Littlechild’s (1970b) earlier paper on marginal cost pricing in the presence of (intertemporally) joint costs, which, in turn, had extended Turvey’s (1969) seminal paper on marginal cost (§4.1.3). To avoid any confusion which might arise from using the term depreciation in the sense of total returns, Turvey had suggested the term ‘*amortisation*’ to relate to the entire payment to capital. Of amortisation, and therefore of “depreciation plus interest—i.e., the capital charges—appropriate to any one year”, Turvey indicated that either “can be obtained as the excess of marginal cost in that year over marginal running costs”.

Turvey (1969), however, had looked at the problem from a different angle than Baumol. His avoided cost approach to determining “incremental system costs” (§4.2.6), meant that “we have solved the problem of marginal costs without having to introduce the concept of depreciation”. Payments to capital (i.e., Turvey’s “depreciation plus interest” or “amortisation”) were a direct outcome of the incremental cost calculation—which was solely based on forward looking costs—rather than the sum of a return of, and a return on, some pre-defined asset value. Consequently, Turvey concluded that: “it is clear that if the problem of defining marginal cost is solved without any depreciation concept the solution can be used to derive such a concept”.

Baumol (1971) suggested that “it is useful to think of the depreciation problem as an intertemporal peak-load pricing problem”, which is how Littlechild (1970b) had formulated the problem. Littlechild (1970b) expressed an optimal pricing rule as “set price in each period equal to marginal operating cost plus marginal (opportunity) cost of capacity”, and an optimal investment rule (§4.1.4) as “purchase equipment up to the point where the sum over all periods of the marginal (opportunity) values of the capacities it provides is equal to its marginal purchase cost”. Littlechild concluded that “price should be set to just fully utilise capacity”. Moreover, he concluded that “the amount set aside for amortisation varies from period to period, depending on demand, and may even be zero in some periods”. The possibility that amortisation could be zero arose because Littlechild assumed that the “marginal capacity cost is zero when there is spare capacity”, echoing the advocates SRMC pricing discussed earlier (§5.3.2).

Yet Littlechild formulated his cost model using linearly homogeneous short run and long run production functions. As Panzar (1976) later demonstrated (§4.1.2), when the peak load pricing problem is formulated with more realistic cost conditions—namely that short run production functions often

exhibit decreasing returns to scale, while the opposite is often the case for the long run costs of capacity—off-peak periods contribute to capacity costs, and capacity is *not* fully utilised during peak periods. The corollary is that, for an asset exhibiting economies of scale (and scope), amortisation payments are not necessarily zero in periods where there is some spare capacity. Nevertheless, Baumol (1971) was also writing prior to Panzar’s result when he commented on “the irrationality of a depreciation policy that demands the same contribution toward the cost of an asset in periods of heavy and of light usage”. Consequently, like Littlechild, Baumol concluded the following.

During any years in which there is unused capacity, the long-run marginal cost of the firm’s output should cover only operating costs (i.e., in such a period, it is equal to short-run marginal cost) and includes absolutely no contribution towards depreciation (Baumol, 1971).

7.2.2 *Economic Depreciation in Contestability Theory*

Given that contestability theory did away with the notion of pricing strictly on the basis of marginal cost (§2.1.7), any link between depreciation and long run marginal cost was not mentioned later by Baumol in the standard contestability text (i.e., BPW). Moreover, in that text, Baumol and his colleagues used the term “economic depreciation” in Hotelling’s original sense, meaning the change in economic asset value over a particular period, rather than in relation to the total amortisation payment.⁸ Apart from requiring a return on the economic asset value, the firm will also require an additional payment from consumers so that some provision is set aside to cover the future replacement cost of the asset every N years, where N is the asset lifetime. As Baumol and his colleagues explained, this provision to cope with finite asset lifetime is the accounting concept of depreciation (D).⁹ Payments to capital will thus comprise a rate of return component and a depreciation component as follows (BPW, 13G3).

$$PC_i = rV_i + D_i \tag{7.2}$$

where: PC_i are the payments to capital determined at the end of the i th year, comprising: (i) a return, at a rate r on the economic asset value, assessed at the beginning of each year; plus (ii) D_i , the depreciation provision set aside at the end of each year.

⁸ “Amortisation” is sometimes used in the sense of “annualised” (i.e., constant) payments to capital. Turvey (1969) used the term in its more general sense. Similarly, “economic depreciation” is sometimes used to refer to a depreciation schedule which relates to a *constant* stream of payments to capital. However, for constant payments to capital to be efficient requires that consumer demand and preferences remain constant with time (§7.2.3).

⁹ Hence, even if asset capacity—in other words, its service potential—remains constant over its finite lifetime (i.e., the asset is non-deteriorating during its lifetime), depreciation is still an essential concept. The asset effectively deteriorates only once by suddenly becoming valueless at the end of the N th year.

After defining economic asset value (in use) and payments to capital, as in (7.1) and (7.2) above, BPW (p. 384) then proceed to note that “economists generally recognise that accounting depreciation rules often bear little relationship to the underlying economic relationships”, but that “what is less clear is a set of depreciation rules that *is* consistent with economic theory”. As Hay and Morris (1993, p. 429) indicate, accounting depreciation provisions are typically based on the historic cost (i.e., the original purchase price) of an asset and are set in order to total up to the *historic* cost figure over the estimated lifetime of the asset. In contrast, BPW (p. 385) link their definition of depreciation directly to their definition of economic asset value as follows, for $1 \leq i \leq X$ (BPW, 13G6 and 13G7).

$$V_{i+1} = V_i + I_{i+1} - D_i \quad (7.3)$$

$$\text{thus } \sum_{j=1}^X D_j = V_1 - V_{X+1} + \sum_{j=2}^{X+1} I_j \quad (7.4)$$

The economic asset value is brought up to date in each period (i.e. year) by adding investment costs and subtracting depreciation. Hence, there is a direct relationship between a firm’s valuation, depreciation and pricing decisions. As BPW (p. 386) state: “if the firm decides on a set of outputs, input purchases, and prices over time, it automatically decides, implicitly, on the streams of payments to capital and depreciation charges. ... For any stream of payments to capital there is an equivalent stream of depreciation charges, and vice versa, where the equivalence is based on the relationships” in (7.1), (7.2) and (7.3). Where there is zero economic profit, $V_1 = I_1 = K$, where K is the initial capital outlay, and where the firm salvages the asset in the $X+1$ th year, the asset effectively no longer exists—thus $V_{X+1} = 0$. Hence, from (7.4) (i.e., BPW, 13G8):

$$\sum_{j=1}^X D_j = V_1 + \sum_{j=2}^{X+1} I_j = \sum_{j=1}^{X+1} I_j = K - S_{X+1} \quad (7.5)$$

Therefore, under conditions of zero economic profit, not only does the discounted sum of a set of payments to capital equal the discounted sum of all investment costs (including the initial outlay), as discussed above (§7.1.3), but the undiscounted sum of the corresponding set of depreciation payments is equal to the undiscounted sum of all investment (and disinvestment) costs. Equations (7.2)-(7.5) then become consistent with standard *accounting* practice. Baumol and his colleagues (BPW, pp. 385-386) suggest that this is particularly significant for *economic* analysis because these expressions are derived from the economic asset value given in (7.1).¹⁰

¹⁰ Note that the expressions (7.1)-(7.5) from BPW (p. 385) are general, and even hold in the presence of inflation and changes in technology. However, BPW implicitly assume zero economic profit in all cases, since in their formulation, the r in (7.2) is used as the discount rate (d) in (7.1).

Nevertheless, under conditions of zero economic profit, there are actually an infinite number of different streams of depreciation provisions which cover the cost of capital equipment over its lifetime (BPW, p. 387), and which satisfy equations (7.1)-(7.5). Schmalensee (1989) refers to this result relating to depreciation schedules as the “invariance principle”, (a result that requires the firm to be guaranteed zero economic profit in the presence of no technological change or competition). Hence, the common nominal straight-line depreciation approach, whereby the annual depreciation provision remains constant, can satisfy these equations.¹¹

Baumol and his colleagues (BPW, p. 394) suggest that simultaneous determination of an “optimal” depreciation stream and the corresponding payments to capital requires some overall objective function for the investment (or firm) to be specified, including the cost and demand functions relating to the investment. For instance, an optimal set of intertemporal prices could be derived subject to a profit maximisation objective under various regulatory and competitive constraints (from the point of view of the firm in question), or subject to a welfare maximisation objective (from society’s perspective), a viewpoint to which Turvey (1969) also subscribed (§4.1.4). Based on the underlying cost and demand functions, different optimal depreciation streams would correspond to the different optimal price schedules arising from either case.

Notwithstanding this approach for deriving economic depreciation from economic asset value, a rather curious recommendation from Baumol’s (1971) earlier work creeps into the standard contestability text. This is particularly surprising given that one of Baumol’s co-authors is Panzar, and that Baumol had already acknowledged the similarities between the peak load pricing problem and optimal depreciation schedules. In advising how contestability theory can be applied in practice, Baumol and his colleagues draw up a checklist for comparing real markets to the perfectly contestable ideal. The final item is the following statement, which appears to be a holdover from the earlier work based on linearly homogeneous production functions.¹²

Intertemporal price patterns must satisfy the rules of economic depreciation, with no contribution toward recoupment of fixed or sunk costs in periods of excess capacity (BPW, p. 471).

¹¹ Assuming that zero economic profit holds, and thus the entire stream of undiscounted depreciation payments equals the undiscounted sum of all investment costs, increasing (or decreasing) the depreciation provision in a particular year, does not make (or reduce) a firm’s supply of funds in that year. Unless tax or dividends are changed, any increase in depreciation will be exactly offset by a fall in retained earnings (Hay and Morris, 1993, p. 379).

¹² While Baumol (1971) allowed the production process as a whole to exhibit economies or diseconomies of scale, capacity costs were assumed to be linearly homogeneous. In discussing what would happen if this assumption were relaxed, Baumol proposed using the Ramsey pricing approach to adjust the required payments to capital (§6.3.5).

7.2.3 Economic Depreciation under Constant Payments to Capital

Rogerson (1992) explains that “economists have long recognized the fact that the goal of fairly reimbursing the firm for its asset purchases can be accomplished using any depreciation schedule so long as a fair rate of return is paid on the remaining book value”. Given this freedom—as evidenced by Schmalensee’s invariance principle—Rogerson explains that the role of depreciation is often seen as “optimally smoothing consumer expenditures over time”, as do Bertram *et al.* (1992).

The conventional role of a depreciation charge is to allow a business to retain sufficient cash to maintain the owner’s real capital intact, in terms of being able to acquire equivalent replacement assets as previously acquired assets wear out. In theory, continual ploughing-back of depreciation allowances into replacement investment should allow the business to operate in perpetuity, neither losing nor accumulating cash. In practice, accounting conventions commonly do not ensure that enough money is accumulated in reserves to allow all assets to be replaced, and sustainability in the strict sense is not guaranteed in this way. In addition, the time profile of actual replacement investment usually bears no resemblance to the steady rate of accumulation of depreciation reserves (at least within the period represented by the lifetime of the assets). ... The aim of depreciation conventions is to smooth out through time the acquisition of the funds for replacement investment, in order to avoid abrupt shocks to profitability and price when major items of equipment wear out (Bertram *et al.* 1992, pp. 91-92).

As discussed earlier (§6.4.3), Rogerson (1992) rejected the notion that just any depreciation schedule would be economically acceptable, and concluded that from a welfare basis the optimal depreciation schedule is the “real constant depreciation schedule”—namely, one derived from *constant* payments to capital (in real terms). Although not derived with the objective of smoothing consumer expenditures, Rogerson’s result does in fact have that characteristic. The price path is smooth, although the actual depreciation provision component of each real constant payment to capital does not remain constant, but increases over time.

Where there is no inflation, no change in technology, and the asset deteriorates but one time (entirely at the end of its lifetime), it is straightforward to derive the corresponding equations for economic asset valuation and depreciation over the lifetime (N) of a single asset with initial cost K . In the simplest case, where demand ceases at the end of the N th year, and zero economic profit requires that $r = d$, the constant payments to capital (\overline{PC}) are related to the initial capital outlay as shown in (7.6a), derived from (7.1). The term B , defined earlier in (6.1), is the uniform present worth factor. Consequently, the payments to capital in each period (i.e., each year) are as shown in (7.6b).

$$K = \sum_{j=1}^N \overline{PC} \left(\frac{1}{1+d} \right)^j = \overline{PC} \sum_{j=1}^N \left(\frac{1}{1+d} \right)^j = \overline{PC} \frac{1 - \left(\frac{1}{1+d} \right)^N}{d} \quad (7.6a)$$

$$\Rightarrow \overline{\text{PC}} = \frac{K}{B_N} \quad (7.6b)$$

However, as Bertram *et al.* (1992, p. 145) explain, when an asset has a finite lifetime, commitment to providing an *ongoing* service of *constant* service potential based on the utilisation of that asset, involves not only the initial cost of the asset, but the asset's replacement costs indefinitely into the future (§6.2.2). If ongoing network service is to be provided in perpetuity, then at the beginning of every $kN+1$ th year, for $k \geq 0$, a capital outlay will need to be made, such that $I_{kN+1} = K$. In all other years, the reinvestment costs (I_i) are zero, but depreciation is accumulated to allow the reinvestments of K in every $kN+1$ th year, for $k > 0$. Assuming zero economic profit, the required discounted sum of payments to capital, in the year of initial investment, can be found from the discounted sum of investment costs as follows (e.g., Bertram *et al.*, 1992, p. 145) in (7.7a). The term Γ , defined in (7.7b), is the present worth factor for an infinite series comprising uniform terms at periodic intervals T . The constant payments to capital in this 'perpetual demand' case (7.7c) are the *same* as in the 'finite demand' case in (7.6b).

$$\sum_{j=1}^{\infty} \text{PC}_j \left(\frac{1}{1+d} \right)^j = \sum_{j=1}^{\infty} I_j \left(\frac{1}{1+d} \right)^{j-1} = \overline{\text{PC}} \sum_{j=1}^{\infty} \left(\frac{1}{1+d} \right)^j = \sum_{k=0}^{\infty} I_{kN+1} \left(\frac{1}{1+d} \right)^{kN} = K(1+\Gamma_1) = \frac{R}{dB_N} \quad (7.7a)$$

$$\Gamma_i \equiv \sum_{k=1}^{\infty} \left(\frac{1}{1+d} \right)^{kT-i+1} = \frac{1/d - B_{T-i+1}}{B_T}; \text{ for } 1 \leq i \leq T \quad (7.7b)$$

$$\Rightarrow \overline{\text{PC}} = \frac{1}{\sum_{j=1}^{\infty} \left(\frac{1}{1+d} \right)^j} \cdot \frac{K}{dB_N} = \frac{K}{B_{\infty} dB_N} = \frac{dK}{dB_N} = \frac{K}{B_N} \quad (7.7c)$$

Substituting these constant values for the payments to capital—from either (7.6b) or (7.7c)—into (7.2) and (7.3) provides the expressions (7.8) and (7.9) for economic asset value and economic depreciation respectively, in each year of the asset's lifetime. In words, the economic asset value is equal to the initial asset cost, multiplied by the ratio of: the uniform series present worth factor, for a period encompassing the years to the next asset replacement; to, the uniform series present worth factor for the entire period of the asset's lifetime.¹³

$$V_i = \sum_{j=i}^{\infty} \overline{\text{PC}} \left(\frac{1}{1+d} \right)^{j-i+1} - \sum_{k=1}^{\infty} I_{kN+1} \left(\frac{1}{1+d} \right)^{kN-i+1} = \frac{K}{dB_N} - K\Gamma_i = K \frac{B_{N-i+1}}{B_N}; \text{ for } 1 \leq i \leq N \quad (7.8)$$

$$D_i = \overline{\text{PC}} - dV_i = \frac{K}{B_N} (1 - dB_{N-i+1}) = \frac{K}{B_N (1+d)^{N-i+1}}; \text{ for } 1 \leq i \leq N \quad (7.9)$$

¹³ Equations (7.8) and (7.9) are also associated with the expression "single vintage limit" (e.g., Bertram *et al.*, 1992, p. 146). This relates to the economic value of an asset which has to be replaced at the end of its lifetime in its entirety.

The annual payments to capital can be viewed as the *long run marginal (or incremental) cost* of capacity (§7.2.1), while the annual depreciation payments can be viewed as the *option value* of delaying investment (§7.1.1). Although it was initially assumed that payments to capital are constant, this is not in fact required if it is considered that payments to capital should equal long run incremental costs. This is because the long run incremental costs are themselves constant, as can be shown from rearranging expression (7.7a).

7.3 Asset Value and Intertemporal Subsidy-Free Prices in Theory and Practice

7.3.1 Deriving Economic Asset Value from Intertemporal Anonymously Equitable Prices

Bertram *et al.* (1992, p. 100), argue that, because there are an infinite number of depreciation methodologies which still allow the full recovery of the cost of capital equipment (§7.2.2), owners *and* users of network assets should be largely indifferent between alternative valuation procedures when the choice is viewed at the time of installation of the assets, and there is a guarantee that the rule chosen will be consistently applied. Depending on the firm's objective function, this may be correct (§9.2.2). However, *consumers* will not be indifferent to the choice of valuation methodology, unless they are associated with those assets over their *entire* lifetime.

The same result as that above (§7.2.3) can also be derived from an intertemporal cross subsidy perspective. Consider two consecutive groups of consumers with identical characteristics demanding a service which requires the input of some capacity costing K .¹⁴ Both groups of consumers have the same aggregated demand for capacity q , and the second group of consumers begins consumption after a "period 1" of $T < N$ (years) and this demand lasts indefinitely, for "period 2". The total revenue/cost equality relationship is thus as shown in (7.10a), where ρ is the discount factor (e.g., BPW, p. 409) for any period of X years as defined in (6.3).

$$P_1q + \frac{P_2q}{1 + \rho_T} = \frac{K}{dB_N}; \text{ for } 1 \leq T < N \quad (7.10a)$$

$$\text{where: } \rho_X \equiv (1 + d)^X - 1 \Rightarrow X = \log_{(1+d)}(1 + \rho_X) \quad (7.10b)$$

If capacity costs for the incumbent firm, potential entrants, and consumers are all statically symmetric, then the intertemporal cross subsidy problem can be formulated either as a competitive entry or self-production problem. The stand alone cost of self-production for the first consumer group might appear to be simply K . In other words, the imputed value of self-production (i.e., P_1q) should be less than or equal to the entire cost of constructing the required capacity a single time. (It only needs to be

¹⁴ So that the subsidy-free and anonymously equitable prices are equivalent, declining average incremental costs of capacity will be assumed (§4.3.4).

constructed once, and not subsequently replaced, because $T < N$). However, this ignores the fact that at the end of T years, the original asset has remaining service potential, and thus may have some value at the end of the first period (S_1) to the *subsequent* group of consumers in the second period. The possibility that the asset can be resold at the end of T years results in a *net intertemporal stand alone cost* expression for the first group of consumers, as is shown in (7.11).¹⁵ (Note that the subscripts on P and S relate to a period and the end of that period respectively, whereas the subscripts on other terms relate to the year).

$$P_1q \leq K - \frac{S_1}{1 + \rho_T}; \text{ for } 1 \leq T < N \quad (7.11)$$

It is important to note that this *net intertemporal stand alone cost* equation does *not* require an assumption that the capacity is *perfectly fungible* (§3.5.1). The capacity may in fact have no *alternative* use and be entirely irreversible. However, the capacity still has value in its *existing* use, but to a different (later) set of consumers who demand a service based on the utilisation of that capacity. (If the formulation is of a network, the capacity could relate to any immobile and durable item of network equipment for which demand is location specific). The cost of self-production for the second group of consumers is simply the cost of providing the asset in perpetuity, and equals the original total cost. Hence, the (net or gross) intertemporal standalone cost equation for the second group is as shown in (7.12).

$$P_2q \leq \frac{K}{dB_N} \quad (7.12)$$

This second group of consumers would be willing to pay for the asset constructed by the first group of consumers if the resale value at the end of period 1 (S_1) satisfies expression (7.13). This recognises that if the second group purchased the now second-hand asset, it would still need to replace that asset for the first time after another $N-T$ years. And for resale to be feasible—in other words, desirable from the perspective of the second group of consumers—the combined resale cost, plus the discounted future stream of future asset replacements, would have to “dominate” the cost of constructing a new asset at the end of T years, and replacing it indefinitely into the future.¹⁶

$$P_2q \leq S_1 + \frac{K}{dB_N(1 + \rho_{N-T})} \leq \frac{K}{dB_N} \quad (7.13)$$

¹⁵ As discussed earlier (§4.3.6 and §6.1.2), the concept of “net IC” (and, by complementarity, “net SAC”) already exists, and relates to the inclusion of cross-elasticities of demand in the *static* SAC and IC bounds. Here the terms ‘net *intertemporal* SAC’ and ‘net *intertemporal* IC’ are introduced to distinguish these concepts from the earlier terms. Henceforth, however, and particularly in Chapters VIII and IX, these intertemporal concepts will simply be referred to as net IC (NIC) and net SAC (NSAC), with the intertemporal nature of the bounds taken as a given.

¹⁶ “Dominate” is meant in the same sense which BPW (p. 194) use the term.

With some algebraic manipulation, it can be seen that the resale value S_1 satisfies (7.14). Expressing the resale value in this form demonstrates that the asset's resale value is equivalent to the economic asset value (7.8) for constant demand in the face of no changes of technology and inflation, except that the resale expression is an inequality. However, if the second group of consumers are willing to pay the total cost exhibited by the RHS of (7.13), then they will be indifferent to paying the amount that maximises the LHS of (7.14), and it is in the interests of the first set of consumers to push up the resale price to the maximum possible value, in order to minimise their own net stand alone cost (NSAC) of supply. Given the useful identity in (7.15) which is derived from (7.6b) and (7.10b), the resale expression (7.14) can also be rearranged as (7.16).

$$S_1 \leq \frac{B_{N-T}}{B_N} K \quad (7.14)$$

$$B_X = \frac{1}{d} \left(\frac{\rho_X}{1 + \rho_X} \right) \quad (7.15)$$

$$S_1 \equiv \frac{B_{N-T}}{B_N} K = \frac{1}{dB_N} \left(1 - \frac{1 + \rho_T}{1 + \rho_N} \right) K = \left(\frac{\rho_N - \rho_T}{\rho_N} \right) K \quad (7.16)$$

Expressions (7.10)-(7.14) describe the set of SAC constraints on subsidy-free revenues for the two groups of consumers. The equivalent complementary IC expressions are not presented, but can be easily derived by subtracting (7.11) and (7.13) in turn from (7.10a). Combining all these expressions together, the bounds on subsidy-free revenues are as presented in (7.17) and (7.18), and are in fact equalities rather than inequalities.

$$P_1 q = \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K ; \text{ for } 1 \leq T < N \quad (7.17)$$

$$P_2 q = \frac{1}{dB_N} K \quad (7.18)$$

However, for the bounds to satisfactorily ensure that the revenues are anonymously equitable for any and every possible combination of sequential consumer groups, then the constraint expressions would also have to relate to any and every possible sequence of consumers. This requires a number of sets of infinite constraint expressions to be developed, in a similar manner to Faulhaber and Levinson's (1981) static conventional peak load pricing problem, presented earlier (§4.3.3).

To derive the anonymously equitable revenues, period 1—which is of some arbitrary length less than N years—is entirely divided into two sub-periods termed 1α and 1β , of variable period length α and β years respectively (where $\alpha + \beta \equiv T < N$, $\alpha > 0$, and $\beta \geq 0$), for every possible combination of self-producers (or entrants) during the entirety of period 1. Since the subsidy-free revenue constraints are

now equalities, no total revenue/cost equation is required, as equation (7.10a) is automatically satisfied by (7.17) and (7.18). The infinite set of subsidy-free revenue constraints are shown in equation (7.19), as well as (7.18) above.

$$P_{1\alpha}q \equiv P_{\alpha}q + \frac{P_{1\beta}q}{1 + \rho_{\alpha}} = \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K; \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (7.19)$$

Additional constraints are now introduced by allowing self production (or entry) at any time before the original asset's lifetime is over. During period 1, constraints relating to the stand alone costs of self producers (or entrants) are described by expression (7.20a). The resale value at the end of period 1, to consumers in period 2, of the asset which consumers could construct at some point during period 1, is given in (7.20b).

$$P_{1\beta}q \leq K - \frac{S_1}{1 + \rho_{\beta}}; \quad \forall \beta: 0 \leq \beta \leq T \quad (7.20a)$$

$$\text{where: } S_1 \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_{\beta}}{1 + \rho_N} \right) K \quad (7.20b)$$

However, this infinite set of constraints is not the only possible constraining influence on prices in period 1. In parallel with these constraints, which model self-production (or entry) during period 1 with asset resale occurring at the end of period 1, is a set of constraints representing the consumer group whose demand commences at the beginning of period 1, but sells up these assets before the end of period 1. The resultant set of constraints is shown in (7.21a) with the corresponding resale equation in (7.21b).

$$P_{1\alpha}q \leq K - \frac{S_{1\alpha}}{1 + \rho_{\alpha}}; \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (7.21a)$$

$$\text{where: } S_{1\alpha} \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_{\alpha}}{1 + \rho_N} \right) K = \left(\frac{\rho_N - \rho_{\alpha}}{\rho_N} \right) K \quad (7.21b)$$

Expressions (7.20a) through (7.21b), when linked with (7.19), successfully protect the interests of all possible configurations of consumers anonymously coming and going throughout period 1.¹⁷ Combining all these constraints results in the pair of sets of equalities in (9.6a) and (9.6b) which fully describe the conditions constraining period 1 prices (where $\alpha + \beta \equiv T$, $\alpha > 0$, and $\beta \geq 0$).

¹⁷ It is assumed that, apart from the length of time for which particular consumers demand capacity, their characteristics are identical.

$$P_{1\beta}q = \frac{1}{dB_N} \left(\frac{\rho_\beta}{1 + \rho_\beta} \right) K; \quad \forall \beta: 0 \leq \beta \leq T \quad (7.22a)$$

$$P_{1\alpha}q = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right); \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (7.22b)$$

Transforming these equations into (per unit) payments to capital (p_t) provides (7.23), since a price can be considered to be a per unit revenue for a sub-period of arbitrary length i .

$$P_{1\alpha}q = \int_{i=0}^{\alpha} p_t q \left(\frac{1}{1+d} \right)^i = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right) K; \quad \forall \alpha: 0 \leq \alpha \leq T \quad (7.23)$$

Because the equality in (7.23) holds for any sub-period length between 0 and T years, it is clear that the price for any sub-period of the *same* length must be the same. A more familiar way of looking at this subsidy-free price is to transform it into annual payments to capital (i.e., annual prices). Over the course of period 1, annual payments to capital p_t , made at the end of the t th year, relate to the total revenue collected at the beginning of period 1, as shown in (7.24a). Rearranging provides the annual payment to capital defined below in (7.24a) and (7.24b) as p_t , which from (7.23) clearly must remain constant throughout all of period 1.

$$\frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K = P_1q = \sum_{t=1}^T p_t q \left(\frac{1}{1+d} \right)^t = B_T p_t q = \frac{1}{d} \left(\frac{\rho_T}{1 + \rho_T} \right) p_t q \quad (7.24a)$$

$$p_t = \frac{K}{qB_N}; \quad 0 \leq t \leq T \quad (7.24b)$$

This outcome shows that, of all the infinite sets of price paths which can satisfy the subsidy-free revenue constraint for period 1 given in (7.19), only a *single* price path provides a set of subsidy-free and anonymously equitable prices, and that this is the same constant payment to capital price path as found by using Rogerson's (1992) Ramsey pricing approach (§7.2.3), given the same underlying assumptions. Consequently, as is implied by the resale equation (7.16), the economic asset value derived from the intertemporal subsidy-free pricing approach is the same as that under the Ramsey pricing approach. Moreover, this result also has important implications for the application of constrained market pricing in practice, since it implies that the concerns regarding great pricing freedom (§5.1.3) may be rendered null and void as long as intertemporal factors are appropriately taken into account.

7.3.2 Firm Value versus Asset Value: the Role of Accumulated Depreciation

Small and Ergas (1999) explain how reductions in either or both of total willingness to pay for the services of the asset, and the total service life of the asset reduce the value of the firm holding the asset. However, if willingness-to-pay remains constant, and service life remains the same, the fact that the economic value of an *asset* held by a firm depreciates with time does not necessarily imply that the

firm's value in its entirety devalues by the same amount. If the depreciation provisions are set aside, then at the end of an original asset's lifetime the accumulated depreciation will be equal to the initial capital outlay, as is seen from (7.5). Therefore, in the absence of inflation or changes of technology, the investment is self-sustaining, since the accumulated depreciation can be used to purchase a new (replacement) asset with the same service potential (i.e. capacity). And although the asset's value decreases over its lifetime, the entire value of the investment as a whole, or the value of the incumbent firm's entire operations, remains constant. This is because the value of the firm as a whole comprises not only the (decreasing) value of the physical asset, but also the (increasing) value of accumulated depreciation, as show in (7.25).

The overall economic investment value (EIV) of a service provided through the construction of an asset, for any year between the current year and the X th year (for $X \leq N$), is: (i) the present value of the returns on the economic asset value (V); plus (ii) the present value of the returns on accumulated depreciation; plus (iii) the discounted future value of the investment/firm/operation *as a whole*; as follows, for $1 \leq i \leq X \leq N$.

$$\text{EIV}_i = \sum_{j=i}^X rV_j \left(\frac{1}{1+d} \right)^{j-i+1} + \sum_{j=i+1}^X \gamma \left[\sum_{k=1}^{j-1} D_k \right] \left(\frac{1}{1+d} \right)^{j-i+1} + \left[S_{X+1} + \sum_{k=1}^X D_k \right] \left(\frac{1}{1+d} \right)^{X-i+1} \quad (7.25)$$

where: γ is the rate of return applied to the accumulated depreciation. The second term of the RHS of (7.25) is derived presupposing that accumulated depreciation is not only set aside to be used for asset replacement in the $N+1$ th year, but that a return can be made on this quantity in the meantime. The final discounted term is the value of the investment in the asset salvage (or resale) year, which comprises not only the physical asset's salvage value but also the accumulated depreciation provision. From (7.3) and (7.5), assuming zero economic profit, no inflation and no change in technology, the following equality holds for $1 \leq i \leq X$, irrespective of the depreciation method applied (although the corresponding values for V_i and D_k will of course differ in accordance with which method is used).

$$K = V_j + \sum_{k=1}^{j-1} D_k \quad (7.26)$$

Therefore, for an asset subject to zero economic profit, the overall investment/firm value given in (7.25) remains the same as the initial capital outlay K in every year, as long as: (i) the return obtainable on accumulated depreciation is the same as that received relating to the physical asset (i.e., $\gamma = r = d$); (ii) any alternative use/activity for which the accumulated depreciation provision is temporarily utilised is perfectly fungible; and (iii) the salvage value of the asset is "economic" (i.e., equals its equivalent economic asset value). Although the physical asset's value reduces with time due to the burgeoning replacement requirement, the investment's value as a whole remains intact, thanks to the accumulated depreciation provision. Depreciation is of course not necessarily accumulated as cash reserves; any real

firm is unlikely to ever maintain cash reserves equivalent to accumulated depreciation (§7.2.3). Along with retained earnings, depreciation provides an inexpensive internal source of funds compared to external debt financing (e.g., Hay and Morris, 1993, p. 428). Such funds can be used to expand the firm's activities and to finance capital works. However, as Sidak and Spulber (1997, p. 319) point out, in a competitive market, price always includes compensation for economic costs such as the interest forgone by the firm when it uses internal funds instead of borrowing those funds from a bank. Hence, the opportunity cost of capital associated with internal funds is also positive.

The equality of economic investment value with asset replacement cost holds as long as not only the physical asset is assumed to be perfectly fungible, but any alternative use/activity (with a return satisfying $\gamma = r$), for which the accumulated depreciation is temporarily utilised, is also assumed to be perfectly fungible. This makes sense, because as Baumol (1971) explains (§7.1.1), an investment is only worthwhile to an investor if at the end of the investment period (X) the investor receives back (at least) the asset's initial purchasing power, plus appropriate compensation for the use of the investor's funds (i.e., the opportunity cost of capital). The first two terms of the RHS of (7.25) are the investor's compensation for use of funds, and the last term is the sum of economic asset value and accumulated depreciation at the time of salvage. From (7.26), under conditions of perfect fungibility, this last term is thus equivalent to the asset's initial cost K , and consequently (where there is no inflation or change of technology) is equivalent to the investor's initial purchasing power (or capital outlay).

7.3.3 *Optimal Real World Depreciation: Front-Loaded or Back-Loaded?*

Rogerson (1992) concluded that utilities would have an incentive to undercapitalise if they applied any depreciation schedule more accelerated than a “real constant depreciation schedule” (§7.2.3)—namely, the schedule presented in (7.9), but under the stricter assumptions of symmetric costs, no demand growth, no changes in technology, and no inflation. In practice, rather than following the (real) constant depreciation schedule which results in “*back-loaded*” capital cost recovery, Burness and Patrick (1992) indicate that depreciation schedules are typically “*front-loaded*”, meaning that recovery in any period cannot exceed that of any previous period (of the same length). Since typical front-loaded depreciation schedules—such as nominal or real straight line depreciation—are, by definition, “accelerated” relative to the back-loaded real constant depreciation schedule, Rogerson predicted that most depreciation methods generate incentives for undercapitalisation.

Burness and Patrick examine the optimal recovery of capital costs from two perspectives: firstly, that of a profit-maximising firm acting under a rate of return constraint; and secondly, that of a regulator attempting to maximise welfare subject to the regulated firm's revenues recovering its economic costs. Burness and Patrick find that for a regulated firm operating under a rate of return constraint, capital cost recovery either begins immediately, and continues at the most rapid rate (subject to the ROR constraint) until full recovery of cost occurs, or is zero for an initial period and then occurs at the maximum rate for

the remainder of the asset lifetime until full recovery occurs. The firm is concerned with perpetuating the rate base as long as possible when the rate of return is above its cost of capital, so as to maximise profits, or to depreciate the rate base as quickly as possible where the rate of return is lower than the cost of capital, in order to minimise losses. Although Burness and Patrick find that the socially optimal path of recovery differs from that of the firm, in either case, optimal recovery requires “back-loading”, as long as the firm can make a positive or zero economic profit. Burness and Patrick suggest that part of the reason why firms in practice appear to be averse to back-loading, and prefer to front-load their depreciation schedules could be due to regulatory uncertainty (§6.3.3).

Crew and Kleindorfer (1992) provide another explanation as to why firms prefer to front-load their actual depreciation schedules. They suggest that where a regulated firm faces either competitive entry or technological progress, then the firm will have incentives for front-loading cost recovery. In fact, under conditions of competition and technological progress, front-loading of capital is *essential* if the regulated firm is to remain viable.

If the introduction of accelerated capital recovery is delayed by regulators, they may effectively vitiate any opportunity of the firm to recover its invested capital. The breathing space, or period of time, that the regulators can delay introducing the application of efficient capital recovery without ultimately compromising the firm’s ability to recover its invested capital is called the “Window of Opportunity” (WOO). This same window of opportunity requires that the level of depreciation initially be set optimally. There are limited opportunities in the future, under technological change and competition, to rectify mistakes made now (Crew and Kleindorfer, 1992).

In particular, in the case of price-cap regulation (§6.4.1), Crew and Kleindorfer indicate that “if depreciation is set solely based upon the status quo, the initial price cap may be set at too low a level to allow full cost recovery”. Crew and Kleindorfer define a “WOOPS” point as that “time at which the window of opportunity is past”, the latest time at which the firm can still earn its cost of capital. In other words, if a firm is constrained first by regulatory controls on capital cost recovery, and then by competition, the firm may be unable to recover its initial capital outlay. One way of looking at this is that the regulatory restrictions on depreciation have made the firm’s intertemporal price path unsustainable (§3.3.3). Crew and Kleindorfer contrast their result with Schmalensee’s (1989) and suggest that one interpretation of their results is that the invariance principle must be reinterpreted to be a “constrained invariance principle” under conditions of competition and technological change. Thus, at least when the “WOOPS” is greater than zero, there still exist a large number of possible depreciation schedules that can ensure full recovery of capital given the firm’s cost of capital. Nevertheless, in any event, an accelerated depreciation schedule is required to ensure cost recovery.

Most interestingly, Turvey (1969)—although not discussing depreciation in the context of regulation—had also recognised the need for accelerated depreciation in the face of technological change, but had expressed this need in *marginal cost* terms. Turvey explained that the expectation of lower marginal costs in the *future*, due to technological progress, somewhat paradoxically raises *today's* marginal costs. Consequently, Turvey—like Baumol (1971)—provided an argument for accelerating depreciation today in the face of lower costs tomorrow. Rather than suggesting that disallowing a firm's "entitlement" to at least a zero economic profit is to unfairly "ignore its investment-backed expectations" (§5.3.2), Turvey indicates that "the rise in marginal cost will necessarily raise the amortisation on existing capacity of all vintages". If an asset's "economic lifetime is now shorter than it would have been had expectations not changed, amortisation in each year of its life must now be greater than under the old set of expectations". Apart from technological progress, a similar argument could be used in the face of a hypothetical optimal asset configuration in the face of changing *demand* patterns (§6.3.2).

The reason for this apparent paradox is that, as shown earlier, marginal cost is the present worth of the cost consequences of bringing forward the installation of new capacity and of postponing the scrapping of old capacity. The cost consequences of bringing forward the installation of new capacity are now less favourable than they were, since they now include the sacrifice of a greater cost improvement [once the new technology becomes available] than they did before. ... To achieve a net increment in capacity [before that new technology becomes available] will therefore require yet more postponement of scrapping and/or advancement of installation, both of which are more expensive than with the previously optimal policy. Hence, marginal cost is increased. The expectation that future available technology is going to be better than current technology to a greater extent than hitherto expected means that the cost of expanding capacity soon instead of later includes a greater sacrifice of future improved technology. Thus it is that an improved expectation of jam to-morrow can be said to raise the cost of bread today. It certainly makes sense in resource allocation terms, since to set price equal to a marginal cost which rises because of newly anticipated technical progress will discourage demand, so postponing the expansion of capacity until advantage can be taken of increased superior techniques (Turvey, 1969).

While the expectation of lower costs in the future justifies accelerated depreciation, Turvey was not advocating that a firm necessarily be allowed to fully recover its costs. He pointed out that the total lifetime amortisation will most likely still be lower than previously expected, leading to some "obsolescence". Turvey notes that allowing the firm to recover such obsolescence—which could be viewed as "stranded cost" (§3.5.1)—is *not* consistent with marginal cost pricing. Turvey also addressed the issue of "whether amortisation can be said to be based on historic cost or replacement cost", and made it clear that the answer is *neither*. Like Baumol (1971), Turvey pointed out that is no reason why, in the face of unforeseen changes in prices or technology, lifetime depreciation should necessarily sum up to an asset's initial acquisition cost or replacement cost. Turvey's later collaborative work (i.e., Turvey and Anderson, 1977), where he requires that price incentives should be directed at

tomorrow's consumers (§4.1.3), has perhaps been taken to be an indication of Turvey's advocacy of replacement-cost pricing. However, Turvey's (1969) original paper on marginal cost made his position clear.

This is not to deny that a change in the price of new assets will change marginal cost, amortisation and replacement cost in the same direction. It will, of course, do so. The point is simply that to set prices equal to marginal costs is not the same thing as earning a revenue sufficient to cover replacement-cost depreciation. Except by chance, these are alternative and mutually incompatible pricing principles. Marginal-cost pricing, to put it crudely, means paying for assets when they are being used, not before they are acquired (Turvey, 1969).

This conclusion contrasts with that of Schramm (1991) who—in his critique of strict LRMC pricing (§4.1.5)—wrote that, where “already existing levels of service will be maintained indefinitely”, then “replacement costs akin to depreciation, but based on future, rather than historic costs, will always be charged against current users, so that, at the time when replacements are needed, the required funds will be available to provide these replacements at no additional costs to future users”. This appears to imply the accumulation of capital for future investments *in advance*. Such “pre-financing” of future investment, though, is different from basing *current* prices—comprising return on capital, plus the depreciation provision—on *current* replacement costs. For instance, Sidak and Spulber (1997, p. 318) state that “the regulated firm's cash flows must be large enough for the firm to replace its capital over time. Thus, a regulated firm's rates of return and depreciation are adjusted so that its cashflows approximately equal those that would result from the use of replacement costs rather than book costs for its invested capital”. This neither implies pre-financing, nor that accumulated depreciation will exactly equal asset replacement cost. However, it does imply a rejection of historic-based calculation of depreciation payments.

7.3.4 Opportunity Cost-Pricing of Specialised Assets: Alternative Uses versus Alternative Users

Zijl and Irwin (2001) provide an extreme example of front-loaded depreciation in discussing “opportunity cost” pricing.

Even the opportunity-cost approach will equate the expected present value of costs and revenues. ... When the assets are sunk, the expected downward revaluation in the period after investment is equal to the entire value of the investment. Thus the revenue-setting equation for opportunity-cost pricing would require users to pay for the entire cost of a sunk investment in the year in which the investment was made (Zijl and Irwin, 2001).

Zijl and Irwin appear to accept the position that where an investment has no alternative use, then the opportunity costs of the investment are zero. They would imply that the stand alone cost faced by current consumers is the entire stand alone cost of the “sunk” investment in its entirety. However, there may be a substantial mismatch between the length of the period over which current consumers demand

the associated service, and the lifetime of the investment. If current consumers have good reason to believe that demand for the service will still remain even after they require the service no longer, then—rationally—they will not be willing to pay the entire cost of the asset as the asset utilisation price. They will only be willing to pay the net intertemporal stand alone cost (§7.3.1). If *ownership* of the asset were to be transferred to themselves, then the consumers might be willing to pay that entire cost of the asset, because they could then lease or resell the asset to future consumers. Consequently, where demand for a service requiring the utilisation of a specialised asset exists, that asset can be considered to always have an alternative use; use by consumers who themselves own the asset in question.

Williamson (1986b) appears to be one of the few to distinguish between alternative *uses* of assets, and alternative *users* (§3.5.1). The opportunity cost of capital is usually determined from its best alternative use. However, this does not necessarily mean that the opportunity cost of capital is zero if there is no alternative use, because alternative users may exist who are willing-to-pay for the assets in their *current* use. Value is effectively the intersection of marginal opportunity cost and demand. If there is no demand, then regardless that a firm has incurred opportunity costs, it would be inefficient to compensate that firm for its now valueless and stranded investment. However, if demand still exists for an asset, even if it has no value in some alternative use, it cannot be said that the value of the asset is zero. A competitive market price—which Sidak and Spulber (1997) consider to be the appropriate benchmark for opportunity cost (§6.3.1)—still exists even for specialised assets, as long the willingness-to-pay of current and future consumers exceeds total costs.

For instance, firms that are undervalued in the sharemarket are subject to takeover; which results in a change of *ownership*, rather than a change of use. Even if no other firms exist in the market, the users of the assets themselves could always take over the existing firm which provides them with goods or services based on the utilisation of those assets. Of course, the complexity and level of transaction costs associated with such a move would be likely to increase with the number of users. Nevertheless, some parts of New Zealand’s electricity supply industry are already owned by their users, namely EMCO (§2.3.4) and trust-owned ELBs (§2.3.2). The transaction costs in the latter are minimised by having consumers vote for members of a trust board who appoint a board of directors to manage the company. Evans and Quigley (1998) suggest that, in general, “joint ventures provide the efficient approach to vertical integration in the presence of natural monopoly and downstream competition on product cost and variety”. And in later collaborative work (i.e., Evans *et al.*, 2000), they suggest that the pre-reform arrangement of having consumers match owners in New Zealand’s ESAs “went a considerable way to improving allocation over that of a local investor-owned monopoly”.¹⁸ In fact, one of the Baumol group

¹⁸ On the other hand, Evans *et al.* (2000) highlight that there are certain institutional inefficiencies inherent in such an arrangement, and that, at least for pre-reform ESAs, dynamic inefficiencies arose because the opportunity cost of capital was not fully accounted for in prices (§2.2.3 and §6.3.1).

implied that such a common ownership structure might make such markets more contestable: “One way to avoid the exercise of monopoly power is to have the sunk costs borne by a government or municipality, ... or by mandating that sunk costs be shared by a consortium ... rather than to have the sunk costs incurred by the firm that is supplying the services” (Bailey, 1981).

7.4 Optimised Deprival Valuation (ODV) Methodology in New Zealand

7.4.1 *The Roots of New Zealand’s Deprival Valuation Methodology*

Efficient pricing is one of the key tenets of New Zealand’s power sector reform process (§2.3.1). However, assessing whether distribution line charges are efficient or not has been plagued by the circularity problem described above (§7.1.2). Apart from the use of the optimised deprival valuation (ODV) methodology by the Ministry of Economic Development (and previously the Ministry of Commerce) as a benchmark to determine the exercise of monopoly power (§2.4.3), the majority of electricity line businesses (ELBs) have actually been setting their line charges on the basis of their ODV. Consequently, the methodology involved in determining the ODV warrants closer examination. For instance, the Inquiry into the Electricity Industry (Ministry of Commerce, 2000) estimated that for a typical ELB, 82% of all costs were driven by recovery of their ODV, and receiving a rate of return based on those valuations. This “significance of line asset values for the price consumers pay for their power”, led the Government to request the Commerce Commission to undertake a review of the ODV methodology (Hodgson, 2000b)—(a review that is not yet completed; §1.3.1).

Notwithstanding recent concerns about its fairness (§2.4.6), the general principles of the ODV benchmarked approach have been in place since 1994. Before selecting the ODV as the valuation method to be applied, the New Zealand Ministry of Commerce assessed the suitability of a range of valuation methodologies for this regulatory purpose, by evaluating them against criteria of: (i) ability to reveal monopoly pricing behaviour; (ii) consistency in providing benchmarked comparisons of value; and (iii) low compliance cost (Ernst and Young, 1994, p. 1). However, the Ministry also expressed that, in a more general business context, an ELB’s valuation should also contribute to: (a) business sustainability (i.e., allow the firm to set prices at a level sufficient to allow it to maintain its operations); (b) a reasonable return, (commensurate with perceived risk and sufficiently attractive for new investment where business expansion is warranted); and (c) efficiency (i.e., the firm should have incentives to utilise the lowest cost asset base and the most efficient configuration of assets); (Ernst and Young, 1994, pp. 4-5).

As discussed earlier (§7.1.2), Hay and Morris (1993, pp. 429-432) discuss that within the accounting profession there are many possible approaches to assessing the value of an asset, including: (i) [depreciated] “historic cost”; (ii) “value in purchase” (i.e., the asset’s [depreciated] replacement cost, DRC); (iii) “value in use” (i.e., the “economic” net present value (EV) of future cash flows obtainable from the use of the asset); and (iv) “value in sale” (i.e., “net realisable value”), NRV. For the same asset,

the magnitudes of these values will not necessarily be the same as each other. (In addition, if the firm is publicly listed, its “stock market valuation” will provide yet another assessment of value).

Hay and Morris explain that, historically, the various approaches mainly arose from differences in agreement within the accounting profession on how to appropriately account for the impact of inflation on company value. Particularly under historic cost valuation, companies may be showing apparently healthy rates of return but, during periods of high inflation, be unable to maintain their level of operations and thus be steadily going into insolvency. The general view is that inflation should cause asset value to “*appreciate*”. Hay and Morris discuss the various attempts to grapple with this problem in the UK, which was motivated by the high-inflation environment prevalent during the 1970s. The UK debate is of particular interest in the New Zealand context, because reports generated by both the UK accounting profession and the UK government during the past 25 years have influenced the approach taken to valuing ELBs in New Zealand. The most significant of these reports from a New Zealand perspective are the 1975 Sandilands Report, and the 1986 Byatt Report from the UK Treasury.

The Sandilands Report follows an approach generally termed “replacement cost accounting” (RCA). The report takes the view that, if the basis for the appropriate concept of value is the “value of assets to the business”, and that this is given by the “deprival value”, DV, (i.e., the maximum loss a firm will suffer if deprived of an asset), then historical cost is irrelevant. Ignoring stock market valuation, of the other approaches listed above, Hay and Morris (1993, pp. 430-431) indicate that the Sandilands Report tended to favour employing value in purchase—the replacement cost. If either the present value of future cash flows or the net realisable value were higher than the replacement cost then, on being deprived of the asset, the firm would replace it (for use or resale respectively), and the deprival value would be the replacement cost. Only if this latter value were the highest would the firm not replace the asset, and consequently the cost to the company of being deprived of the asset would be the higher of the other two values. The replacement cost in question is the “written-down” (i.e., depreciated) value of the “written-up” (i.e., current) replacement cost, the latter being found from reference to a current price index of assets for the industry in which it operates (as such accounting for appreciation due to inflation). Therefore, this implies that the asset’s valuation (V) is as follows:

$$V = DV = \min[\text{DRC}, \max(\text{EV}, \text{NRV})] \quad (7.27)$$

The Byatt Report, which is directly referenced in the New Zealand Ministry of Commerce’s rationale for the valuation methodology applied to ELBs (Ernst and Young, 1994, p. 7), is specifically aimed at developing accounting policies for nationalised industries.¹⁹ Like the Sandilands Report, it

¹⁹ In describing the concept of deprival value, the Ministry of Commerce (Energy Policy Group, 1994c, p. 6) defined it as the *minimum* loss that a business would suffer if deprived of an asset. However, in either case, the valuation rule is given by

implicitly favours replacement cost valuation. Hay and Morris (1993, p. 432) note that this is on account of RCA's applicability to the public sector, but also point out that such an approach stems from a contestable markets framework. Firms in a contestable market are restricted to earning only normal returns (i.e., zero economic profit) since they are vulnerable to hit and run entry. Where potential entrants consider entering the market, the replacement cost of capital represents the cost of their entry.

Hay and Morris discuss that RCA is criticised by Edwards *et al.* (1987) as missing the real purpose of accounting data, which they consider is to act as a signal to a firm as to whether it should carry out further investment or, instead, actually divest. Edwards and his colleagues advocate an alternative measure to RCA, and to the various other proposals, called "real terms accounting" (RTA), which is based on a set of "value-to-owner" rules. The appropriate valuation depends on the "opportunity cost" of capital as given by the value to the asset's owner. They consider that this value will be equal to the *greatest* of the (depreciated) replacement cost, the present value of expected future earnings, or the net realisable value. Hay and Morris (1993, p. 432) note that the opportunity cost concept directly answers the question whether or not the firm should expand or contract its operations, through a comparison of the opportunity cost of the firm's assets and the actual cost of funds. In contrast, Hay and Morris profess that the RCA approach as presented in the Byatt Report is unable to provide guidance for investment decisions similar to the RTA approach, in markets which are not contestable.

7.4.2 The Optimised Deprival Valuation (ODV) Methodology

New Zealand's ODV methodology was first officially detailed in the 1994 ODV Handbook (Energy Policy Group, 1994b) but by August 2001 had been revised three times, including versions published in April 1999 (Energy Markets Regulation Unit, 1999), and October 2000 (Energy Markets Regulation Unit, 2000c). These changes were made to keep up with the pace of structural changes in the industry, particularly the legal separation of energy retailing and line services from 1998 (§2.4.5). Consequently, the governing regulations were subsequently revised and reissued in 1999, and further amended during 2000 (§2.4.7).

The stated aim of the ODV methodology is to value an ELB's network assets "at the level at which they can be commercially sustained in the long term, and no more". An ELB's ODV "should be equal to the loss to the owner if they were deprived of the assets and then took action to minimise their loss" (Energy Markets Regulation Unit, 2000c: s2.2). In the words of one of the principal authors of the original ODV Handbook, the ODV methodology is "soundly underpinned theoretically and is capable of giving a valuation outcome *consistent with contestable market outcomes*" (Wilson, 2000b, p. 4; emphasis

(7.27). This is almost identical to Baumol's (1971) definition of the "opportunity value" of an asset as "the *minimum* loss to which a firm would be subjected by its disappearance" taking into account associated "changes in both costs and revenues".

added).²⁰ The ODV is not required to be used for either taxation purposes or as the basis for setting line charges. Consequently, book values may differ from the ODV value (ODV Handbook, s2.3).²¹

In most cases, an ELB's ODV is equivalent to the optimised depreciated replacement cost (ODRC) of its network assets. The ODRC is the *replacement* cost of the existing system fixed assets at modern equivalent asset (MEA) value, which have been *optimised* from an engineering standpoint, and *depreciated* according to their age. To develop the ODRC first entails the preparation of a detailed asset register of the distribution network, and the assets are valued at their current replacement cost (RC) based on the current state-of-the-art (i.e., MEAs). To ensure consistency across the industry, the ODV Handbook presents tables containing maximum values for most MEAs. These RC values are subsequently depreciated, using straight line depreciation, based on the remaining life of the actual assets that these MEAs would replace, to provide the network's overall DRC value (s3.18). Again, in order to achieve industry-wide consistency, the Handbook also prescribes maximum asset lifetimes for depreciation purposes.

A system optimisation procedure is then applied in order to “determine a value of system fixed assets that is the counterpart to the market value of the assets of a business in a competitive market”. The Handbook considers that this “is the value of the assets on which such a business could earn a normal rate of return commensurate with the risk that business faces” (s2.11). Optimisation consists of “removing any surplus assets or excess capacity from the network configuration, given the required level of service and network capacity” (s2.12), assuming that the boundaries of the ELB, as well as the location and number of existing consumers, are fixed.²² The procedure comprises three stages: (i) identifying stranded assets, those which are no longer required to supply line services to consumers; (ii) optimising the system configuration, meaning that assets over and above those required to meet standard quality of supply criteria should be “optimised out”; and (iii) optimising elements of the system, meaning that any *excess capacity* should be optimised out, particularly if assets with lower capacity and a lower replacement cost would be adequate to meet existing and projected supply (s3.31). Once the optimised system has been determined, any parts of the “notional” optimised network which differ from

²⁰ As critics of the ODV highlight, there is a problem in justifying prices on the basis of the underlying asset valuation.

²¹ Unless otherwise indicated, section numbers cited in the ODV Handbook refer to the October 2000 edition (Energy Markets Regulation Unit, 2000c).

²² This is somewhat similar to the “scorched earth” or “scorched node” approach taken to determine network costs in the US telecommunications industry (e.g., Sidak and Spulber, 1997, p. 421; Parsons, 1998). Parsons (1998) explains how there are a range of computer “cost proxy models” available to determine least (incremental) cost configuration of local exchange service on the basis of public domain data on households, businesses, terrain, and road networks over a relatively small geographic area. Sidak and Spulber criticise these models as follows: “Such assumptions are only meaningful if that is indeed the relevant decision, such as might be the case in rebuilding a local exchange network that had been seriously damaged by war or natural disaster”.

the real network should be revalued, based on the RCs of the notional MEAs, and depreciated to reflect the service potential (§7.1.1) of the real assets which they replace (s3.53).²³ The optimisation process thus seeks to disallow cost recovery of “inefficient’ and/or “gold plated” pre-reform investment decisions (§2.2.3), as well as to provide an incentive for future investments to be optimal. As such, the ODV method is intended to promote both *productive* and *dynamic* efficiency.

7.4.3 Economic Value (EV) of Network Segments

However, the ODV of any particular “segment” of a network is not simply its ODRC. The Handbook (s2.5) requires that the ODV be determined from the minimum of that segment’s ODRC and its “economic value” (EV). The EV of the system fixed assets in any network segment is defined (s2.19 and s3.77) as the maximum of: (a) the present value (PV) of the after-tax cashflows (ATCF) attributable to that segment, less any *initial* investment in non-system fixed assets (NSFA) and working capital (WC) associated with the asset; and (b) the segment’s net realisable value (NRV), in other words, its “*scrap value*”. The Handbook (s3.80) indicates that the NRV of an asset is its value in its “best alternative use”, which is considered to be the potential proceeds, less the costs, of disposing of an asset. This is particularly applicable to assets which may have become *stranded* due to consumer disconnection, or where the revenue associated with a particular network segment is so low that the ELB would be better off dismantling that segment and selling off its assets piecemeal.²⁴

Consequently, the overall ODV methodology, shown in (7.28) and (7.29) below, is broadly similar to the approach outlined in the 1975 Sandilands Report, as shown in (7.27), with the additional requirement that the network also be subjected to an “optimisation out” of surplus assets and excess capacity.

$$V = \text{ODV} = \min[\text{ODRC}, \text{EV}] \quad (7.28)$$

$$\text{where: } \text{EV} = \max[\text{PV}(\text{ATCF}) - \text{NSFA} - \text{WC}, \text{NRV}] \quad (7.29)$$

The future ATCF attributed to the segment are not to be determined from the actual revenue, based on existing tariffs, but on the “profit maximising revenue” that would be earned from the “(long run) profit maximising tariffs that could potentially be charged” (s3.72). By definition, profit maximising tariffs can be no greater than “maximum *sustainable* tariffs”—tariffs higher than this would cause consumers to disconnect, should alternative sources of supply be less expensive (s3.73). The

²³ The concept of “service potential” is explicitly defined in the 1999 Disclosure Regulations as follows: “‘Service potential’, in relation to an asset, means the output or service capacity of that asset, determined by reference to attributes such as physical output capacity, associated operating costs, useful life, and quality of output”.

²⁴ Allowing an ELB to value a segment at its NRV protects the ELB from the requirement, in s62 of the *Electricity Act 1992*, that it maintain existing consumer connections until March 2013, unless consumers agree to disconnection.

Handbook (s3.74) indicates that a range of such sources should be considered, including: (i) disconnection from the network with electricity supply from a local generator; (ii) substitution of all or part of the electricity supply with other fuels; and (iii) direct supply from the transmission grid or from a neighbouring ELB, in other words, “bypass supply” (§3.3.1).²⁵ Since a consumer’s motivation to disconnect is actually based on the final *bundled* tariff which the consumer is charged by its energy retailer (i.e., comprising both line charge and energy charge), the maximum sustainable tariff is to be assessed considering the unit cost of the next best alternative source of *delivered* energy.²⁶

Further, ATCF is to be calculated by applying an “avoidable (incremental) cost allocation methodology” (ACAM). The ACAM approach, which is outlined in the Electricity Information Disclosure Handbook (Energy Markets Regulation Unit, 2000b), bears a strong resemblance to the language introduced by Faulhaber (1975) regarding incremental and standalone cost test (§4.3.2).²⁷ The ODV Handbook states (s3.78) that each segment should be treated as “*incremental*” to the rest of the network, including other segments, which are tested separately. The rest of the network is treated as the “standalone” business. Consequently, no physically *common* capacity costs of the upstream zone substation will be included in the incremental cost of the network segment, which as discussed earlier (§5.3.3), could be a fallacious assumption. The ACAM approach (s3.79) makes an assessment of the expenses, revenues, assets and liabilities that would be avoided by the ELB if it did not operate its incremental *business* (i.e., the segment under examination).

7.4.4 The Evolution of the ODV Methodology

As noted above, the ODV methodology has gone through numerous revisions, and at the time of writing, is again under review. Over time, the rules have become both more detailed as well as more stringent, as successive governments have become rather less willing to allow light-handed regulation a free rein. At the time of the release of the October 2000 edition of the ODV Handbook, the Minister of Energy announced that “the revised ODV Handbook introduces considerably more rigour to the valuation process and is significantly more prescriptive than its predecessors” (Hodgson, 2000a). These changes were introduced even though the previous year’s disclosures indicated that: (i) the overall ELB ODV value had declined by NZ\$100 million (around 2.4%); (ii) the average and median ELB return on

²⁵ Furthermore, the profit maximising tariff (excluding energy costs but including transmission costs) is not to exceed 30c/kWh (s3.76).

²⁶ The potential constraint of embedded subnetwork competition (§3.3.1-§3.3.2) on the line charges for new consumers is not directly alluded to, because the Handbook is written with the valuation of *existing* assets in mind. Nevertheless, embedded subnetwork competition has the potential to limit the maximum revenue obtainable from any new subnetwork, and as such has possible implications on the incumbent’s investment decisions.

²⁷ The Disclosure Handbook (Energy Markets Regulation Unit, 2000b) defines the SAC of a service as the cost that would be incurred if that was the sole service provided, and the IC of a service as the cost that is incurred in supplying the service, excluding all costs which would be incurred if that service were not supplied.

investment (ROI) were only 4.1% and 5.1% respectively; and (iii) line charges for the typical domestic and commercial consumer had reduced by about 2.8% and 3.7% respectively, in real terms (Energy Markets Regulation Unit, 2001).

Minor methodological changes included changing the optimisation procedure from being the first, to being the last step in the ODRC calculation. But more prescriptive elements included that, for instance, in optimising the network the security of supply criterion (§3.1.3) should be limited to $(n-1)$, with the exception of supply to CBDs, or when a specific consumer non-standard contract exists requiring a high level of security (s3.42). Notably, the optimisation of the network must be supported by extremely detailed load forecasting. The existing load and load forecasts at the end of the “relevant planning period” must be provided for “each part of the network”, namely: “each point of connection” to Transpower’s grid; “each zone substation”; as well as “each individual distribution feeder” (s7b). Moreover, the relevant planning period was itself shortened from ten to five years for distribution assets, and for distribution transformers no future load growth should be permitted (s3.37).

7.5 Criticisms of the ODV Methodology

7.5.1 *The Windfall Critique*

Building on their earlier critique of the ODV methodology (i.e., Bertram *et al.*, 1992)—which primarily focused on the valuation of Transpower—Bertram and Terry (2000, p. ii) consider that basing line charges on ODV values has allowed the ELBs to realise windfall capital gains from the post-reform revaluation of their network assets. This has resulted in a substantial wealth transfer from consumers to ELB shareholders.²⁸ As Bertram and Terry rightly point out, and as the current ODV Handbook still makes clear (s2.3), the Government never *mandated* that actual line charges be based on ODVs, although it was also clear that ODVs *could* be used as the basis for determining tariffs (Energy Policy Group, 1994e). Nevertheless, the valuations were primarily intended as a benchmark for performance comparison, not for price setting.

Yet some industry commentators consider that it would have been somewhat naïve for the Government to expect that ODV would not eventually become the basis for determining line charges. Prior to the application of the method, it was suggested that unless line charges were to be derived from a rate-of-return on ODV asset value, then allocative inefficiencies would arise (Saha, 1993). This is because the ELB business value, based on earnings valuation, would be less than the *allowable* asset value, and consumers and suppliers would decide between energy sources on the basis of prices which

²⁸ Bertram and Terry (2000, p. ii) do point out that in some cases, depending on the ELB ownership structure (e.g., trusts; §2.4.1), a considerable portion of this windfall profit has made its way back to consumers in the form of rebates. However, they suggest that there is nothing to ensure that this will continue in the future, and point out that there are “major leakages”, such as tax on the surplus.

might not reflect the true costs of the product. Consequently, for ELBs in private ownership there would likely be pressure from shareholders to maximise both returns and share value, implying the desirability of using ODV as the rate base, since it acts as the *de facto* cap on profit maximisation.

Before the method was introduced, its proponents acknowledged that ODV values of distribution networks would most likely be double their historical book values (e.g., Wilson, 1994). And while the Ministry predicted that efficiency gains would more than offset revenue requirements for any line business that set charges indexed to their ODV (Energy Policy Group, 1994e), the example financial model presented by the Ministry among its original ODV documentation similarly assumed ODVs to be twice the representative ELB's current book value. This doubling of overall ELB asset value (from NZ\$2 billion to NZ\$4.2 billion over the period 1992 to 1999) is borne out by subsequently analysis, and over the same period, ODV pricing has clearly become the standard pricing policy (Bertram and Terry, 2000, p. 7 and p. 21; §4.4.3).

This substantial increase in ELB asset value lies at the heart of the most severe criticisms of ODV. Bertram and Terry (2000, pp ii-iii) consider that the revaluation gains allowed due to the change in the regulatory regime should be treated as income. They state that, for a natural monopoly in particular, which prices its services directly from the value of its assets, the treatment of capital gain is of great importance. Consequently, Bertram and Terry consider that ELBs have been allowed to earn a "return on" (and "return of") capital that it has never actually invested in the business.²⁹ Hence, the primary problem facing consumers is that ELBs are able to "write-up" their asset values using the ODV methodology, and then use these unfairly high values as a justification for capturing monopoly rents. The true returns that ELBs have been making over the past few years should be calculated by combining operating surpluses with the capital gains from asset revaluations. For the six years following corporatisation, Bertram and Terry estimate the industry rate of return to have ranged between 16% and 23% post-tax, more than double that which the Ministry of Economic Development considers appropriate for ELBs (i.e., 7.5% to 10%). Consequently, Bertram and Terry (2000, p. 5) observe that many submissions from consumer groups to the Inquiry into the Electricity Industry in 2000 focused on the issue of wealth transfer from consumers to ELB shareholders as a result of the application of the ODV methodology. On the other hand, business interests countered by suggesting that critics of ELB line charges were confusing monopoly profits with economic rents (Kerr, 1999, p. 5), and that "high (accounting) rates of return may reflect timing issues, valuation issues or inframarginal economic rents" (NZBR, 1999, p. 4).

²⁹ Expanded returns not only come from compensation for the use of capital employed (i.e. return *on* capital), but from increased depreciation charges, in other words, return *of* capital (Bertram and Terry, 2000, p. 12).

Bertram and Terry point out that the Ministry itself acknowledges that windfall profits can result from revaluation. However, this is highlighted inconsistently with respect to the difference between economic value (EV) and ODRC, rather than to that between historic book value and ODV: “If a network segment valued at EV was revalued upward to ODRC, then the ELB may be able to raise prices to all consumers and justify the price increase on the basis that returns are below or equal to normal returns based on the ODRC of the whole network. The result is a ‘windfall’ gain to the shareholders of the ELB at the expense of the consumer’ (Energy Markets Regulation Unit, 2000a, p. 9).

7.5.2 *The Circularity Critique*

The next area of criticism is more general, and relates to the concepts and assumptions underlying the ODV methodology, rather than to any one-off wealth transfer resulting from its introduction. Bertram and Terry (2000, p. 17 and p. 24) protest that “deprival” is not a cost concept, but a concept from the world of insurance and damages estimation; one altogether unsound for pricing electricity line services. By allowing it, the regulator legitimises monopoly pricing regardless of its impact on economic welfare (Bertrand and Terry, 2000, p. 23). They suggest that the problem with the “efficient prices” justification of ODV-indexed pricing is that this is based on prices set in a hypothetical long term competitive market with balanced supply and demand, and short-lived non-lumpy assets; a picture not applicable to distribution networks.

When inappropriately applied to price setting for natural monopolies, the ODV methodology is considered to be “circular” (e.g., Bertram *et al.*, 1992, p. 144; §7.1.2). “Because the asset value is endogenously determined, the firm can always appear to be earning no more and no less than the competitive rate of return” (Simon Terry Associates, 2001).³⁰ From the perspective of the asset owner there is a “virtuous circle” of increased asset values leading to increased line charges which, in turn, underpin the increased asset value. Only once some limit to future revenues has been exogenously imposed, such as the pressure from truly competitive market conditions, can the circle be broken and deprival value become anchored to a fixed rate base. Bertram and Terry (2000, p. iii) state that “as deprival value rests on revenue expectations, it cannot at the same time be used as the basis for setting revenues. For this reason its use for rate base purposes was rejected by the US Supreme Court”. However, this latter argument against the use of ODV relies more on legal precedent than economic justification. In fact, in one of the dissenting opinions of the Supreme Court case to which Bertram and Terry are referring, Justice Jackson—famous for his role as the chief US prosecutor at the Nuremberg

³⁰ On the other hand, Wilson (2000a), for one, suggests the ODV approach actually *breaks* the circularity: “The very choice of an asset based valuation method such as ODRC recognises that using a discounted free cash flow model of valuation, which would appear to be appropriate in a competitive market, is inappropriate in a natural monopoly situation as it would lead to a circularity of argument. This is because a natural monopoly is, in the absence of regulatory intervention, unconstrained in its tariff setting and hence is free to determine its level of prices which in turn determine cash flow based valuations”.

trials—cautions against such a critique: “The unfortunate effect of judicial intervention in this field is to divert attention of those engaged in the process from what is economically wise to what is legally permissible”.³¹

³¹ Bertram and Terry are referring to the landmark US Supreme Court case of *Federal Power Commission vs Hope Natural Gas Co.* (US Supreme Court, 1944). It is important to note that this case involved not a question of economics, but a point of law based around disputes over accounting conventions. The issue was whether a specific reduction in natural gas rates ordered by the Federal Power Commission had resulted in rates that could be deemed to be *not* “just and reasonable”, given that the governing legislation (the Natural Gas Act 1938) gave the Commission the power to order a “decrease where existing rates are unjust, unlawful, or are not the lowest reasonable rates” [emphasis added]. Justice Jackson concluded his dissenting opinion by recommending that the case be returned to the Commission: “This problem presents the Commission an unprecedented opportunity if it will boldly make sound economic considerations, instead of legal and accounting theories, the foundation of federal policy”. Earlier he makes it clear that he sees the Commission’s order to be based on accounting conventions rather than economic realities: “Even as a recording of current transactions, bookkeeping is hardly an exact science ... the fallacy of using [accountancy] as a sole guide to future price policy ought to be apparent. However, our quest for certitude is so ardent that we pay irrational reverence to a technique which uses symbols of certainty, even though experience again and again warns us that they are delusive. Few writers have ventured to challenge this American idolatry”.

In any event, Bertram and Terry’s interpretation of the Court’s *majority* ruling is not necessarily the same as others. For instance, the Energy Information Agency of the US Department of Energy summarises the implications of the Hope decision thus: “In settling the ... case, the Supreme Court closed a longstanding dispute by *allowing either original or replacement cost accounting in utility ratemaking*, so long as just and reasonable rates result” [emphasis added] (EIA, 2000, Ch. 4). Similarly, Williamson and Mumssen (2000, pp. 15 and 28) state that the Hope case “established that utilities should earn the opportunity cost of capital applied in other comparable uses – in order to attract investment ... by giving the company a revenue which covers both a return on capital (profits) and return of capital (depreciation)”. The majority decision of the US Supreme Court made it clear that, the “heavy burden” of proof was on Hope Natural Gas to demonstrate that the Commission’s order was invalid because it had resulted in rates that were “*unjust and unreasonable*” in their consequences. However, it was not proposed that “just and reasonable” be assessed in terms of economic efficiency. The Supreme Court pointed out that, in preparing the Natural Gas Act, Congress had “provided no formula by which the ‘just and reasonable’ rate is to be determined. ... [Congress] has not expressed in a specific rule the fixed principle of ‘just and reasonable’”. The Court simply indicated that “the rate-making process under the Act ... involves a balancing of the investor and consumer interests”, and did not rule on what the appropriate method for determining rates or a rate base under the standard of ‘just and reasonable’ might be. The majority decision did not address the validity of specific valuation methodologies, and did not explicitly mention deprival value. In fact, the Court ruled that: “Under the statutory standard of ‘just and reasonable’ it is the result reached not the method employed which is controlling. ... It is not the theory but the impact of the rate order which counts. If the total effect of the rate order cannot be said to be unjust and unreasonable, judicial inquiry under the [Natural Gas] Act is at an end. The fact that the method employed to reach that result may contain infirmities is not then important. ... The conditions under which more or less might be allowed are not important here. Nor is it important to this case to determine the various permissible ways in which any rate base on which the return is computed might be arrived at”. Trebing (2000) summarises the decision as follows: “In the *Hope* case ... the method of regulation was subordinated to the doctrine of end result. That is, the end result of regulation was the criterion for judging the reasonableness of a regulatory action and not the regulatory techniques that were employed”.

Apart from the methodological issue, another important point is that Hope Natural Gas was not solely a network utility, but a firm engaged in the production and wholesaling of natural gas (a “depleting” natural resource), not just its

But even business interests have tended to provide general criticisms of the ODV methodology, with the New Zealand Business Roundtable (NZBR, 1999, p. 2) expressing concern that “the drift in New Zealand into rate-of-return regulation of line businesses based on optimised deprival values is a threat to incentives to control costs and provide the appropriate quality of supply”. However, they suggest that problems with ELBs should not be attributed to any naturally monopolistic characteristics, but to the regulatory environment; one that creates incentives to inflate ODV values, and hence prices (Kerr, 1999, p. 3). Hence, their solution is not to substitute a different valuation methodology for ODV, or even to substitute incentive regulation for *de facto* ROR regulation, but to reduce the level of regulatory intervention in the power distribution sector entirely (§3.3.4).

Similar—if not so extreme—criticisms have also been lodged by some of the ELBs themselves. Mercury Energy Lines Business (1999, pp. 21 and 24) criticises the implicit rate of return regulatory regime as including incentives to over-invest in capital. In addition, Mercury acknowledges that line businesses have better information about their cost structures than regulators will ever be able to obtain, and face incentives to hide and distort information. This might seem a somewhat surprising argument coming from within the industry, but as discussed earlier (§3.1.5), such comments were made in response to the threat of heavier-handed regulation.

transportation. Hence, its relevance to ODV, in light of the subsequent changes in utility industry structure, is somewhat diminished. Interestingly, in his dissenting opinion, Justice Jackson presaged the future evolution of regulatory theory and practice for network industries (i.e., functional unbundling, with subsequent deregulation of the energy production and wholesaling aspects of the business, while still regulating transportation; §2.2.2): “Hope’s business has two components of quite divergent character. One, ... is essentially a transportation enterprise [and t]he other ... supply of natural gas”. The writer suggests that the price of gas in the field should be fixed in the same manner as any other commodity, and that such a price would not be calculated to produce a fair return on any rate base. Consequently, the opinion goes on to criticise the nature of rate base pricing, particularly in regard to the *non*-network part of the business: “It is necessary to a ‘reasonable’ price for gas that it be anchored to a rate base of any kind? Why did courts in the first place begin valuing ‘rate bases’ in order to ‘value’ something else? ... Does anybody imagine that [a costly producer] can get or ought to get for his gas five times as much as [a less costly producer] because he has spent five times as much? The service one renders to society in the gas business is measured by what he gets out of the ground, not by what he puts into it, and there is little more relation between the investment and the results than in a game of poker. ... No one seems to have questioned that the rate base method must be pursued”.

The Court’s majority decision did however address the broad notion of circularity in rate setting outlined by Bertram and Terry, implying that a ‘fair’ rate base should be derived from fair prices, rather than vice versa: “‘Fair value’ is the end product of the process of rate-making not the starting point. ... The heart of the matter is that rates cannot be made to depend upon ‘fair value’ when the value of the going enterprise depends on earnings under whatever rates may be anticipated”.

There are two distinct issues here. First, is the ODV method itself “unfair”? In other words, does it allow monopoly rents to be taken by ELBs? Second, is the *revaluation* the “unfair” aspect of ODV? Even if the method itself could be considered fair, say if applied to a newly-established ELB on a greenfields basis, perhaps the upward revaluation to ODV for an older ELB could nevertheless be considered to allow substantial “unfair” windfall gains.

In some ways, although appearing to consider the method unfair on both grounds, Bertram and Terry’s critique of ODV is more relevant to the second one. For instance, they cite the 1995 decision in the International Court of Justice regarding Heathrow Airport user charges that: “an asset’s economic rate of return for a single period is its economic income (that is the combined sum of the net cash flow received during the period and the change in the value of assets over the period) expressed as a proportion of the value of the assets at the beginning of the period. ... A fundamental difference between an accounting rate of return and an economic rate of return is that the former takes no account of unrealised capital gains or losses whereas the latter includes in the relevant ‘return’ any appreciation or depreciation in the value of the assets the profitability of which is being measured” (Bertram and Terry, 2000, pp 14-15). Consequently, they propose that, in the year the revaluation from historic value to ODV occurs, raising the annual economic rate of return (i.e., the ARP or ROI) far above the WACC, the simplest way to adjust the rate of return to acceptable levels would be to credit the full amount of the revaluation against required revenue in the period of the revaluation. This would effectively result in a one-off rebate to consumers of the cash amount of the book valuation.³²

On the other hand, in discussing ODV’s overall applicability, Bertram and Terry (2000, p. 13) consider that “the generally-accepted principle underlying calculation of the required revenue which a natural monopoly can legitimately collect from its customers is that the owners of the enterprise should be entitled to maintain their financial wealth in real terms (that is, they should suffer no loss of wealth as a result of entry into the business), while recovering operating costs and the market rate of return on their net financial exposure in the enterprise”. Further, they acknowledge that, over time, as a monopoly enterprise is expanded and its existing assets are replaced, these new assets would appear in the accounts at current cost less depreciation. Thus there will be a continual trend for the asset base to tend toward DRC, and financial accounts would reflect ODRC, assuming no technological change that renders parts of the distribution system obsolescent. They maintain, however, that this is not an argument in support of inflating the asset base to ODRC at some mid point in the life of the assets with consequent distributional

³² This argument is effectively the mirror image of Sidak and Spulber’s (1997) argument that utilities facing stranded costs as a result of industry deregulation are entitled to compensation (\$6.3.4); this compensation being the difference between the present discounted value of net earnings under regulation and those expected under competition. (They also suggest that a higher discount rate should be applied in calculating the present discounted value of earnings under competition because of higher risk).

impacts for owners and consumers (Bertram and Terry, 2000, p. 22). Nevertheless, it does imply that the ODV approach would be acceptable had there had been such a large divergence between historic book valuations and the initial ODV valuations.

Bertram and Terry (2000, p. 23) state that the ODV approach will generally result in the value of the network being valued at ODRC, except in those few segments where tariffs are constrained in some way by consumer willingness and ability to pay, and thus those segments must be valued at EV instead. And ODRC is viewed as the maximum rent that could be theoretically extracted without providing the opportunity for a new entrant to duplicate the facilities and enter the market. But assuming that there are no other barriers to entry apart from the level of an incumbent ELB's line charges, if a competitor is unwilling to enter the market at line charges that provide a lower profit than return on its ORC value (depreciated over time), then this would seem to indicate that the incumbent is making a less than normal profit. If ODRC valuation is acceptable in the long term, or on a greenfields basis,³³ then if book value at some point in time has fallen to less than ODRC, in the absence of technological change, this would imply that the assets at some stage have become undervalued. But only in a world without inflation or changes in technology, and where ODV and traditional accounting conventions both use the same depreciation policy (e.g., straight line depreciation) and taxation policy, would historical book values and DRC values be the same.³⁴

7.5.3 *The Optimal Spare Capacity Critique*

Criticism of the ODV has also related to some of the finer points regarding the impact of optimisation rules on dynamic efficiency. A 1999 Ministry of Commerce discussion paper on the appropriate optimisation rules to include in the ODV methodology had the following view concerning spare capacity in a power distribution network.

Provision of built-in capacity, over and above that required to meet existing load requirements under normal operating conditions is, if this additional capacity is included in the asset value, a cost premium on the existing consumers. This is because, by virtue of its natural monopoly, an ELB is in a position to achieve a return on all recognised assets, irrespective of whether these are

³³ This is perhaps what Bertram and Terry (2000, p. 25) mean when they suggest that: "Only at the end of a full cycle of realised asset replacement conducted under agreed (or legislated) new full-cost-recovery rules, could there be legitimacy for ODRC-based charges".

³⁴ One of the dissenting opinions of the Court in the Hope case (see fn. 31), pointed out this equivalence (Justice Reed, US Supreme Court, 1944): "Historical cost, prudent investment and reproduction cost were all relevant factors in determining fair value. Indeed, disregarding the pioneer investor's risk, if prudent investment and reproduction cost were not distorted by changes in price levels or technology, each of them would produce the same result". "Prudent investment" was taken to mean "the sum originally put into the enterprise, either with or without additional amounts from excess earnings reinvested in the business", while "reproduction cost" was consider to be "the minimum amount necessary to create ... a modern plant capable of rendering equivalent service". In other words, reproduction cost is similar to deprival value.

currently required. ... In contrast to premium [i.e., spare] capacity provided to ensure supply reliability, capacity built into a network solely to meet future load growth is of no immediate benefit to existing customers. The question then arises as to why existing customers should pay a premium to cover the cost of assets from which they derive no benefit. A firm in a competitive environment would not be able to make a return on such assets from existing customers: (Ministry of Commerce, 1999, pp. 3 and 5).

While stating that “a rigorous regulatory approach would not provide for any premium assets” (i.e., spare capacity), the Ministry’s discussion paper acknowledged that some spare capacity to accommodate future load growth “is prudent business practice”. The Ministry suggests that “the amount of premium capacity to meet future load requirements that can be left in the network after optimisation is thus a compromise between a ‘no frills’ approach based on economic theory and a more pragmatic approach that recognises the need to provide network owners with a development incentive”. The clear implication is that *economic theory would suggest that any spare network capacity could not exist in a competitive market.*³⁵ Given that spare capacity can be efficient if a program of anticipatory construction is optimal (§6.2), the merits of this claim are investigated in Chapter IX. The Ministry’s discussion paper noted that the ODV’s optimisation rules at the time (i.e., Energy Markets Regulation Unit, 1999) permitted a planning period of ten years across all sectors of the network, and concluded that “this is inappropriate, even if a development incentive is allowed”. One option presented in the paper was that assets installed to meet future load growth be entirely excluded from the ODV asset valuation, in which case “the maximum capacity of any segment of the optimised network shall be determined by the existing load”.

Criticism of this viewpoint came from one of the initial authors of the first ODV Handbook. Providing comments on this discussion paper, Wilson (2000a) warned that the proposition “that all spare capacity should be optimised out, irrespective of how it arises, is out of tune with the achievement of long-term dynamic efficiency in network investment”. Further, the view that spare capacity is of no benefit to existing consumers implies that they “are short sighted in their views and in fact that they may not remain customers for a sufficient length of time or have an interest in long-term optimality. We do not accept that this view is appropriate or proven, let alone desirable”.

Optimisation of investments must have regard to the full life cycle costs of the assets concerned, not merely the initial cost. This implies the need for a long-term view on network optimality, rather than an assessment of the immediate position. In the line business case, it requires the consideration of projected future growth in demand and the long term optimal planning of

³⁵ This viewpoint was reiterated in a later discussion paper of the Ministry: “In a competitive environment ELBs would not be able to charge customers prices that would achieve a normal rate of return on ... premium assets” (Energy Markets Regulation Unit, 2000a).

investment to best match future demand in addition to serving the present situation (Wilson, 2000a).

Notwithstanding submissions such as Wilson's—that international best practice in subtransmission planning has a horizon of around 15 years—the current ODV Handbook was revised to allow for a shorter planning period (§7.4.4).³⁶ When optimising the distribution system under the current ODV Handbook, the maximum capacity of any part of the optimised network is to be determined by the forecasted load growth at the end of the “relevant planning period”, and in no case should optimised capacity exceed existing capacity (s3.35). The relevant planning period for transmission networks (defined as those where voltages are above 33 kV), subtransmission networks, and zone substations, is stated as being 10 years, whereas for HV and LV distribution (i.e., 11kV and 400V) it is only five years (s3.37).³⁷

On the other hand, the potential optimality of capacity replacement is addressed in the Handbook's rules regarding depreciation. If “a class of assets is routinely replaced as part of the evolution of the system before its technical life expires, then this should be taken into account in assessing the [total asset lifetime] for that class of assets” (s3.25). This suggests that depreciation can be *accelerated* to ensure that when an asset is (optimally) replaced prior to its asset lifetime, the accumulated depreciation is equivalent to the asset's replacement cost (§7.3.3).

7.5.4 The Tomorrow's Costs Critique

As part of its Power Package of October 2000 (§2.4.7), the New Zealand Government announced that it would charge the Commerce Commission with reviewing the ODV methodology. Consequently, the Commission's draft Price Control Study of airports (i.e., Commerce Commission, 2001; §5.3.2) was scrutinised with interest by power sector commentators, given that airports had been using the ODRC approach to value some specialised airfield assets. In the draft Study, the Commerce Commission (2001, p. 8) made it clear that the appropriate pricing benchmark was competitive market outcomes, and that a key pricing principle was that “today's consumers should only bear today's costs”.

The Commission then grappled with the pros and cons of historic or replacement cost valuation, the latter being already implemented by a number of New Zealand airport companies through the

³⁶ Wilson (2000a) also pointed out that, given New Zealand's low demand growth rate, even a *longer* planning period might be appropriate.

³⁷ For distribution transformers, no future load growth is permitted, and these transformers must be optimised in terms of only partial capacity utilisation based on current network loadings (s 3.37). This is not however of as much concern as the planning horizon for other assets, since as mentioned earlier, load growth in a network is more typically due to an increase in the number of consumers, rather than the consumers increasing their existing demand substantially. Such assets can be considered more “individual” or dedicated” (§3.1.1).

(voluntary) adoption of the ODRC methodology. The Commission noted that the ODRC approach is intended to “mirror conditions in a competitive or contestable market inasmuch as the firm does not make a return on inefficient investments”. Nevertheless, in this particular context, the Commission rejected the ODRC methodology on the basis that it did not meet the “today’s costs” principle, and recommended what could perhaps be described as an optimised depreciated *historic* cost approach.

The Commission’s preliminary view is that specialised airfield assets should be included in the asset base at historic cost. The assets should also be depreciated and optimised as appropriate. The use of replacement cost would run contrary to the Commission’s view that today’s acquirers of airfield activities should only bear today’s costs. Historic cost is consistent with the fundamental principles adopted by the Commission. It provides investors with a return on the amounts invested, and preserves incentives to invest in the future. Investors are compensated for inflation through the use of a nominal WACC (Commerce Commission, 2001, p. 11).

Support for the broad thrust of Commission’s findings came from Simon Terry Associates (2001), the organisation associated with Bertram and Terry’s (2000) circularity and windfall critiques of the ODV methodology discussed above (§7.5.1 and §7.5.2). Reiterating these earlier critiques of ODRC (and by implication ODV), Simon Terry Associates affirmed that the Commission’s view that “replacement-cost based pricing unjustifiably loads potential future costs onto today’s users of the service” as being “well grounded in the relevant economics literature”.

On the other hand, the draft ruling drew critical submissions from the airport companies themselves, who hired such regulatory experts as Alfred Kahn (§5.3.2), as well as from the ELBs, who saw this draft ruling as setting a possible precedent for their own industry. Representing one of the airport companies, Zijl and Irwin (2001) used a similar logic to Bertram and Terry’s windfall critique, but *in reverse*: “[the Commission] appears not to have appreciated the risk created by ‘midstream’ changes in valuation approach applied to investments already made. The arguments for choosing between historic-cost and ODRC approaches at the outset of an investment in a sunk asset are subtle, and different approaches might reasonably be chosen in different circumstances. But the argument against *changing* the approach in midstream are strong”.

However, Zijl and Irwin also objected to the Commission’s view that the ODRC methodology “involves an element of pre-financing”, and that in “accepting ODRC, today’s consumers will pay for some of tomorrow’s costs”. The Commission expressed concern that “there is no guarantee that any pre-financing of future replacement is set aside and kept for that purpose and current users have no guaranteed rights in the future”. Zijl and Irwin point out that the difference in the time profile of revenues under ODRC and the historic-cost approaches does not relate to pre-financing.

While Zijl and Irwin claimed that ODRC valuation does not contribute to tomorrow's costs and therefore the Commission's rejection of the methodology on the grounds that it does not meet the "today's costs" principle, Kahn (2001) rejected the "today's costs" principle itself. Kahn did not address the validity of the ODRC methodology directly. Rather, he examined what should or should not be included in economically efficient prices, and approached the issue from a marginal cost perspective (§4.1).

Ideally efficient rates would (and regulated rates should) track short-run marginal costs (including congestion costs); and that where short-run marginal cost pricing is impractical, efficient prices would be based on their surrogate, long-run incremental costs. The latter include costs associated with the construction of additional capacity when and as those costs become reasonably predictable. Correspondingly, I have therefore rejected the proposition that costs associated with increases in capacity, either (i) actively in construction or (ii) by general agreement necessitated by existing at clearly contemplated levels of demand, do not belong in regulated rates. I have therefore explicitly rejected the opposing propositions, such as that "today's users should bear only today's costs" or that plant not yet in service is not "used [or] useful" when these propositions are interpreted to exclude the causal responsibility of today's demand for the need to build up additional capacity. ... The incremental capital costs required to hold (short run) congestion costs to economically optimal levels *are* today's costs (Kahn, 2001).

Parts of Kahn's argument also bore some resemblance to Wilson's above (§7.5.3), since it recognised that today's consumers should not be exempted from contributing to the costs of capacity associated with either "clearly contemplated levels of demand", or "plant not yet in service".

The Commission nowhere seems to pose the question of whether or the extent to which the costs of "pre-financing of new ... investments that will be 'used and useful' should or should not be part of the marginal costs on the basis of which efficient prices are to be set. ... The asserted principle that "today's users should only bear today's costs" is meaningless in an industry with lumpy, long lived assets and fails to recognise the causal effect of today's demand on the investment requirements of tomorrow, and accords neither with fairness nor economic efficiency (Kahn, 2001).

Nevertheless, in this particular instance Kahn recommended that airport facilities be priced on the basis of long run incremental costs (§4.1.5) because congested use of those facilities was, in his view, *already* occurring. Kahn concluded that charging the present value of the costs associated with capacity expansion—capacity which in this case has not *yet* been completed and thus cannot be used even partially—serves as a "concededly superior surrogate" to short run congestion costs. Spare capacity is not really spare because it can be used immediately if demand were to increase, rather there is an opportunity cost associated with tying up capital associated with "prudent" expansion "within the

timeframe proposed by the airport”. Kahn thus argues that including *construction* costs in the asset valuation base moves “the time profile of rates closer to the profile of marginal costs”.

The next two Chapters turn to a question similar to that which Kahn suggested be posed: to what extent should the cost of today’s investments in assets that will only become utilised at some future date, be included in today’s efficient and anonymously equitable prices? This question has also been foreshadowed by the earlier general discussion of intertemporal anonymous equity (§6.4.2), and more specifically in the discussion of the second of Baumol and Sidak’s (1994a) demand sequencing scenarios (§6.2.1). Kahn emphasised that “the concept of ‘cost’ has no meaning in either economics or logic except in terms of *causation*”. As the discussion of Baumol and Sidak’s sequencing of demands example similarly explained, today’s consumers, simply by engaging in consumption today, may optimally require that spare capacity be also built today, for future use by future consumers. Should today’s consumers defer their demand today, then they will also defer the need to incur the costs of spare capacity.

CHAPTER VIII

A TWO-GOOD/TWO-PERIOD MODEL OF INTERTEMPORAL SUBSIDY-FREE AND SUSTAINABLE REVENUES

Perfect contestability is a theoretical benchmark that is by its very construction immune from considerations of strategic behaviour by dint of its assumption of the absence of economically sunk costs and irreversible commitments necessary for entry: William Baumol, John Panzar and Robert Willig (1988, p. 490)

Subsidy-free prices do no more than ensure that the production and sale of each commodity makes all consumers at least as well off as they would otherwise be: US pioneer of subsidy-free pricing, Gerald Faulhaber (1975)

The previous Chapters have emphasised the importance of the time dimension to efficient and “fair” prices for power distribution services. It may be quite acceptable for current consumers to contribute toward the costs of assets serving future consumers as long as costs and investment decisions are intertemporally interdependent (§6.4.2). But how much should that contribution be? In Chapter VII (§7.3.1) a simple two-period model of intertemporal subsidy-free prices was developed where demand is constant. This Chapter lays the groundwork for the development of a simple model of intertemporal subsidy-free *prices* in the presence of demand growth (Chapter IX), by first outlining a model of intertemporal subsidy-free *revenues*. The model is an extension of the model which Baumol and his colleagues used to investigate intertemporal unsustainability (BPW, Chs. 13-14). Their model serves a useful starting point because it was derived within the context of contestability theory, where sustainability and the absence of cross-subsidies are viewed as but two sides of the same coin.

8.1 A Single-Good/Two-Period Model of Contestable Natural Monopoly

8.1.1 BPW’s Model of Intertemporal Unsustainability under Capacity Expansion

As BPW themselves point out, every industry that exists for more than an instant in time can be interpreted as a multiproduct industry, because a single good produced in two or more different periods can be considered a different product in each period (BPW, p. 371). However, this can result in some terminology problems. Consequently, throughout Chapters VIII and IX, the *time* of consumption is not considered to be a characteristic of a “*good*”, whereas the term “*product*” is used to distinguish between two different goods consumed during the same period, as well as to distinguish between a single good consumed in two different periods. Therefore, BPW’s model of intertemporal unsustainability, which BPW themselves actually describe as a single-*product*/two-period model (BPW, p. 398), is hereafter referred to as a single-*good*/two-period model, comprising one good, but two products.

BPW use their model to prove that there is no set of “stationary” product “prices”—actually more correctly “revenues”, since they relate to receipts over an extended period of time—which allows

an incumbent firm to prevent entry at the time when it is efficient to expand capacity, even if that firm can provide the least cost intertemporal supply configuration (BPW, pp. 407-413). Key assumptions of their model are that: (i) demand is growing and is for a non-storable good; (ii) there are declining average (incremental) costs of capital construction; (iii) entry and exit are free; and (iv) construction costs are “sunk” and are the determining cost component. More specifically, this latter “*sunk costs assumption*” is articulated by BPW as requiring that: “the constructed facilities have no other valuable use outside the industry in question, so that once built, these facilities are sunk” (BPW, p. 407). As discussed earlier (§3.5.2), BPW use the term “sunk” in a number of ways. In this case they are implying that the facilities are *imperfectly fungible*, and then only from the perspective of alternative uses *outside* the industry in question. Hence, although it might appear that by constructing a model involving “sunk” costs, BPW are themselves violating conditions of perfect contestability, as will be seen shortly (§8.1.6), this is not necessarily the case.

In their model, the initial demand for a good (q), starting in year 1 (i.e., the beginning of period 1), is supplemented by demand for a second good (q') in year $T+1$ (i.e., the beginning of period 2). Demand for the initial good is the same in the first and second periods (i.e., $q = q_1 = q_2$). Given the terminology outlined in the previous paragraph, consumption of the initial good in each period can be considered as the demand for two distinct products. The asset (or assets) required to supply either or both goods has the cost function $K(y)$, where y is the vector of product demands. All assets have a finite lifetime of N years, and asset capacity does not deteriorate during its lifetime—in other words, its service potential remains constant (§7.1.1). Costs of operating the plant are ignored, and the plant is assumed to be sufficiently durable to operate with undiminished capacity for the two periods of analysis (BPW, p. 407). This latter assumption is termed here the “*finite demand assumption*”. In addition to the costs of producing capacity, there are also *fixed costs*, $F \geq 0$; one-off costs incurred for the establishment of any firm.

BPW focus on the case where the optimal construction configuration is a program of *capacity expansion* (§6.2). The total cost (TC), in present value terms, of supplying the entire market demand under such a construction configuration is given in equation (8.1), from BPW (p. 408). Given the declining average costs assumption, if fixed costs are non-zero, then the least cost industry supplier will be an intertemporal natural monopoly (BPW, 14A2). (However, when capacity expansion is optimal, and $F = 0$, *two* firms can supply the entire market at the *same* cost as the natural monopolist). The natural monopolist incurs fixed costs F and capacity costs $K(q)$ at the beginning of period 1, and capacity costs of $K(q')$ at the beginning of period 2, which is T time intervals (i.e., years) in length. Consequently, the second period’s capacity costs are discounted by $(1+\rho_T)$, where ρ_T is the discount factor defined earlier in (6.3).

$$TC_{CE} = F + K(q) + \frac{K(q')}{1 + \rho_T} \quad (8.1)$$

The only alternative construction configuration that can potentially be least cost is a program of *anticipatory construction*, for which the total cost of supplying the market is as shown in equation (8.2), where $K(q, q')$ is effectively $K(q+q')$, since each product relates to the same good (BPW, 13M4). Capacity expansion is the least cost construction configuration when condition (8.3a) holds (BPW, 14A3). Written in terms of the length of period 1, the condition for capacity expansion to be optimal can be restated in terms of T as shown in (8.3b). Note that in BPW's example, the finite demand assumption precludes *capacity replacement* from being a possible least cost construction configuration. This assumption is relaxed shortly (§8.3).

$$TC_{AC} = F + K(q, q') \quad (8.2)$$

$$\rho_T > \rho_{AC \rightarrow CE} \equiv \frac{K(q') - [K(q, q') - K(q)]}{K(q, q') - K(q)} = (1 + d)^{T_{AC \rightarrow CE}} - 1 \quad (8.3a)$$

$$T > T_{AC \rightarrow CE} \equiv \log_{(1+d)} \left(1 + \frac{K(q') - [K(q, q') - K(q)]}{K(q, q') - K(q)} \right) \quad (8.3b)$$

8.1.2 Sustainability of Revenues in the BPW Model

For an intertemporal configuration of revenues to be *sustainable*, BPW state that it is necessary for: (i) total costs to equal total revenues, in present value terms (i.e., total industry profits are “normal”); and (ii) the configuration of revenues to be ‘*undominated*’¹ (BPW, p. 409). BPW indicate that sustainability is usually described as a Bertrand-Nash concept (§2.1.7), which means that “stationary” revenues are available to deter entry without their magnitudes ever changing in response to actual or threatened entry. However, in an intertemporal analysis the Bertrand-Nash premise does not preclude prices from changing over time (BPW, p. 372). Rather, the Bertrand-Nash assumption, applied to a perfectly contestable intertemporal market, requires that revenues for every product in the current and all future periods are published in advance, and are not subsequently changed. This implies that both the incumbent and potential entrants have access to perfect information (§5.1.1).

The BPW model only involves two products, and thus a sustainable configuration of revenues only relates to two revenues, P_1 , the revenue received in period 1, and \hat{P}_2 , the revenue received in

¹ A vector of prices (or revenues) yielding a given quantity of profit is *undominated* if there does not exist any other price vector that yields higher profits (BPW, p. 194).

period 2. This latter revenue is obtained from the now greater demand for the same good as that supplied in period 1. Consequently, the total revenue/cost equality condition is given in (8.4) (BPW, 14A4).²

$$P_1q + \frac{\hat{P}_2(q+q')}{1+\rho_T} = F + K(q) + \frac{K(q')}{1+\rho_T} \quad (8.4)$$

After presenting the previous assumptions and expressions, BPW demonstrate that, unless the fixed costs F are sufficiently large, any efficient industry configuration that involves capacity expansion is unsustainable, in other words, no set of sustainable revenues exists. To demonstrate unsustainability does not require an examination of *all* possible entry strategies, it only requires that the incumbent is unable to prevent entry for at least a *single* entry strategy, for any possible set of its own pre-announced revenue requirements for each current and future product. BPW examine two entry strategies (BPW, pp. 410-411): (i) an entrant (termed a “*Type 1 Entrant*”; BPW, p. 426) that plans to construct q units of capacity in period 1, and to sell q units in both periods 1 and 2; and (ii) an entrant (termed a “*Type 2 Entrant*”; BPW, p. 427) that plans to supply the entire market demand ($q+q'$) in period 2 alone. The constraints that Type 1 and Type 2 entry place on the incumbent monopolist’s revenues are presented in (8.5) and (8.6) respectively (BPW, 14A7 and 14A9).

$$P_1q + \frac{\hat{P}_2q}{1+\rho_T} \leq F + K(q) \quad (8.5)$$

$$\hat{P}_2(q+q') \leq F + K(q, q') \quad (8.6)$$

For these conditions to hold, and to also be consistent with the total revenue/cost equality in (8.4), requires fixed costs to be non-zero and to satisfy the sustainability condition (8.7) presented below (BPW, 14A5). If this condition is *not* met, then, in a perfectly contestable market, this model of an intertemporal natural monopoly is unsustainable. Revenue earned by the efficient monopolist in the first period cannot both be: (i) sufficiently low to keep out a Type 1 entrant; and (ii) sufficiently high to permit the selection of a second period revenue requirement that will keep out a Type 2 entrant, as well as allowing the monopolist to recover its total capital and fixed cost outlay (BPW, p. 413). Consequently, no equilibrium solution exists.

$$\frac{F}{q+q'} \geq \left[\frac{K(q')}{q'} - \frac{K(q, q')}{q+q'} \right] \quad (8.7)$$

² BPW’s presentation is slightly more general in that capacity is not assumed to necessarily match demand, hence y is not demand, but capacity. This suits the use of the model for a power distribution network since it is the demand for capacity which is relevant (§3.1.3). But in the absence of lumpy plant that is indivisible, and if K is a concave function, efficiency requires either the construction of all required capacity in the beginning of the first period, or construction only of the capacity immediately needed at the time (BPW, p. 408).

BPW (p. 411) note that the LHS of (8.7) is the *average cost* of the potential Type 2 entrant, whereas the RHS is the incumbent firm's *incremental cost* of expansion. However, they do not point out that this expression is effectively the standard sufficiency condition for the sustainability of a single product natural monopoly, namely that average costs must exceed marginal (i.e., incremental) costs (§3.4.4).

8.1.3 *Subsidy-Free Revenues derived for the BPW Model under Capacity Expansion*

Sustainability is an *entry* and *equilibrium* concept from the point of view of the *firm*; an undefined number of firms contest with each other to supply a given market (cognisant that some consumers may be willing and able to self-produce). It is a concept of more interest when the conditions of perfect contestability do *not* hold. On the other hand, as discussed earlier (§4.3.5 and §6.4.2), subsidy-free bounds on prices are a *self-production* concept from the point of view of the *consumer*, always derived within a perfectly contestable market benchmark. Subsidy-free prices are determined neither to explain nor to predict actual market behavior, but rather to provide guidance to regulators where the real world falls short of the perfectly contestable ideal (Baumol and Sidak, 1994, p. 43).

If the industry structure is optimally served by a natural monopoly, then only the price offerings of a *single* firm need be considered. From a consumer perspective, *which* firm wins the entry contest to become the monopolist is irrelevant, since all potential entrants are considered to be (statically) symmetric. Any “contest” is notionally between the “winning” natural monopolist firm, and the consumers themselves. However, this contest occurs *before* that firm may enter; it has not yet gained the right to become the industry's incumbent supplier. The Bertrand-Nash premise is now rolled out, meaning that this single potential entrant publishes its revenue requirements for all current and future products, and promises not to renege on the associated prices once the capacity is constructed. In deriving subsidy-free revenues, consumers or consumer groups are assumed to have access to the same technology at the same cost as firms intending to supply the market, and have no loss in utility from self-producing.

In the BPW model, under capacity expansion, bounds on subsidy-free revenues can be derived from (8.4)-(8.6), since these expressions can relate not just to entrants but to consumer group coalitions of self-producers as well. As noted above, to test unsustainability only requires that a single successful entry strategy be found. However, unless *all* possible entry strategies are considered it can only be concluded that the derived sustainability condition is a *necessary* condition for the existence of sustainable revenues; the condition is not necessarily a *sufficient* one. Similarly, evaluating all self-production possibilities is required in order to derive bounds on subsidy-free revenues that are both necessary and sufficient. Expressions (8.4)-(8.6) do not include all possible self-production cases. For instance, they do not account for a consumer existing in period 1 who constructs q units of capacity in period 1 to meet its demand in period 1 only. However, this possibility, represented by $P_1q < F+K(q)$ is

clearly dominated by other self-production possibilities (e.g., 8.5). Consequently, although all possible cases should be examined, not all self-production cases will actually contribute to the bounds, since some will be dominated by lower SAC self-production possibilities.

By using Faulhaber’s (1975) approach (§4.3.2) to deriving bounds on subsidy-free revenues—which has been demonstrated in earlier examples (§4.3.4, §5.2.2 and §7.3.1)—necessary subsidy-free bounds on the revenues obtained from the two products are presented in (8.8) and (8.9). The left-hand term of each expression is the intertemporal (per unit) IC associated with that product, and the right-hand term is the product’s intertemporal (per unit) SAC. The bounds are only necessary conditions for *intertemporal cross subsidies* not to exist; they are not sufficient because these bounds must always be considered in combination with the total revenue/cost equality (8.4). For instance, once the revenue obtained from product 1 has been selected by the monopolist, consistent with (8.8), the required revenue for product 2 is then found through substitution of the selected product 1 revenue into (8.4). Expression (8.9) simply provides the bounds on the outcome of that substitution. The reverse applies if the monopolist selects the product 2 revenue first.

$$\frac{1}{q} \left(\left(\frac{\rho_T}{1 + \rho_T} \right) F + K(q) - \frac{K(q, q') - K(q')}{1 + \rho_T} \right) \leq P_1 < \frac{1}{q} \left(F + K(q) - \frac{qK(q')}{q'(1 + \rho_T)} \right) \quad (8.8)$$

$$\frac{1}{q'} K(q') \leq \hat{P}_2 < \frac{1}{q + q'} (F + K(q, q')) \quad (8.9)$$

Further, these bounds are only valid if sustainable revenues are able to exist under conditions of perfect contestability. Therefore, expression (8.7) must also hold. However, in many of the other cases presented later in this Chapter, the sustainability condition effectively reduces to $F \geq 0$, or F greater than a negative value. When this occurs, sustainable revenues can *always* exist with respect to all products (since F is assumed to be non-negative).

8.1.4 Intertemporal Subsidy-Free Revenues and the Sequencing of Demands Example

It is enlightening to compare the intertemporal bounds on subsidy-free revenues (8.7)-(8.9) with expressions (5.5)-(5.7) in the *static* two good example (§5.2.4). Whereas the fixed costs (F) do not appear in the incremental cost equations of the subsidy-free bounds in the static example, they do in the intertemporal example (at least for the first product). Effectively the result above can be seen as a reworking of the second scenario of Baumol and Sidak’s (1994a, p. 69) sequencing of demands example (§5.3.1 and §6.2.1). Baumol and Sidak state that regardless of which product is consumed first, neither product’s incremental costs contribute to these fixed costs. Since incremental costs are by definition subsidy-free, this would mean that it is “fair” for the revenues collected from the sale of one (but not both) products to include no fixed cost contribution. The above example demonstrates that this is not correct for the product consumed first. The result above also reinforces that incremental costs should

contribute to joint costs which are *not* fixed. The cost of the joint capacity K —which is not fixed, but is subject to increasing returns to scale—is included in the incremental costs of both products in both the static and the intertemporal examples.

Consumers of the first product tie up resources incurred in both the fixed cost and capital cost outlay. In the words of Alfred Kahn, there is a clear cost “causation” involved (§7.5.4). If consumers of the first product chose to delay their consumption, then the fixed costs would not be incurred until a later date. There is an opportunity cost associated with meeting demand now, and a possible option value associated with deferring demand until later (§5.3.4).

8.1.5 Sustainable and Subsidy-Free Revenues under Anticipatory Construction Optimality

Although BPW discuss anticipatory construction, they do not present the sustainability condition or subsidy-free revenue bounds for the case where such a construction configuration is optimal, since their key interest is in demonstrating that unsustainability can occur in an expanding industry. Consequently, their explanation for unsustainability problems under capacity expansion is somewhat perplexing. BPW (p. 473) attribute unsustainability to “the opportunity cost of tying up resources by building spare capacity in anticipation of the demand for capacity in the future”. However, under capacity expansion, capacity is expanded at the time of future demand growth. Consumers in the first period do tie up resources relating to both fixed costs and capacity costs, but not in expectation of later demand. In fact, under capacity expansion, the costs of capacity cannot even be considered “joint” (§3.4.2), since they are clearly attributable to the two products. The only costs which are joint are the fixed costs (F).

It is under anticipatory construction where some spare capacity exists because it is optimal (§6.1.1), and the cost of capacity is “joint” to both products. The approach to deriving subsidy-free revenues under anticipatory construction is very similar to that taken for capacity expansion. However, the results are markedly different. For anticipatory construction to be optimal requires expression (8.3a) to fail, and the total revenue/cost equality is now derived from (8.2) as shown in (8.10) below. It is noteworthy that, under anticipatory construction optimality, a natural monopoly will always be the least cost supplier of this single-good market, even if the fixed costs incurred to establish the firm are zero.

$$P_1q + \frac{\hat{P}_2(q + q')}{1 + \rho_T} = F + K(q, q') \quad (8.10)$$

With the exception of a firm or self-producer supplying the entire market, for which equation (8.10) would apply, the possibilities for entry (or for self-production) are the same as under capacity expansion. Consequently, a necessary and sufficient condition for the existence of sustainable prices in a perfectly contestable market is that presented in (8.11a) in terms of F , and rearranged in terms of ρ_T in (8.11b). Unlike (8.7), the right-hand side of (8.11a) can go negative. Further, when F is zero, the right-

hand side of expression (8.11b) is positive. Therefore it can be seen that if period 1 is sufficiently short, sustainable revenues will always exist under anticipatory construction, and as such this construction configuration is inherently more sustainable than capacity expansion.

$$F \geq \frac{1}{q'} [(q + \rho_T(q + q'))K(q, q') - (1 + \rho_T)(q + q')K(q)] \quad (8.11a)$$

$$\rho_T \leq \frac{(q + q')K(q) - qK(q, q') + q'F}{(q + q')[K(q, q') - K(q)]} \quad (8.11b)$$

This result is consistent with BPW's "intuitive explanation of intertemporal unsustainability" (BPW, pp. 420-421). They indicate that strong cost complementarity in production is the key influence ensuring that the production of a multiplicity of outputs is most efficiently undertaken by a single firm. Further, there are few cases offering complementarities as strong as those between the production of current and future output, when both are produced by the same firm utilising much of the same plant in both periods. BPW suggest that one would expect this to give an incumbent firm a commanding advantage over any firm choosing to supply output in only one of the two periods. Nevertheless, they point out that this complementarity advantage can be overtaken and overwhelmed by the growing opportunity cost over time during which resources are tied up, by being constructed in advance.

However, BPW's explanation is considerably more relevant to anticipatory construction than to capacity expansion, since cost complementarity in the capacity expansion case is limited to the fixed costs. Where anticipatory construction is optimal—therefore some spare capacity is present in period 1—there is a trade-off between the complementarity advantage from the economies of scale inherent in the capacity cost function, and the cost of time. Consequently, even where anticipatory construction is the least cost configuration, the greater the time interval between initial asset construction and subsequent demand growth, the greater the cost of time. Eventually this opportunity cost becomes sufficiently high for a firm producing in both periods to become vulnerable at its break-even product revenue levels to takeover of some or all of its market by a firm producing only in period 2 (BPW, p. 421).

Necessary subsidy-free revenues under anticipatory construction are shown in (8.12) and (8.13)—again assuming that the sustainability condition holds; (8.11a) in this case—derived from (8.5), (8.6) and (8.10). If F is zero, but the time interval between initial asset construction and subsequent demand growth is sufficiently small—it satisfies (8.11b)—then the range between unit IC and unit SAC for each product provides some flexibility to the monopolist in setting its revenue requirements, while still making a normal profit.

$$\frac{1}{q} \left(\left(\frac{\rho_T}{1 + \rho_T} \right) F + \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \right) \leq P_1 < \frac{1}{q} \left(F + K(q) - \frac{q[K(q, q') - K(q)]}{q'} \right) \quad (8.12)$$

$$\frac{1}{q'}(1 + \rho_T)[K(q, q') - K(q)] \leq \hat{P}_2 < \frac{1}{q + q'}(F + K(q, q')) \quad (8.13)$$

8.1.6 Anticipatory Construction with Fungible Assets

BPW assume that the assets in their model are (perfectly) non-fungible *outside* the industry. Nevertheless, given the assumptions of BPW's model, under capacity expansion the assets are actually effectively non-fungible *within* the industry as well. This is because, as BPW demonstrate, if revenues are unsustainable when capacity expansion is optimal, there is no feasible resale price between firms for an asset constructed at the beginning of period 1, and put up for sale at the end of that period (BPW, p. 424).

The situation changes dramatically, however, under anticipatory construction optimality. A potentially dominant entry strategy might be for a firm to build sufficient capacity for period 2 (i.e., $q+q'$) at the beginning of period 1, supply the period 1 demand (i.e., q), but then at the end of period 1 sell its asset to a new firm wanting to supply the entire market in period 2 only. Hence, even if a firm is a natural monopoly, that does not preclude firm or asset ownership changes over time. A feasible “resale” or “salvage” price $S(q, q')$ does exist at the end of period 1, as long as expression (8.14) is satisfied. The resultant entry/self-production constraint and resale price are shown in (8.15a) and (8.15b) respectively. This approach is similar to that taken in the earlier “no demand growth” model of intertemporal cross subsidy (§7.3.1).

$$\rho_T \leq \frac{(q + q')K(q) - qK(q, q')}{(q + q')[K(q, q') - K(q)]} \quad (8.14)$$

$$P_1 q \leq F + K(q, q') - \frac{S(q, q')}{1 + \rho_T} \quad (8.15a)$$

$$\text{where: } S(q, q') \equiv K(q, q') \quad (8.15b)$$

However, this entry strategy is no more dominant than that of a Type 1 or Type 2 entrant. Therefore, the sustainability conditions in (8.11a) and (8.11b) do not change. On the other hand, if this entry constraint is considered a self-production constraint—one relating to a self-producer that builds sufficient capacity for period 2 at the beginning of period 1, supplies its own demand in period 1, and then sells the asset to the self-producing consumers in period 2—the bounds on subsidy-free revenues become more restrictive if resale is feasible, than the case where the asset is assumed to be perfectly non-fungible. The new subsidy-free revenue bounds are presented in (8.16) and (8.17). (If resale is infeasible, then the subsidy-free bounds in (8.12) and (8.13) still apply). However, although the sustainability condition does not change, and therefore the firm is inherently no less sustainable, under perfect contestability the bounds on sustainable revenues and subsidy-free revenues are the same. Therefore, a monopolist that is sustainable where assets are perfectly non-fungible can still be sustainable

should assets become fungible within the industry. However, this monopolist will have less flexibility in selecting required revenues for each product, since the range of subsidy-free and sustainable revenues is much narrower.

$$\frac{1}{q} \left(\left(\frac{\rho_T}{1 + \rho_T} \right) F + \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \right) \leq P_1 < \frac{1}{q} \left(F + \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \right) \quad (8.16)$$

$$\frac{1}{q + q'} K(q, q') \leq \hat{P}_2 < \frac{1}{q + q'} (F + K(q, q')) \quad (8.17)$$

It is interesting to note that the resale feasibility condition (8.14) is the same as the sustainability condition (8.11b), if F is set to zero. In fact, this is the case under any of the construction optimality and demand cases examined in this Chapter, as long as sustainability still holds for zero fixed costs. Consequently, resale feasibility is a sufficient, but not necessary, condition for the existence of sustainable revenues (given that F is always non-negative), and sustainability is a necessary, but not sufficient condition, for resale to be feasible. Furthermore, the sustainability condition is independent of whether assets are considered fungible or non-fungible.

This is a marked departure from BPW's formulation, since it introduces the notion of *net intertemporal SAC* (NSAC) and *net intertemporal IC* (NIC), as presented earlier (§7.3.1). Assuming that assets are even partially fungible can considerably tighten the bounds on subsidy-free and sustainable revenues. In fact, in the presence of no fixed costs, only a *single* revenue value associated with each product is free of cross-subsidies and is sustainable—since the left-hand sides and right-hand sides of (8.16) and (8.17) are the same if F is zero. Where resale is feasible, the net intertemporal stand alone cost of each product is equivalent to its net intertemporal incremental cost; a result of great significance in the search for efficient prices that ensure cost recovery.

As long as a resale value exists, it is not important what entity, inside or outside the industry, purchases those assets. Fungibility can arise both from alternative *users* and not just alternative *uses* (§7.3.4). Allowing for asset resale between periods is also consistent with the assumptions of a perfectly contestable market. Moreover, if assets can be resold, then sunk costs no longer remain. In fact, placing a notional restriction on resale, where resale is in fact feasible, could itself be seen as violating the benchmark of perfect contestability, since such a restriction is inconsistent with the assumption of no barriers to entry. Supply or self-production with resale is simply another “entry” mechanism which constrains a consumer's willingness-to-pay for a product, and consequently their perceived (i.e., net) stand alone cost of consuming that product (§5.1.1). As noted earlier (§4.3.6), Sidak and Spulber (1997, p. 341) have observed that requiring prices to be below SAC is “superfluous”, because in a competitive market prices could never be above SAC. However, a clear definition of SAC in determining price ceilings still needs to be outlined. It is the assertion of this thesis that the appropriate value of “SAC”

from a consumer perspective is the intertemporal stand alone cost of self-production *net* of resale to subsequent consumers (where feasible). Even if notional consumer willingness-to-pay exceeds this net SAC value, a rational consumer would not be willing to pay more than the lower value, as long as that consumer: (i) has access to perfect cost information; (ii) experiences no disutility from self-producing in a consumer coalition; and (iii) transaction costs involved in forming the coalition are negligible.

8.2 An Extended Two-Good/Two-Period (TGTP) Model of Natural Monopoly

8.2.1 Extending BPW's Model to Two Goods

BPW's model relates to a single good, whereas the extension of the model presented in this and the following sections relates to two goods; one for which the demand remains constant in both periods, and a second good which is consumed in the second period only. The two goods, however, are considered similar enough that the units of their demands can be added; consequently, the model relates particularly well to goods which are identical in all respects, except the time and *location* at which they are consumed. Such a two-good/two-period (TGTP) model can be representative of a simple power distribution network comprising *three* products, one in the first period, and two in the second period. (For example, the capacity could be a zone substation supplying two downstream loads in different locations, with the second load only requiring supply in the second period).³ An additional assumption is included in the model—the initial demand in period 1 is greater than the demand growth (i.e., the *second* good) in period 2, and thus $q > q'$. This is more restrictive than BPW's formulation, but does not seem an unreasonable assumption for modelling a distribution network, where initial demand for capacity—such as at a new zone substation—is likely to be high (§6.2.2).

New total cost equations for anticipatory construction and capacity expansion, (8.18) and (8.19), simply recognise that there are now two products in period 2, and two associated revenues, P_2 and P_2' . However, the capacity expansion optimality condition remains the same as for the single-good model (8.3a).

$$P_1q + \frac{P_2q + P_2'q'}{1 + \rho_T} = F + K(q, q') \quad (8.18)$$

³ The assumption that unit demands can be directly added is not a necessary aspect of the model, but does serve to simplify it. The model can be extended to look at the case where the joint capacity required to supply the demand for both goods, and not just the *cost* of capacity, is also subject to increasing returns to scale. Such a situation is typical in a power distribution network due to the effects of diversity (i.e., the peak demand for both goods never occurs simultaneously). Furthermore, transformer costs, and also zone substation costs, typically satisfy the declining average costs assumption (§3.6). Supplying downstream loads would require additional investment than just the joint substation costs, but these additional costs would be dedicated specifically to the two loads and as such could be directly added to the net IC and net SAC bounds. This would neither change the sustainability condition, nor the magnitude of the range between subsidy-free revenue bounds (although the *magnitude* of the bounds themselves would change).

$$P_1q + \frac{P_2q + P_2'q'}{1 + \rho_T} = F + K(q) + \frac{K(q')}{1 + \rho_T} \quad (8.19)$$

Similar to the single-good model, entry possibilities still include Type 1 and Type 2 entrants under either construction configuration. However, there are a number of other entry/self-production possibilities relevant to either the determination of the sustainability condition, or to the determination of the bounds on subsidy-free prices. The first, common to both construction configurations when assets are considered to be perfectly non-fungible, is termed “*Type 3 entry*” (8.20). Other constraints are: (8.21) applicable to both construction configurations if assets are perfectly non-fungible; (8.22a)-(8.23b) for anticipatory construction where assets are fungible; and (8.24a)-(8.25b) for capacity expansion where assets are fungible.

$$P_1q + \frac{P_2'q'}{1 + \rho_T} \leq F + K(q) \quad (8.20)$$

$$P_2'q' \leq F + K(q') \quad (8.21)$$

$$P_1q \leq F + K(q, q') - \frac{S(q, q')}{1 + \rho_T} \quad (8.22a)$$

$$\text{where: } S(q, q') \equiv K(q, q') \quad (8.22b)$$

$$P_1q + \frac{P_2'q'}{1 + \rho_T} \leq F + K(q, q') - \frac{S(q)}{1 + \rho_T} \quad (8.23a)$$

$$\text{where: } S(q) \equiv (1 + \rho_T)K(q) - \rho_T K(q, q') \quad (8.23b)$$

$$P_1q \leq F + K(q, q') - \frac{S(q, q')}{1 + \rho_T} \quad (8.24a)$$

$$\text{where: } S(q, q') \equiv K(q, q') \quad (8.24b)$$

$$P_1q + \frac{P_2'q'}{1 + \rho_T} \leq F + K(q, q') - \frac{S(q)}{1 + \rho_T} \quad (8.25a)$$

$$\text{where: } S(q) \equiv (1 + \rho_T)K(q) - \rho_T K(q, q') \quad (8.25b)$$

8.2.2 Sustainability and Resale Feasibility in the Two-Good Model

Sustainability conditions under the two-good model, for either fungible or non-fungible assets, are (8.26a) or (8.26b) for anticipatory construction, and (8.27) for capacity expansion.

$$F \geq (1 + 2\rho_T)K(q, q') - 2(1 + \rho_T)K(q) \quad (8.26a)$$

$$\rho_T \leq \frac{2K(q) - K(q, q') + F}{2[K(q, q') - K(q)]} \quad (8.26b)$$

$$F \geq 2K(q') - K(q, q') \quad (8.27)$$

Interestingly, the necessary and sufficient sustainability condition for capacity expansion (8.27), is simply a *sufficient*, but not necessary, condition for sustainability under anticipatory construction. Therefore, just as in the one-good model, anticipatory construction is shown to be inherently more sustainable than capacity expansion. But of even more interest is that (8.28) provides a *sufficient sustainability condition* for both construction configurations, indicating that some capacity cost functions will always be sustainable, even for zero fixed costs under capacity expansion—an unsustainable situation in the single-good model. Consequently, it can be seen that *the two-good model is inherently more sustainable than the single-good model*.

$$2K(q') \leq K(q, q') \quad (8.28)$$

The resale feasibility expressions for anticipatory construction and capacity expansion provide another set of sufficient conditions for sustainability. Since, like the single-good case, these are simply found by setting the value of F in the relevant sustainability condition to zero, these can be found by substituting zero for F in (8.26b), under anticipatory construction, and directly from (8.28), under capacity expansion.

8.2.3 Subsidy-Free Revenues

Necessary subsidy-free revenue conditions are somewhat more complex under the two-good model. In particular, when assets are non-fungible, single-product SAC constraints, such as (8.21), potentially become dominant for certain capacity cost functions, providing a complex subsidy-free revenue region. For the two-good model, subsidy-free bounds where assets are considered non-fungible are *not* the same as subsidy-free bounds where assets are considered fungible but resale is infeasible, unlike the corresponding single-good cases.⁴ But since non-fungible cases explicitly disallow the possibility of resale, and are therefore considered to violate the perfect contestability assumption (§8.1.6), in this thesis they are considered to be of less interest. However, the subsidy-free revenue bounds under anticipatory construction, where assets are potentially fungible, are presented in (8.29)-(8.31), where resale is feasible, and (8.32)-(8.34) where resale is infeasible.

$$\left(\frac{\rho_T}{1 + \rho_T} \right) F + \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \leq P_1 q < F + \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \quad (8.29)$$

$$(1 + \rho_T)K(q) - \rho_T K(q, q') \leq P_2 q < F + (1 + \rho_T)K(q) - \rho_T K(q, q') \quad (8.30)$$

⁴ This is because the single-product SAC constraints that affect the sustainable and subsidy-free region where assets are considered non-fungible, all occur in the region where constraints due to feasible resale dominate when the asset is considered fungible.

$$(1 + \rho_T)[K(q, q') - K(q)] \leq P_2'q' < F + (1 + \rho_T)[K(q, q') - K(q)] \quad (8.31)$$

$$\left(\frac{\rho_T}{1 + \rho_T}\right)F + \left(\frac{\rho_T}{1 + \rho_T}\right)K(q, q') \leq P_1q < F + 2K(q) - K(q, q') \quad (8.32)$$

$$(1 + \rho_T)[K(q, q') - K(q)] \leq P_2q < F + (1 + \rho_T)K(q) - \rho_T K(q, q') \quad (8.33)$$

$$(1 + \rho_T)[K(q, q') - K(q)] \leq P_2'q' < F + (1 + \rho_T)K(q) - \rho_T K(q, q') \quad (8.34)$$

The corresponding bounds under capacity expansion are (8.35)-(8.37) and (8.38)-(8.40) where resale is feasible and infeasible respectively.

$$\left(\frac{\rho_T}{1 + \rho_T}\right)F + K(q) - \frac{K(q, q') - K(q')}{1 + \rho_T} \leq P_1q < F + K(q) - \frac{K(q, q') - K(q')}{1 + \rho_T} \quad (8.35)$$

$$K(q, q') - K(q') \leq P_2q < F + K(q, q') - K(q') \quad (8.36)$$

$$K(q') \leq P_2'q' < F + K(q') \quad (8.37)$$

$$\left(\frac{\rho_T}{1 + \rho_T}\right)F + K(q) - \frac{K(q, q') - K(q')}{1 + \rho_T} \leq P_1q < F + K(q) - \frac{K(q')}{1 + \rho_T} \quad (8.38)$$

$$K(q') \leq P_2q < F + K(q, q') - K(q') \quad (8.39)$$

$$K(q') \leq P_2'q' < F + K(q, q') - K(q') \quad (8.40)$$

These single-product revenue bounds, (8.29)-(8.40), are only *necessary* conditions for subsidy-free revenues, they are not *sufficient*. For example, if capacity expansion is optimal, and resale is feasible, expression (8.36) indicates that it is possible for an incumbent monopolist to recover (per unit) revenue P_2 from consumers of the original good in period 2 up to an amount $NSAC_2$ (i.e., $F + K(q, q') - K(q')$), without any cross subsidies existing (although any revenue obtained above this level from these consumers is a clear indication of cross subsidy). Nevertheless, as soon as the revenue target for that particular product is selected, this immediately constrains further the revenues that can be obtained for the other two products (i.e., P_1 and P_2'), since subsidy-free conditions also require total revenue to equal total cost (8.18). As Faulhaber (1975) pointed out in his three-product (but single-period) example, the stand alone and incremental revenues must also be compared with the stand alone and incremental costs for each *pair* of products, as well as to the total revenue/total cost constraint, for sufficiency to be achieved. The tests are combinatorial (§4.3.4). Consequently, there are a whole corresponding set of *dual-product* subsidy-free revenue bounds. However, these bounds are not presented here, simply because, for a three-product model like the one presented here, they can be readily derived from the total revenue/cost equality, and the single-product revenue bounds.

8.2.4 *The Difference Between Sustainable and Subsidy-Free Revenues*

The two-good model, like the single-good model, demonstrates that, where resale is infeasible or fixed costs are high, an incumbent monopolist has quite a high degree of flexibility in setting revenue targets. However, a two-good market under the presented assumptions is more likely to be sustainable than a single-good market. Nevertheless, where resale is feasible, and fixed costs are zero, NIC equals NSAC for all products, and the monopolist is restricted to setting a *single* sustainable and subsidy-free revenue value for each product, whether the market comprises one or two goods.

But it is also important to recognise that subsidy-free revenue bounds should always be considered within a perfectly contestable framework. Consequently, although subsidy-free revenues are also sustainable under assumptions of perfect contestability, real barriers to entry (or other restrictions on firm behaviour) may change the actual sustainability of a set of revenues, whereas the bounds on subsidy-free revenues do not change. For example, if the three-product monopolist in this case were unable to *distinguish* between the demand of the two different goods at different locations, the monopolist would have no choice but to charge the same price to consumers of both goods (§5.2.2 and §8.4.1). (In a real network industry, inappropriate metering of downstream capacity might cause such a situation). As a result, if capacity expansion were optimal, and (8.28) satisfied, a monopolist incurring zero fixed costs would actually become unsustainable should prices to all consumers be restricted to being the same in the second period. However, the fundamental structure of the model has not changed. There are still two goods comprising three products, it is simply that the monopolist is now constrained to charge only two different revenues because consumers are only distinguishable by the *period* in which their demands occur. (This is expanded upon below; §8.4.1). The artificial restriction on price-setting behaviour decreases sustainability; but the subsidy-free revenue bounds remain unchanged.

8.3 **Relaxing the Finite Demand Assumption: “Perpetual Demand”**

8.3.1 *Total Costs and Construction Optimality under “Perpetual Demand”*

BPW’s finite demand assumption (§8.1.1) requires that assets in the model are sufficiently durable to operate with undiminished capacity for the two periods of the analysis. Relaxing this assumption, and assuming instead that demand continues to exist in perpetuity (hereafter termed the “*perpetual demand assumption*”) has a significant impact on characteristics of the model. Firstly, there are a completely new set of construction optimality conditions, involving a third possibly optimal construction configuration. Secondly, there are a new set of NSAC and NIC bounds associated with each product, since both cost characteristics, demand characteristics and the resale value of assets, all change.

The first effect is that a new construction configuration, other than capacity expansion or anticipatory construction, becomes potentially optimal. This new configuration is *capacity replacement* (§6.2.1) and occurs when a least cost construction program consists of initially constructing capacity at a cost of $K(q)$, and subsequently constructing new capacity, at a cost of $K(q,q')$ at the beginning of

period 2, to serve the entire market. It is assumed that period 2 begins *before* the end of the original asset's useful lifetime. This is an interesting outcome, because such a configuration might be considered to involve a duplication of assets. An existing asset with remaining service potential is being scrapped in favour of a new asset. Nevertheless, as discussed earlier (§6.2) and re-verified below, discarding useful capacity can in fact be optimal. Consequently, as discussed earlier (§3.2.1)—and as was recognised by Broadman and Kalt (1989)—duplication of assets may be part of an optimal investment program.

New total revenue/cost equations for anticipatory construction (8.41), capacity expansion (8.42), and capacity replacement (8.43) respectively (under the two-good/two-period model) are given below. The new term that appears in all total revenue/cost equations is B_N , the uniform series present worth factor (6.1) for a period of N years, which reflects the need to replace assets indefinitely as they come to the end of their lifetime (i.e., after N years). Because demand is perpetual, the capacity constructed at the beginning of period 2 must be renewed at the end of its lifetime if demand is to continue to be served. The other construction configurations also require a program of asset renewal in perpetuity.

$$P_1q + \frac{P_2q + P_2q'}{1 + \rho_T} = F + \frac{1}{dB_N} K(q, q') \quad (8.41)$$

$$P_1q + \frac{P_2q + P_2q'}{1 + \rho_T} = F + \frac{1}{dB_N} \left[K(q) + \frac{K(q')}{1 + \rho_T} \right] \quad (8.42)$$

$$P_1q + \frac{P_2q + P_2q'}{1 + \rho_T} = F + K(q) + \frac{1}{dB_N} \cdot \frac{K(q, q')}{1 + \rho_T} \quad (8.43)$$

Under *constant* returns to scale, whether or not demand is finite or perpetual, capacity expansion is always the optimal construction configuration, and sustainable revenues always exist, given the assumptions of perfect contestability. For increasing returns to scale, anticipatory construction is more likely to be optimal should demand growth occur in relatively early years. Under perpetual demand however, increasing returns to scale also have the effect of making capacity replacement a likely optimal configuration, should demand growth not ensue until nearer the end of the original asset's lifetime. The new construction optimality conditions under perpetual demand for anticipatory construction (8.44), capacity expansion (8.45), and capacity replacement (8.46) respectively, are as follows.

$$\rho_T \leq \min \left(\frac{K(q') - [K(q, q') - K(q)]}{K(q, q') - K(q)}, \frac{\rho_N K(q)}{(1 + \rho_N)K(q, q') - \rho_N K(q)} \right) \quad (8.44)$$

$$\frac{K(q') - [K(q, q') - K(q)]}{K(q, q') - K(q)} < \rho_T \leq \frac{(1 + \rho_N)[K(q, q') - K(q)] - K(q)}{K(q)} \quad (8.45)$$

$$\rho_T > \max \left(\frac{(1 + \rho_N)[K(q, q') - K(q)] - K(q)}{K(q)}, \frac{\rho_N K(q)}{(1 + \rho_N)K(q, q') - \rho_N K(q)} \right) \quad (8.46)$$

The transition time $T_{AC \rightarrow CE}$ between anticipatory construction optimality and capacity expansion is unaffected by the relaxation of the finite demand assumption, and is still as is shown in (8.3b). The corresponding transition conditions between capacity replacement and the other two construction configurations are as shown in (8.47a)-(8.48b) below

$$\rho_{AC \rightarrow CR} \equiv \frac{\rho_N K(q)}{(1 + \rho_N)K(q, q') - \rho_N K(q)} \equiv (1 + d)^{T_{AC \rightarrow CR}} - 1 \quad (8.47a)$$

$$T_{AC \rightarrow CR} \equiv \log_{(1+d)} \left(1 + \frac{\rho_N K(q)}{(1 + \rho_N)K(q, q') - \rho_N K(q)} \right) \quad (8.47b)$$

$$\rho_{CE \rightarrow CR} \equiv \frac{(1 + \rho_N)[K(q, q') - K(q')] - K(q)}{K(q)} \equiv (1 + d)^{T_{CE \rightarrow CR}} - 1 \quad (8.48a)$$

$$T_{CE \rightarrow CR} \equiv \log_{(1+d)} \left(1 + \frac{(1 + \rho_N)[K(q, q') - K(q')] - K(q)}{K(q)} \right) \quad (8.48b)$$

There is an interesting characteristic of these new optimality conditions. For sufficiently large increasing returns to scale, and/or where the assets have relatively short lifetimes, it is possible that capacity expansion is *never* the least cost construction configuration, irrespective of the moment (during the extent of the original asset's lifetime) that the demand growth occurs. Such a result occurred in the example presented earlier (§6.2.4), and is demonstrated in the example plot of optimal construction conditions shown in Figure 6.2.

Feasibility of capacity expansion in the TGTP model requires that the right-hand side of expression (8.45) is greater than the left-hand side, and therefore that the asset lifetime satisfies expression (8.49) below. Where capacity expansion is feasible, and expression (8.49) holds, the optimality of anticipatory construction or capacity replacement is found from the first term in the respective minimum (8.44) and maximum (8.46) expressions above. Should capacity expansion not be feasible, then the second term in those expressions will be the appropriate minimum or maximum value respectively for testing whether anticipatory construction or capacity replacement is optimal.

$$\rho_N > \frac{K(q) + K(q') - K(q, q')}{\frac{K(q) \cdot K(q')}{K(q, q')} + K(q, q') - K(q) - K(q')} \quad (8.49)$$

8.3.2 Impacts on Stand-Alone Costs, Incremental Costs and Resale Feasibility

As discussed earlier (§5.1.1), SAC and IC are not only defined by the structure of costs (i.e., the nature of the capacity cost function), but also by the state of demand. This complicates the regulatory determination of efficient and subsidy-free prices, and typically regulators derive bounds on prices based on cost data alone, leaving the “self-interest” of the regulated firm to take demand conditions into

account (Baumol and Sidak, 1994, p. 51; §4.3.1). Consequently, price floors and ceilings are likely to only be necessary, but not sufficient, constraints on efficient prices.

Assuming that demand continues in perpetuity has a big impact on NSAC and NIC, since the asset lifetime now becomes a crucial factor impacting the cost of supply, due to the need for regular asset renewal. Further, the resale value of assets change, since intuitively, a potential purchaser is more likely to exist if demand is expected to continue for the foreseeable future. Therefore, entry strategies and self-production possibilities also change; even in just a three-product model, a very large number of resale possibilities emerge to be investigated, particularly given that there is a new optimal construction configuration to be considered.

Perpetual demand does not affect the standard Type 1 and Type 2 entry strategies substantially, except due to the need to account for the greater level of costs involved in supply due to perpetual asset renewal. However, it does impact the associated resale feasibility conditions and consequently the actual asset resale values, where resale is feasible. Resale is affected by the changed demand conditions, and resale values and resale feasibility conditions now become functions of asset lifetime (indicated through the inclusion of the ρ_N discounting term). For example, the Type 1 entry or self production constraint under anticipatory construction is presented in (8.50a), with the corresponding resale equation in (8.50b). The corresponding resale feasibility condition under anticipatory construction is shown in (8.51). These contrast with the Type 1 entry constraints with resale relevant under finite demand, given in (8.22a) and (8.22b), and with the resale feasibility condition under finite demand, derived from (8.26b) by setting F to zero.

$$P_1 q \leq F + K(q, q') - \frac{S(q, q')}{1 + \rho_T} \quad (8.50a)$$

$$\text{where: } S(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_T}{1 + \rho_N} \right) K(q, q') \quad (8.50b)$$

$$\rho_T \leq \frac{(1 + 2\rho_N)K(q) + K(q') - (1 + \rho_N)K(q, q')}{2(1 + \rho_N)K(q, q') - (1 + 2\rho_N)K(q) - K(q')} \quad (8.51)$$

On the other hand, Type 3 entry is somewhat different under perpetual demand, and the new expression is shown in (8.52). Some spare capacity associated with the original asset becomes available in the second period, until such a time as the original asset arrives at the end of its useful lifetime, at which point a new asset of appropriately-sized capacity is constructed. However, it is not always feasible to sell “access” to this spare capacity. Possible feasible resale scenarios—involving a consumer coalition of first period consumers and consumers of the second good in the second period that sell assets or access to spare capacity to consumers of the first good in period 2—are very dependent on which construction configuration is optimal. For instance, the related NSAC constraint under anticipatory construction is

(8.53a) and (8.53b), and this constraint dominates (8.52).⁵ The resale feasibility constraint is the same as for other entry or self production constraints, as already shown in (8.51).

$$P_1q + \frac{P_2q'}{1 + \rho_T} \leq F + K(q) + \frac{1}{dB_N} \cdot \frac{K(q')}{1 + \rho_N} \quad (8.52)$$

$$P_1q + \frac{P_2q'}{1 + \rho_T} \leq F + \frac{1}{dB_N} K(q, q') - \frac{S^*(q)}{1 + \rho_T} \quad (8.53a)$$

$$\text{where: } S^*(q) \equiv \frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \quad (8.53b)$$

Of particular interest is the fact that, under capacity replacement, the resale value of the asset constructed to serve the market demand of period 1 (i.e., q) is *zero* at the end of that period. This is the key characteristic of a program of capacity replacement. Because the asset is *replaced* before the end of its potential lifetime, it has no value at the beginning of period 2, when an asset of sufficient capacity to serve the entire market demand of period 2 (i.e., q and q') is constructed, (assuming that there is no *external* resale market).

8.3.3 Sustainable and Subsidy-Free Revenues under Perpetual Demand

Where resale is feasible, single-product bounds on subsidy-free revenues under perpetual demand do not differ substantially from the comparable expressions derived under conditions of finite demand. The exception is that NIC and NSAC values must now account for the increased costs due to perpetual asset renewal. For example, subsidy-free bounds, where resale is feasible and anticipatory construction is optimal, are presented in (8.54)-(8.56); these are similar to the corresponding finite demand expressions in (8.29)-(8.31), with the exception of the *asset renewal term* (i.e., $1/dB_N$).

$$\left(\frac{\rho_T}{1 + \rho_T} \right) F + \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \leq P_1q < F + \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \quad (8.54)$$

$$\frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \leq P_2q < F + \frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \quad (8.55)$$

⁵ Usually, a resold asset has a finite lifetime. However, the asterisk by the resale value S in (8.53a) indicates that the resale is for use in perpetuity. Replacement of the asset is actually paid for in perpetuity by the seller, since the resale amount is sufficient to cover the appropriate proportion of the purchaser's future asset renewal costs (discounted back to the time of resale). Since the asset costing $K(q, q')$ is effectively indivisible, and only a portion of the asset is being sold, S^* can be considered an "access price". Depending on the nature of the asset, providing indefinite access to a portion of an indivisible asset does not necessarily cause a problem in reality. For instance, if the asset were a zone substation supplying two downstream lines/feeders, the line capacity could be constrained to ensure that demand does not exceed purchased capacity (e.g., through an appropriate installed capacity for the feeder, or through metering that activates a current limiter).

$$\frac{1}{dB_N}(1 + \rho_T)[K(q, q') - K(q)] \leq P_2'q' < F + \frac{1}{dB_N}(1 + \rho_T)[K(q, q') - K(q)] \quad (8.56)$$

However, where resale is infeasible, the subsidy-free bounds differ substantially between the finite and perpetual demand cases. For example, when anticipatory construction is optimal, but resale is infeasible, the subsidy-free bounds are (8.57)-(8.59), in contrast to (8.32)-(8.34).

$$\left(\frac{\rho_T}{1 + \rho_T}\right)F + \frac{1}{dB_N}\left(\frac{\rho_T}{1 + \rho_T}\right)K(q, q') \leq P_1q < F + K(q) - \frac{1}{dB_N}\left(K(q, q') - K(q) - \frac{K(q')}{1 + \rho_N}\right) \quad (8.57)$$

$$(1 + \rho_T)\left(\frac{1}{dB_N}\left[K(q, q') - \frac{K(q')}{1 + \rho_N}\right] - K(q)\right) \leq P_2q < F + \frac{1}{dB_N}[(1 + \rho_T)K(q) - \rho_T K(q, q')] \quad (8.58)$$

$$\frac{1}{dB_N}(1 + \rho_T)[K(q, q') - K(q)] \leq P_2'q' < F + (1 + \rho_T)K(q) - \frac{1}{dB_N}\left[\rho_T K(q, q') - \left(\frac{1 + \rho_T}{1 + \rho_N}\right)K(q')\right] \quad (8.59)$$

Since sustainability conditions are closely related to the conditions for resale infeasibility, the sustainability conditions are also very different under perpetual demand. For example, contrast the sustainability condition for anticipatory construction under perpetual demand, (8.60a) or (8.60b), with the corresponding condition under finite demand, (8.26a) or (8.26b). As before, the resale feasibility condition (8.51) can be derived from the sustainability condition (8.60b) by setting F to zero.

$$F \geq \frac{1}{dB_N}\left[(1 + 2\rho_T)K(q, q') - \frac{1 + 2\rho_N}{1 + \rho_N}(1 + \rho_T)K(q) - \frac{1 + \rho_T}{1 + \rho_N}K(q')\right] \quad (8.60a)$$

$$\rho_T \leq \frac{(1 + 2\rho_N)K(q) + K(q') - (1 + \rho_N)K(q, q') + \rho_N F}{2(1 + \rho_N)K(q, q') - (1 + 2\rho_N)K(q) - K(q')} \quad (8.60b)$$

Capacity expansion is still a less sustainable configuration than anticipatory construction, and may require non-zero fixed costs for sustainability, under some capacity cost functions and demand growth scenarios. But again, unlike the single-good case, sustainability does not *always* require non-zero fixed costs when capacity expansion is optimal. However, the biggest difference between the perpetual and finite demand cases is the presence of a possible new optimal construction configuration, and the possibility that capacity expansion is *never* the least cost configuration. Capacity replacement is an interesting case, since when this configuration is optimal, sustainable revenues are more likely to exist when demand growth occurs toward the *end* of the original asset's lifetime, rather than the beginning, as is the case for either anticipatory construction or capacity expansion. This is seen in the sustainability condition for capacity replacement below (8.61).

$$\rho_T > \frac{(1 + \rho_N)K(q, q') - K(q') - K(q) - \rho_N F}{K(q) + K(q')} \quad (8.61)$$

8.4 Intertemporal Unsustainability Revisited

8.4.1 Intertemporal Unsustainability and the Distinguishability of Goods and Consumers

It is clear then that the key reason for the intertemporal unsustainability exhibited in BPW's model of capacity expansion is their implicit assumption that their model relates to a *single* good comprising *two* products: the first product being the capacity demanded in period 1; and the second being the greater level of capacity demanded in period 2. This means that, during period 2, the "price" (i.e., the entire-period revenue) assigned to the demand for the new capacity constructed at the beginning of period 2, cannot be different from the price assigned to the demand for the capacity constructed at the beginning of period 1. Demand for each unit of either capacity is priced the same. But different unit *costs* are associated with the new and the old capacity, both in a static *and* an intertemporal sense. Consequently, demand for the new and old capacity in period 2 could be considered as demand for two distinctly different goods. As such, BPW's model more correctly relates to a two good model comprising *three* products, the same as in the extended BPW model introduced in Section 8.2. Different prices can be assigned to the demand for each of those three products, including demand for both new and old capacity in period 2, without price discrimination arising.

It was discussed above (§8.2.4) that, where capacity expansion is optimal, and the sustainability condition (8.28) satisfied, a monopolist incurring zero fixed costs would actually become unsustainable should prices for both goods be restricted to being the same in the second period. The converse is true for BPW's single good model. Although BPW's original model is always unsustainable in the presence of zero fixed costs, if different prices could be charged for the demand in period 2 associated with the new and old capacity (i.e., for the two different products), then capacity cost functions satisfying (8.28) would always be sustainable. Sustainability depends on the ability of the monopolist to charge different prices for different products.

As was touched on earlier (§5.2.2), pricing problems arise in practice where different products, goods or consumers are not "distinguishable". If capacity is a joint input to a product, essential at a particular time for the production or delivery of a commodity (or perhaps downstream capacity in a network), then consumers will be indifferent as to whether the commodity is jointly associated with the new or the old capacity.⁶ Further, if it is not possible to distinguish whether that commodity is associated

⁶ BPW describe demand as being for a non-storable good (or service) which cannot be produced without a capital outlay for a durable facility of a certain capacity (BPW, p. 407). The model developed in this thesis presents the demand for the non-storable good as separable from the demand for the capacity of the joint durable good required to produce that non-storable good. The focus is on the demand for *capacity* (at a particular location and time), rather than for the commodity associated with that capacity (§3.1.2). This approach has been taken to allow a closer alignment of the model to the purchase of electricity, where the commodity is energy, and the capacity is the connection capacity sufficient to meet peak power

with either the new or old capacity, then it is not possible to charge a different price for the demand for either capacity. And even if it might be possible to distinguish with which capacity the commodity is associated, another distinguishability problem could arise if it is not possible to distinguish which consumers demand which commodity.

An example of the latter case is of power supply through a single transformer to an office building, where multiple tenants are not individually metered for their electricity consumption (i.e., the commodity), and the building owner allocates supply capacity costs (and consumption costs) on a per office basis. If a new transformer is installed to meet growing demand, the owner has no choice but to allocate costs by some method other than demand (unless metering is improved), since individual demand levels are unknown, and consumers are not distinguishable. But even if the consumption of individual tenants were metered, how would it be possible to distinguish which of the tenants are really responsible for growing demand? If all tenants have homogeneous electricity consumption patterns, and both the old and the new transformers provide common supply, again the consumers supplied by the new capacity are indistinguishable from the consumers supplied by the old capacity. Where two goods are effectively indistinguishable, or two goods are consumed by indistinguishable consumer groups, in reality the two would have to be treated as a single good.

This problem of distinguishability is not present in the two-good, three-product model developed in the previous sections. Although the model can be considered representative of a simplified distribution network, a network is not the only possible market that could be consistent with such a model. But the problem of distinguishability is not so relevant in a network, since goods and consumers are distinguishable by location, and thus a separate price can be charged for each good during the same period. An additional problem might arise if arbitrage of network capacity between locations were possible, but arbitrage of access to capacity at different locations will incur its own costs and should be easy to prevent in practice.

8.4.2 *Unsustainability under Different Construction Configurations*

Although unsustainability can also occur in the TGTP model under all construction configurations, as well as under capacity replacement if BPW's finite demand assumption is relaxed, the TGTP model has been shown to be inherently more sustainable than BPW's single-good model (§8.2.2 and §8.3.3). Consequently, as very few markets exist where there truly is just a single good, BPW's concern regarding the possibly widespread nature of intertemporal unsustainability (BPW, p. 429) loses much of its impact. This is particularly so for a network industry, where goods are distinguishable.

demand. Such an approach is particularly applicable to the New Zealand electricity industry where the firms which sell energy from electricity, and network connection for peak power, must be legally separate entities (§2.4.5).

Furthermore, in the TGTP model the alternative construction configurations of anticipatory construction and capacity replacement appear inherently more sustainable than the only configuration examined in detail by BPW, namely, capacity expansion. BPW's assumption that demand is finite, or conversely that asset lifetime has no bearing, means that capacity expansion is more likely to be the optimal construction configuration. As demonstrated above (§8.3.1), when this assumption is relaxed, it is possible that capacity expansion is *never* the least cost construction configuration. Consequently, BPW's single-good assumption and finite demand assumption are both strongly contributing factors to the likelihood of intertemporal unsustainability.

BPW's focus on capacity expansion could well be of less relevance when considering many real world utilities, where anticipatory construction (or even capacity replacement) can frequently provide the optimal configuration (§6.2). Certainly, utilities will need to expand their capacity, but as has been discussed earlier (§3.1.2), network industries often expand more to serve new consumers in new locations (i.e., *greenfields* development) than due to increased demand from existing consumers (i.e., system augmentation). In choosing to expand total capacity to supply these new goods, construction is likely to be anticipatory and (optimally) include spare capacity to allow for future demand growth (§6.1.1). Moreover, anticipatory construction is more likely to be optimal if the current model incorporated into the investment decision the fact that electricity supply is a joint product of connection capacity and energy, rather than just capacity on its own. This is because the cost of losses are subject to diseconomies of scale; hence spare capacity *reduces* the cost of losses (§4.2.6).

8.4.3 The Significance of Sunk Costs and Fixed Costs to Unsustainability

BPW state that, in an intertemporal context, the higher the *sunk costs* involved in the construction of added capacity, the less likely is sustainability, whereas the *fixed costs* incurred in the establishment of a firm tend to improve sustainability (BPW, p. 412). This conclusion, however, is not borne out by the preceding analysis of either BPW's own model, or the extension of their model to two goods.

Sunk costs—in the sense of *ex post* irrecoverable costs (§3.5.4)—can arise in BPW's model of intertemporal unsustainability, but this is *because* the model is unsustainable. Unsustainability causes the irrecoverable sunk (i.e., stranded) costs rather than the other way around. For instance, where there are sufficiently high fixed costs for BPW's single good model to be sustainable—in other words, F satisfies expression (8.7)—the natural monopolist can make a normal profit. As such, all its costs involved in initial capital outlay are fully recoverable; there are no stranded costs. However, should the level of fixed costs be insufficient, and thus no set of sustainable revenues exists, potentially resulting in inefficient entry, the incumbent will not be able to make a normal profit, and as such there will be irrecoverable costs associated with the investment.

BPW attribute their own intertemporal unsustainability result to “sunk” costs—by which they really mean imperfect fungibility—rather than to the constraints they place on an incumbent firm’s ability to distinguish between goods or products and consequently to be able to price them differently. Yet it was demonstrated above (§8.1.6) that, where anticipatory construction is optimal and assets are at least partially fungible—in other words, they have a positive resale value to alternative users—a *tighter* set of bounds on sustainable revenues exists than would be the case if the assets were deemed perfectly non-fungible (i.e., entirely sunk). In the TGTP model, the same result occurs irrespective of which construction configuration is optimal. This is because the ability to resell assets reduces the apparent stand alone cost of supply from a consumer’s perspective to its *net* value.

Although the new analysis provides different conclusions as to the linkage between sunk costs and unsustainability, it does back up BPW’s findings with regard to fixed costs. It has been demonstrated again that fixed costs always improve sustainability, or more precisely, they widen the bounds on sustainable revenues (§8.1.6). However, the fixed costs represented by F are assumed in both the current model and BPW’s model to be perfectly non-fungible. Where conditions of sustainability hold, these fixed sunk costs can also be fully recovered by the sustainable revenues.

The assumption that the fixed costs (F) are perfectly non-fungible (retained throughout this and the following Chapter) is not a particularly limiting assumption, since any *partially fungible* fixed costs—such as the transformer bay, earthing and metering, as in (3.4)—could be subsumed into the construction function K without any impact on the results (as is demonstrated in Gunn, 2002). Since, K is already defined as having decreasing (incremental) average costs, increasing the level of fixed costs *within* K only serves to strengthen the decreasing average cost characteristics of the function.

8.4.4 Unsustainability due to Disequilibrium and Symmetry

The TGTP model can also be thought of as an intertemporal version of Faulhaber’s (1975) presentation of the cross-subsidy problem as a static *co-operative game*, where sustainability is implicitly viewed as an *equilibrium* concept (§4.3.5). Conversely, unsustainability implies the *absence* of equilibrium (BPW, p. 10). Consequently, comments such as unsustainable Ramsey prices can be welfare-superior to sustainable prices (i.e., Teplitz-Sembitzky, 1992. p. iii) have no meaning—Pareto optimality is an equilibrium concept, and by definition, unsustainable prices can never be part of an equilibrium solution (§4.3.6).

Requiring that the potential intertemporal monopolist publish prices for both periods in advance, and not make the initial capital outlay until it has binding contracts with all current and future consumers, would place all current and future consumers (who are also potential self-producers), on a “level playing field” with the potential incumbent monopolist, as well as with all potential entrants. A level playing field can be ensured by requiring that all cards are laid on the table prior to any initial investment, in

other words, all current and future market participants have “perfect information” regarding current and future costs, and that no participants are allowed to subsequently renege on their previously offered (and accepted) prices (i.e., the Bertrand-Nash assumption).

Faulhaber (1975) indicated that when no *core* to his static game exists, then there is no stable supply arrangement. In other words, there is no potential equilibrium.⁷ When no core exists in the intertemporal game, then an incumbent monopolist attempting to supply the entire intertemporal market demand will find that no sustainable set of revenues exists. Consequently, the incumbent will be unable to make a normal profit by recovering its initial capital outlay, and there will be stranded costs. Unless the incumbent can see that all its costs are recoverable *ex ante*, then the investment would not proceed. Therefore, the disequilibrium that arises from playing this intertemporal game could act as a warning to the potential incumbent not to attempt to supply the market in the first place. As BPW (p. 217) note, “if a natural monopoly is unsustainable, then any multi-firm equilibrium is also likely to be unsustainable”. Hence, the multi-firm equilibrium is not really an equilibrium. This may deter entry and leave the incumbent free to manoeuvre after all. As Sassower (1988) points out, Baumol and his colleagues ignore the opportunity costs incurred by a competitor in attempting to enter the market. If unsustainability is really a problem, then potential entrants might recognise that they could become just as unsustainable as the incumbent, should they force the incumbent out of the market, and as such, consider their capital to be better employed elsewhere.

But the lack of an equilibrium might impact any attempt to derive bounds on subsidy-free prices. If no possible equilibrium and no core to the game exists, and no sustainable prices exist, then the inference is that neither can a set of subsidy-free prices. Even though the TGTP model has been shown to be more sustainable than BPW’s original model, there are still many cost functions which are inherently unsustainable under particular demand vectors. Perhaps, however, this problem is less intractable than it first appears.

Problems often arise in resolving the outcomes of static game models due to the implicit or explicit assumption of “symmetry”. If all potential market entrants are “symmetric”, and must make their entry and pricing decisions simultaneously, then the core of the game may well be indeterminate. As Hazledine (1992, pp. 37-38) speculates, oligopoly theory may be harder than it really should be, since the interesting questions in the theory of markets and of the firm boil down to asymmetries between agents, rather than to symmetries. He suggests that the goal should be to “escape from the oligopoly problem”,

⁷ The outcome of the game does not provide the answer as to what the equilibrium revenues actually are, since any set of revenues that satisfies the core of the game can be part of an equilibrium solution. The game simply provides a set of bounds within which sustainable (and subsidy-free) revenues can lie.

and from the dilemmas of strategic interaction between rather similar firms, that have been central to industrial organisation theory.⁸

In the static case, one way of escaping from the problem of symmetry is by deciding *which* of the potential entrants becomes the incumbent, or the market “leader”, since this immediately produces an asymmetry. As Hay and Morris (1993, pp. 94-95) point out, there is a fundamental asymmetry between incumbent and entrant which means that the incumbent is likely to win any game where it needs to “get in first”. By definition, being the “first mover” constrains other market participants to being “followers”.

Clearly, symmetry is of less relevance in an intertemporal context. In the TGTP model, any firm which wins the right to supply market demand at the beginning of period 1 becomes the incumbent firm, and automatically has a “first mover advantage” over potential entrants in period 2, (as well as potential *intra*period entrants *during* period 1; discussed later in §9.1.1). An additional asymmetry arises in the intertemporal world, simply because a firm or self-producer operating in period 1 can potentially sell its assets to firms or self-producers in period 2, whereas of course the converse is not possible. Consequently, there is a fundamental asymmetry in intertemporal models which is not present in static models.

8.4.5 Relaxing the Bertrand-Nash Assumption: Contestability Theory and Strategic Behaviour

At this stage, it is worth examining the problematic Bertrand-Nash assumption, without which conditions of *perfect* contestability do not exist. The applicability of this assumption has been one of the major areas of criticism of contestability theory as a whole (§2.1.7). For example, Hazledine (1992, p. 14) states that empirical examples of contestable markets are rare, since it seems that prices are, in general, not likely to be more difficult to adjust than output, and Shepherd (1984) goes as far as suggesting that the Bertrand-Nash basis of contestability theory appears to rule out the prospect of using the theory to draw any significant normative conclusions.⁹

BPW’s concern regarding the impact of allowing the invisible hand a free reign in an intertemporal context, led them to try and explain away their unsustainability result by highlighting that in the real world, imperfect knowledge and the inapplicability of the Bertrand-Nash assumption may be responsible for the lack of observed real-world market unsustainability (BPW, pp. 426-428). The

⁸ Hazledine (1992, p. 37) notes that: “what oligopoly theorists do is try and predict the outcome of a game between two evenly matched teams. Of course, this is very difficult. It is much easier to predict the outcome if one team is much better than the other, or to not have to predict at all if they are so different that they don’t even have to ‘play’ each other.”

⁹ “If the Bertrand-Nash nonresponse assumption holds, then the model does not admit significant entry, which can influence the incumbent. Total entry, which would entirely duplicate and replace even a monopolist, would be particularly absurd in a Bertrand-Nash model. Yet that case is what ultra-free entry assumes, and where its special superiority over competition in the market is said to occur” (Shepherd, 1984).

problem this provides for the present analysis is that, if relaxing this assumption is required for sustainability to be achieved, then any outcome, although sustainable, will not be consistent with perfect contestability. And if sustainability is required before the bounds on subsidy-free prices can be derived, then those bounds will not be useful as a regulatory benchmark, since they will be inconsistent with perfect contestability.¹⁰

In the TGTP model with perpetual demand, it is fairly straightforward to show that relaxing the Bertrand-Nash assumption, even slightly, will always result in a sustainable outcome.¹¹ Rather than relax it entirely, all that is needed is to allow the potential incumbent to withhold publication of any price until immediately before the period to which that price relates, while still requiring that these prices not be subsequently changed. Hence, the potential incumbent does not need to publish its non-changeable set of prices for *both* periods before entering the market, it only needs to publish its price for the product in period 1 and, assuming that it successfully enters the market, it then does not publish its prices for period 2 products until the very end of period 1. The Bertrand-Nash assumption thus still holds intratemporally within each period, although not intertemporally for the entire period of analysis. However, one might consider that this is actually only a partial relaxation of BPW's implicit "perfect information" assumption rather than a relaxation of Bertrand-Nash, since, although the timing and level of future demand growth is considered known, future prices are not. This is a useful way of looking at the change in assumptions, since perfect knowledge is not stated as being a fundamental premise of perfect contestability.¹²

Given this relaxation of the intertemporal Bertrand-Nash (or perfect information) assumption, a potential Type 1 or Type 3 entrant will never enter the market in period 1 where anticipatory construction is optimal if it does not know what price will be offered by a potential monopolist for either of the two products in period 2. Both types of entrant will recognise that the price they would need to offer in period 2, to forestall entry by a Type 2 entrant, combined with the price offered in period 1 to forestall entry by the potential monopolist (i.e., a revenue less than the NIC of the period 1 product, shown in the

¹⁰ BPW did not derive subsidy-free prices from their single good model, simply because they were concerned with demonstrating that, except in certain circumstances (i.e., with sufficiently high non-fungible fixed costs), their model was actually unsustainable. Thus, implicitly, no subsidy-free prices were considered to exist.

¹¹ BPW themselves go through a similar exercise for their single good model (BPW, pp 426-428).

¹² It is not clear whether BPW's definition of perfect contestability is so strong that it requires what is termed here the "intertemporal Bertrand-Nash assumption"; in other words that all prices for all periods be published in advance of the time of *first* entry, rather than before the time of specific entry relating to a particular price, and then not subsequently changed. For example, the following definition is ambiguous: "each entrant is implicitly assumed to expect that after entry occurs the incumbent will keep his pre-entry prices unchanged for a period sufficiently long to make entry profitable" (BPW, p. 428).

LHS of 8.53), would not allow it to cover its costs. The potential monopolist is thus free to enter.¹³ Because it recognises that other period 1 entrants have no incentive to try and enter, the potential monopolist can safely offer a pre-published price for period 1 up to the NSAC value of the first product (i.e., the RHS of expression 8.53). However, a price higher than this would trigger self-production by period 1 consumers.

Because entry is only potentially forestalled in period 2 by a Type 2 entrant, the now incumbent monopolist is free to charge prices for the two products in period 2 that sum up to the level that would be charged by Type 2 entrant. (This is the same as the NSAC of self-production by all consumers in period 2, and is equivalent to the sum of the LHS of expression (8.54) with the RHS of expression (8.55), or vice versa). Consequently, the revenue received from each product will lie between the bounds presented in expressions (8.54) and (8.55). The relevant bounds are those where it is assumed that resale is feasible. This is because if the potential monopolist does successfully enter in period 1, resale will become potentially feasible after all. Resale only becomes infeasible where it is dominated by the potential offerings of Type 1 and Type 3 entrants. As soon as those types of entrants are out of the picture, a feasible asset transfer price does exist between periods 1 and 2.

As has been discussed earlier (§8.1.6), if fixed costs are zero, then NIC equals NSAC for each product anyway, and there is only one possible sustainable revenue value relating to each product. As such, the incumbent monopolist will earn a normal profit, and the prices offered will be subsidy-free. However, if fixed costs are positive, the incumbent monopolist is able to recover the entire fixed costs in *both* periods, while still being sustainable, and thus make a positive profit. This is because the fixed costs are assumed to be perfectly non-fungible (§8.4.3). Consequently, although relaxing the intertemporal Bertrand-Nash assumption ensures the sustainability of a potential monopolist under anticipatory construction, it does not ensure that prices are subsidy-free. This is because the single-product subsidy-free revenue bounds presented in (8.53)-(8.55) are only necessary, but not sufficient, conditions for no cross-subsidies to exist. As discussed above (§8.2.3), these single-product conditions must be considered in tandem with the relevant total revenue/cost equation (8.41). Since fixed costs can be recovered twice, the total revenue/costs equation is violated, and the monopolist is applying its market power to gain supernormal profits.¹⁴

¹³ As noted above, perfect information of the future timing and level of demand growth is assumed, because otherwise the Type 1 (or 3) entrant might enter after all, not recognising that anticipatory construction is in fact optimal, and it can thus be forced out by the later Type 2 entrant.

¹⁴ It depends whether competition is considered to occur between a *single* potential monopolist and single Type 1, 2 and 3 entrants. Otherwise, it could be argued that if there are a multitude of potential monopolists prepared to enter at the beginning of period 1, then the successful monopolist would have to offer its lowest possible price for the period 1 product, and this would be equivalent to that product's NIC (shown in the LHS of expression 8.53). In such a case, assuming that the

The situation is a little different where either capacity expansion or capacity replacement is optimal. This is because, if prices are not published for period 2 prior to entry at the beginning of period 1, there is little to distinguish potential Type 1 and Type 3 entrants from each other, and from a potential monopolist. This is because any firm would serve the market in period 1 with capacity exactly sized to fit the demand in period 1 alone. This implies that it will not become clear what type of entrant the incumbent firm is until the beginning of period 2, when it will be seen which products it supplies.¹⁵ However, the firm successfully entering in period 1 can only recover its costs if it prices in period 1 like a monopolist, (i.e., the price is at least the NIC of the period 1 product), and in period 2 it supplies the entire market demand. Any possible entrant will recognise that this is the only sustainable entry strategy. Hence, only a firm intending to act as monopolist can (potentially) be sustainable.¹⁶

The preceding subsections have demonstrated that a substantial part of the inherent unsustainability of BPW's single good model results from the assumptions underlying the development of that model (§8.1.1), rather than from the assumptions required for perfect contestability (§2.1.7). Requiring that the price associated with new and old capacity be the same, and assuming that demand ceases before the end of the useful lifetime of any asset, are both strong contributors to intertemporal unsustainability. On the other hand, relaxing these model-specific assumptions does not entirely remove the unsustainability either, whereas partially (or fully) relaxing the Bertrand-Nash assumption does.

The problem is, as Shepherd (1984) puts it, under any departures from the pure conditions of perfect contestability, such as relaxing the Bertrand-Nash assumption, the Baumol group's deductive analysis "becomes speculative". BPW's emphasis is on the incumbent's *prevention* of entry through anticipatory price restraint, so that actual entry need never occur. Shepherd suggests that contestability theory is therefore not relevant to a "post-entry struggle"; rather one should revert to the extensive literature on entry barriers for guidance in estimating the market outcomes. As the quote opening this Chapter indicates, Baumol and his colleagues themselves concede that perfect contestability is a theoretical benchmark that is immune from considerations of strategic behaviour (i.e., BPW, p. 490).

intratemporal Bertrand-Nash assumption holds, the monopolist would only be able to make a normal profit. However, if Bertrand-Nash were completely relaxed, then the successful monopolist could raise the price of the period 1 product to NSAC immediately after entry, (but no higher, as this would trigger self-production by consumers).

¹⁵ For example, under capacity expansion, if the incumbent firm continues in period 2 to supply the original capacity alone, then it is a Type 1 entrant. If it supplies the new capacity alone, then it is a Type 3 entrant. But if it supplies entire market demand, then it is a monopolist.

¹⁶ It might be argued that, in period 2, there is little to distinguish an incumbent monopolist's strategy from that of a Type 2 entrant under either capacity expansion or capacity replacement. This symmetry is removed if the Bertrand-Nash assumption is fully relaxed, since the monopolist can earn up to the NSAC of the period 1 product, leaving it free to forestall Type 2 entry and still make a supranormal profit. The asymmetry partially arises because the potential Type 2 entrant has to contend with the fact that the incumbent's plant has not yet reached the end of its useful life (e.g., Hay and Morris, p. 95).

As such, inferences regarding the sustainability of real-world firms or markets drawn from a model developed within the narrow assumptions of perfect contestability should be made with caution. Although externally-imposed conditions may restrict a real-world incumbent's pricing decisions, potentially resulting in a unsustainable outcome, BPW themselves point out that "casual empiricism" suggests the opposite. Typically there are many factors which may contribute to making real-world markets more sustainable than would be implied simply from an analysis of the single good model under conditions of perfect contestability (BPW, p. 429).

8.5 Toward Intertemporal Subsidy-Free Prices in the TGTP Model

8.5.1 Game Theoretic Determination of Intertemporal Sustainable and Subsidy-Free Prices

Even if the sustainability of a real-world firm cannot be determined with confidence from a particular model, perhaps it is still possible for a set of subsidy-free prices to be derived that are applicable to that firm. By contrast to sustainable prices, subsidy-free prices are a *prescriptive* rather than *predictive* concept, and as such derive from the benchmark of a perfectly contestable market, as has been discussed earlier (§8.2.4). As a guide for prices which recover costs, encourage efficient consumption choices, and are free from the exercise market power, subsidy-free prices derived under assumptions of perfect contestability replace the prices that would be derived from using a perfectly competitive model as a benchmark. Consequently, as a concept, an analysis of a perfectly contestable market has less to say about the potential sustainability of real-world firms and markets than it does about constrained market pricing. The question of what prices *ought to be*, as opposed to what prices *will be* (e.g., sustainable, subsidy-free and/or efficient), is very different. Determining whether the market will constrain prices to be subsidy-free rests on a different set of assumptions from determining the bounds on subsidy-free prices themselves.

However, relaxing the Bertrand-Nash assumption is clearly not the appropriate way of determining subsidy-free prices within a perfectly contestable context. Firstly, although it (at least theoretically) allows prices to be sustainable, full relaxation of Bertrand-Nash may allow the incumbent monopolist make supranormal profits (§8.4.5), and this violates a key requirement for the presence of subsidy-free prices (§4.3.2). Secondly, Bertrand-Nash is a fundamental basis for perfect contestability, and relaxing it steps away from the benchmark.

Yet, as identified above (§8.4.3), a problem arises in the TGTP model because, in some circumstances, namely when resale between periods is infeasible (and fixed costs are not sufficiently high), no core solution can be found. Under such circumstances, the potential monopolist becomes unsustainable if all prices must be published, and transactions must be cleared, for both periods in advance. It cannot *simultaneously* better the offers from Type 1, Type 2 and Type 3 entrants in both periods. Consequently, the potential monopolist has no incentive to enter the market in the first place for fear it will not be able to recover its costs.

But if this problem is placed in a game theoretic context, it becomes clear that although a static game might be indeterminate, an intertemporal one is not. This is the case even if all prices are published and transactions are cleared in advance, as long as the inherent asymmetry of intertemporality is acknowledged (§8.4.4). A net standalone or net incremental cost associated with a particular coalition of market participants is not realistic if it could never exist in an equilibrium context. Moreover, if an equilibrium cannot be achieved, then the market may not be served at all, and this is the worst outcome for all parties. The previous analysis has not accounted for the foregone opportunity cost of not achieving a sustainable price solution at all. Consequently, there is an incentive for all current and future market participants to allow an equilibrium to be achieved. The outcome would still be indeterminate if all parties were truly symmetric. But fortunately they are not.

Firstly, by definition, entities in period 1, whether self-producing consumers or entrants, all have a “first mover advantage” over entities in period 2. Secondly, some participants have less to lose than others, and so may be willing to forgo their bargaining position if the only other alternative is for the market not to be served at all, or if their resultant coalition of self-supply is substantially more costly than a second-best coalition. The strength of potential consumer coalitions is not symmetric. Even if the Bertrand-Nash assumption is maintained in its strictest intertemporal sense, in the payoff matrix for the intertemporal game, all market participants will see that, should their first-best coalitions never result in an equilibrium (i.e., there are no sustainable prices), there is always a second-best coalition that can result in prices that are better for all parties than if no equilibrium were attained.

When the non-fungible fixed costs are zero, the analysis is straightforward. Under circumstances where no sustainable prices appear to exist because resale is infeasible, a potential monopolist can always offer a price to consumers that is just as good (i.e., the same) as would have been the case had resale been feasible, and still make a normal profit. This is because, as noted above (§8.4.4), resale is not inherently infeasible in these cases, it is simply *dominated* due to potential entry involving Type 1 and Type 3 entrants. And this is the crux of determining subsidy-free prices. As the quote at the beginning of this Chapter highlights, “subsidy-free prices do no more than ensure that the production and sale of each commodity makes all consumers at least as well off as they would otherwise be” (Faulhaber, 1975).

Under any optimal construction configuration, the weakest entry (or consumer coalition self-production) strategy is Type 3 entry. It is clearly the most inefficient, since it results in substantial (suboptimal) spare capacity during period 2, which does not occur under the other entry strategies. Viewing Type 3 entry from a self-production (i.e., consumer) perspective rather than an entry (i.e., firm) perspective, provides a similar conclusion. Such a coalition of self-producing consumers would involve consumers in period 1 forming a coalition with consumers of solely the additional demand (q') in period 2. However, the period 1 consumers will recognise that, if they form a coalition with the consumers in period 2 demanding the original level of capacity (q), they can be just as well off as in the

coalition with consumers of the additional demand. The difference is that the Type 3-like coalition can result in a game with no core. Further, consumers in period 1 have the strongest first mover advantage. They can always build capacity and self-produce if successful coalitions cannot be formed. Hence they have sufficient market power to dictate which coalition they would prefer to associate with. On the other hand, the consumers of the additional demand in period 2 will be no worse off than had the game had a core solution from the outset.

Consequently, when fixed costs are zero, the subsidy-free revenues are simply those which are found when resale is feasible with F set to zero. This result holds under any construction configuration. No bounds exist on the subsidy-free revenues, because NIC equals NSAC for each product in the absence of fixed costs (§8.1.6). These revenues are also “sustainable” in a perfectly contestable sense, although what meaning that would have in the real-world version of the market model depends on the real entry barriers, price constraints and many other factors (§8.4.5).

8.5.2 Allocation of Non-Fungible Fixed Costs

When fixed costs are non-zero, the situation is a little more complex, but a core to the game can still be found. Firstly, where the resale feasibility condition holds—from the relevant expression above (§8.3.3), depending on whether anticipatory construction, capacity expansion or capacity replacement is optimal—then the subsidy-free bounds, derived assuming that resale is feasible, constrain revenues. For example, under anticipatory construction optimality, the relevant subsidy-free revenue bounds are shown in (8.63)-(8.65), the same as those derived earlier in (8.54)-(8.56), subject to (8.62) holding.

$$\text{Where: } \rho_T \leq \frac{(1 + 2\rho_N)K(q) + K(q') - (1 + \rho_N)K(q, q')}{2(1 + \rho_N)K(q, q') - (1 + 2\rho_N)K(q) - K(q')} \quad (8.62)$$

$$\left(\frac{\rho_T}{1 + \rho_T} \right) F + \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \leq P_1 q < F + \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \quad (8.63)$$

$$\frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \leq P_2 q < F + \frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \quad (8.64)$$

$$\frac{1}{dB_N} (1 + \rho_T)[K(q, q') - K(q)] \leq P_2' q' < F + \frac{1}{dB_N} (1 + \rho_T)[K(q, q') - K(q)] \quad (8.65)$$

Secondly, where the original condition for sustainability holds (8.60b), but resale feasibility does not, then the subsidy-free bounds, derived assuming that resale is *infeasible*, constrain revenues. (Again, under anticipatory construction, the relevant subsidy-free revenue bounds are shown in (8.67)-(8.69), the same as those derived earlier in (8.57)-(8.59), subject to (8.66) holding).

Where:

$$\frac{(1+2\rho_N)K(q)+K(q')-(1+\rho_N)K(q,q')}{2(1+\rho_N)K(q,q')-(1+2\rho_N)K(q)-K(q')} < \rho_T \leq \frac{(1+2\rho_N)K(q)+K(q')-(1+\rho_N)K(q,q')+\rho_N F}{2(1+\rho_N)K(q,q')-(1+2\rho_N)K(q)-K(q')} \quad (8.66)$$

$$\left(\frac{\rho_T}{1+\rho_T}\right)F + \frac{1}{dB_N} \left(\frac{\rho_T}{1+\rho_T}\right)K(q,q') \leq P_1 q < F + K(q) - \frac{1}{dB_N} \left(K(q,q') - K(q) - \frac{K(q')}{1+\rho_N}\right) \quad (8.67)$$

$$(1+\rho_T) \left(\frac{1}{dB_N} \left[K(q,q') - \frac{K(q')}{1+\rho_N}\right] - K(q)\right) \leq P_2 q < F + \frac{1}{dB_N} [(1+\rho_T)K(q) - \rho_T K(q,q')] \quad (8.68)$$

$$\frac{1}{dB_N} (1+\rho_T) [K(q,q') - K(q)] \leq P'_2 q' < F + (1+\rho_T)K(q) - \frac{1}{dB_N} \left[\rho_T K(q,q') - \left(\frac{1+\rho_T}{1+\rho_N}\right)K(q')\right] \quad (8.69)$$

Finally, where the original sustainability condition does not hold, NIC and NSAC become equal, and are similar to the subsidy-free revenues that would exist for each product assuming that fixed costs were zero, except that the fixed costs are now positive and are recovered entirely from consumers in both periods that demand the original capacity (q). Consumers associated with the growth in demand (q') make no contribution to the fixed costs. This is because, when forced toward a feasible core solution in the face of potential unsustainability, these incremental consumers can stick to a bargaining position of refusing to pay fixed costs, while allowing an equilibrium to be attained. (Under anticipatory construction, the relevant subsidy-free revenue values are shown in (8.71)-(8.73), subject to (8.70) holding. Note that (8.73) does not include any fixed costs, F).

$$\text{Where: } \rho_T > \frac{(1+2\rho_N)K(q)+K(q')-(1+\rho_N)K(q,q')+\rho_N F}{2(1+\rho_N)K(q,q')-(1+2\rho_N)K(q)-K(q')} \quad (8.70)$$

$$P_1 q = \left(\frac{\rho_T}{1+\rho_T}\right)F + \frac{1}{dB_N} \left(\frac{\rho_T}{1+\rho_T}\right)K(q,q') \quad (8.71)$$

$$P_2 q = F + \frac{1}{dB_N} [(1+\rho_T)K(q) - \rho_T K(q,q')] \quad (8.72)$$

$$P'_2 q' = \frac{1}{dB_N} (1+\rho_T) [K(q,q') - K(q)] \quad (8.73)$$

Although it may appear from (8.71) and (8.72) that period 1 and period 2 consumers contribute to an unequal share of fixed costs, this is not the case. However, simply converting the revenue associated with fixed costs into an annual revenue, by amortising the costs (§9.1.1), shows that consumers of original capacity in both periods contribute dF to fixed costs every year, and subsidy-free prices, rather than revenues, are considered in the following Chapter. This is simply the rental value of

non-fungible and non-deteriorating capital, and the present value of this annual revenue stream in perpetuity is equivalent to the initial outlay for non-fungible fixed costs (F).¹⁷

¹⁷ Like the case with no fixed costs, the revenue bounds defined by (8.62)-(8.73), and the similar expressions that can be derived under capacity expansion and capacity replacement, are also sustainable in a perfectly contestable sense, although this result is of only theoretical value.

CHAPTER IX

INTERTEMPORAL SUBSIDY-FREE PRICES AND ECONOMIC DEPRECIATION IN THE TWO-GOOD/TWO-PERIOD MODEL

The concept of 'cost' has no meaning in either economics or logic except in terms of causation:

US regulatory economist, Alfred Kahn (2001)

A shortcoming of the two-good/two-period (TGTP) model presented in Chapter VIII, as well as of BPW's model of intertemporal unsustainability, is that both effectively only consider subsidy-free and sustainable *revenues* relating to each *product*, rather than subsidy-free and sustainable *prices* relating to each possible *consumer* or *group of consumers*. In this Chapter, the TGTP model is extended to relate to intertemporal subsidy-free prices. From this model, economic asset valuations and economic depreciation schedules can be derived. This allows some initial conclusions to be drawn regarding the efficiency and "fairness" characteristics of electricity line business (ELB) prices in New Zealand, as well as of the optimised deprival valuation (ODV) methodology basis for benchmarking total ELB revenue.

9.1 Intertemporal Subsidy-Free Prices in the TGTP Model

9.1.1 Intertemporal Subsidy-Free Prices and Anonymous Equity

Although BPW talk about sustainable or unsustainable "market prices" arising from their model, effectively their prices are "revenues" relating to an entire period comprising many years, and collected solely at the beginning of that period (e.g., BPW, p. 409). And within each period, consumers are considered as a single block distinguished only by their demand for a particular product. As Hay and Morris (1993, p. 94) point out, one of the usual shortcomings of two-period market models in general is that, although allowing for the passage of time, typically they are still essentially static; first-period capacity decisions and second-period entry and output decisions are "once-and-for-all".¹

Since the model presented in the previous Chapter uses the same convention as Baumol and his colleagues, it provides no insight into what a subsidy-free price might be *within* each period. In general, converting revenues to prices or to a set of regular *payments to capital* simply requires *amortisation* of the revenues, the sole restriction being that the present value of the time stream of amortised revenues (i.e., prices or payments to capital) equals the present value cost of capital outlay (§7.1.3). When fixed costs are zero, the present value cost of capital outlay for each product is its net intertemporal stand alone cost (NSAC), and this is equivalent to the subsidy-free revenue required to be obtained for each product at the beginning of the relevant period. However, there are an infinite set of price paths which can satisfy

¹ This means that, in the TGTP model, even if the Bertrand-Nash assumption is relaxed, the incumbent's post-entry price still relates to the entire period. No possible entry is considered again until the beginning of period 2.

the derived subsidy-free revenue conditions. The question is which of these price paths can be considered subsidy-free. Moreover, perhaps it is possible that many price paths can achieve this goal.

One typical approach to amortisation is to simply transform the revenues into *uniform* prices; namely, prices which are the same in any time interval of the same length. If the specified time interval is one year, then the resultant uniform prices are “annualised” prices.² For instance, Perry (1984), as well as Miyazaki (1990), have both extended BPW’s model and consider that BPW’s implicit assumption that revenues are amortised into uniform prices is actually responsible for BPW’s intertemporal unsustainability result. By relaxing this assumption, and by allowing *intrap*eriod entry, Perry demonstrates that any natural monopolist in a contestable market can always find an intraperiod multiple price strategy that not only prevents entry, but also yields a positive profit.³

However, this result is of less interest when it comes ascertaining subsidy-free prices rather than sustainable prices, since if the model allows the monopolist to make a positive profit, then the perfectly contestable market benchmark is violated. Because the total cost/revenue equality condition is not met, subsidy-free bounds on prices cannot be derived from a model such as Perry’s. Nevertheless, Perry does raise an issue that is highly relevant to the determination of subsidy-free prices; namely, that competitor entry should be considered to be possible at *any* time, and not just at the beginning of some pre-defined “period”. This possibility is ignored in BPW’s model and in the TGTP model presented in Chapter VIII, since consumers are implicitly associated with products on a one-to-one basis.

Such a one-to-one treatment is consistent with Faulhaber’s (1975) early work on cross subsidies (§4.3.2). Yet markets having three products do not necessarily have three consumers, or even three consumer groups (§3.1.2). Consequently, to derive intertemporally anonymously equitable prices in the TGTP model requires that entry strategies and self-production possibilities be considered at any instant of time during any period, otherwise a notional barrier to entry has been raised with respect to intraperiod entry and/or self-production. Since perfect contestability requires no constraints on entry (§2.1.7),

² As shown earlier (§7.3.1), it is fairly straightforward to demonstrate that, if there is no demand growth, the only subsidy-free price path is a uniform one. This is somewhat analogous to Stokey’s (1979) result that a monopolist selling a durable good will not seek to engage in intertemporal price discrimination if there is a perfect second-hand market. Consumers with the highest valuation of the good always obtain it.

³ An entirely different way of addressing the unsustainability problem arises from the work of Tirole (1988). In his discussion of contestable markets, Tirole argues that it is better to study the equilibria of an appropriate *game* than the contestable outcomes. Tirole shows that even under uniform pricing a zero profit equilibrium can exist in a two-stage game of natural monopoly where price-setting comes before quantity-setting. Canoy (1994), for one, has extended Tirole’s model to show that differential pricing can be exploited by a natural monopolist to deter entry and make a positive profit in equilibrium even when entry is costless. As explained earlier (§2.1.7, fn. 16), analysis based on Tirole’s work is not explicitly discussed further in this thesis, since New Zealand’s policy reforms in the electricity sector have not been developed within such methodological framework. Rather, they have been derived from concepts inherent in contestability theory.

allowing intraperiod entry is necessary for the subsidy-free prices to be consistent with a benchmark of perfect contestability. However, this approach will not invalidate the existing results, since the already-identified subsidy-free bounds on revenues will still remain necessary, although not sufficient, conditions for the existence of subsidy-free prices.

9.1.2 Assumptions for Deriving Subsidy-Free Prices rather than Revenues

In the subsequent analyses, it is assumed that equilibrium—or more correctly, a core solution—is achieved under all circumstances. Therefore as has been discussed in the previous Chapter (§8.5), sustainable subsidy-free revenue bounds can always exist under conditions of perfect contestability. Consequently, non-fungible fixed costs (F) can be ignored, since the main significance of these costs in BPW’s original model was to demonstrate how fixed costs can transform intertemporal unsustainability into sustainability. The unsustainability problem is now resolved by assuming that market participants will recognise that a “second-best” equilibrium is better than no equilibrium at all, and the attainment of equilibrium will itself result in sustainability. Moreover, non-fungible fixed costs can be considered separable and thus, where present, could be allocated separately from the costs of capacity.

In addition, only cases of perpetual demand where assets are fungible are going to be examined. The assumption of perpetual demand is not as unrealistic as it might first appear. Demand does not really have to be perpetual, all that is required is an expectation from the current industry supplier (or self-producers) that it (or they) can sell their asset(s) to another supplier/self-producer holding similar expectations. Assets are assumed to be fungible, since, as discussed earlier (§8.1.6), artificially restricting intertemporal asset resale between suppliers and/or self-producers raises a notional entry barrier inconsistent with the derivation of subsidy-free prices as a perfectly contestable benchmark. Consequently, the starting point for the derivation of subsidy-free prices is the already-derived subsidy-free revenue bounds, with F set to zero and assets not restricted from being resold. Because the equilibrium assumption ensures sustainability, resale is always feasible (§8.2.2).

9.1.3 Subsidy-Free Prices under Anticipatory Construction

Therefore, under anticipatory construction optimality, the relevant subsidy-free revenue equations can be derived from expressions (8.53)-(8.55) by setting F equal to zero. To derive the subsidy-free prices, a similar approach is taken as in the constant demand example presented earlier (§7.3). Period 1 is entirely divided into two sub-periods termed 1α and 1β , of variable period length α and β years respectively (where $\alpha+\beta\equiv T$, $\alpha > 0$, and $\beta \geq 0$), for every possible combination of entrants or self-producers during the entirety of period 1. Similarly, period 2 is divided into two periods 2δ and 2ϵ : the first is of variable length $\delta > 0$ and covers the years $T \leq t < T + \delta$; and the second is of variable length $\epsilon > 0$ and covers the remaining years $T + \delta \leq t < \text{infinity}$. Since the subsidy-free revenue constraints on subsidy-free prices are now equalities, no total revenue/cost equation is required, as equation (8.41) is automatically satisfied. The resultant subsidy-free revenue constraints are shown in equations (9.1)-(9.3).

$$P_1 q \equiv P_{1\alpha} q + \frac{P_{1\beta} q}{1 + \rho_\alpha} = \frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') \quad (9.1)$$

$$P_2 q \equiv P_{2\delta} q + \frac{P_{2\varepsilon} q}{1 + \rho_\delta} = \frac{1}{dB_N} [(1 + \rho_T)K(q) - \rho_T K(q, q')] \quad (9.2)$$

$$P'_2 q' \equiv P'_{2\delta} q' + \frac{P'_{2\varepsilon} q'}{1 + \rho_\delta} = \frac{1}{dB_N} (1 + \rho_T) [K(q, q') - K(q)] \quad (9.3)$$

Additional constraints are now introduced by allowing entry or self production at *any* time within the two periods. During period 1, constraints caused by entrants or self producers that plan to supply the entire current and future market demand are described by expression (9.4a). The resale value, at the end of period 1, of the asset which they plan to construct at some point during period 1, is given in (9.4b). Although these expressions appear similar to the Type 1 entry with resale constraint used to derive subsidy-free revenue bounds, given in (8.50a) and (8.50b), these differ substantially, since (9.4a) and (9.4b) describe an *infinite set* of constraints, rather than just a single constraint.

$$P_{1\beta} q \leq K(q, q') - \frac{S_1(q, q')}{1 + \rho_\beta}; \quad \forall \beta : 0 \leq \beta \leq T \quad (9.4a)$$

$$\text{where: } S_1(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\beta}{1 + \rho_N} \right) K(q, q') \quad (9.4b)$$

However, this infinite set of constraints is not the only possible constraining influence on prices in period 1. In parallel with these constraints, which model entry during period 1 with asset resale occurring at the end of period 1, is a set of constraints representing an entrant who serves consumers from the beginning of period 1, but sells up these assets before the end of period 1. This entrant constructs assets with a capacity sufficient to supply *entire* market demand for period 2. The resultant set of constraints is shown in (9.5a) with the corresponding resale equation in (9.5b).

$$P_{1\alpha} q \leq K(q, q') - \frac{S_{1\alpha}(q, q')}{1 + \rho_\alpha}; \quad \forall \alpha : 0 \leq \alpha \leq (T - \beta) \quad (9.5a)$$

$$\text{where: } S_{1\alpha}(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\alpha}{1 + \rho_N} \right) K(q, q') = \left(\frac{\rho_N - \rho_\alpha}{\rho_N} \right) K(q, q') \quad (9.5b)$$

This set of constraints can also be considered to represent the net intertemporal standalone cost (NSAC) of self producing consumers whose demand does not last for all of period 1. Although one might expect that such consumers might minimise their SAC by constructing capacity at the beginning of period 1 sufficient to supply only their own demand (i.e., q), under anticipatory construction, there would be no willing purchasers of these assets. Minimising costs requires that the self producers construct capacity sufficient to serve not only themselves, but the entire market demand of period 2 (i.e., q and q').

However, since a feasible resale price exists for these assets at any time during period 1, as shown in (9.5b), building spare capacity at the beginning of period 1 dominates other self production possibilities. Similarly, the entry strategy described by (9.4a), which also requires building spare capacity (although now at some point during period 1) dominates other entry possibilities.

Under anticipatory construction optimality, expressions (9.4a) through (9.5b), when linked by (9.1), successfully protect the interests of all possible configurations of consumers coming and going during period 1, and any possible coalitions of those consumers (assuming that the total market demand during period 1 remains at q). Consequently, these constraints ensure anonymous equity. Combining all these constraints results in the pair of sets of equalities in (9.6a) and (9.6b) which fully describe the conditions constraining period 1 prices (where $\alpha + \beta \equiv T$, $\alpha > 0$, and $\beta \geq 0$).

$$P_{1\beta}q = \frac{1}{dB_N} \left(\frac{\rho_\beta}{1 + \rho_\beta} \right) K(q, q'); \quad \forall \beta: 0 \leq \beta \leq T \quad (9.6a)$$

$$P_{1\alpha}q = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right) K(q, q'); \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (9.6b)$$

These expressions still define subsidy-free revenues, although now the revenues recover costs for a period ranging anywhere from 0 to T years in length. Transforming these equations into a price p_t —in a similar manner to (7.23)—provides (9.7), since a price can be considered to be a revenue for a sub-period of arbitrary length t .

$$P_{1\alpha}q = \int_{t=0}^{\alpha} p_t q \left(\frac{1}{1+d} \right)^t = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right) K(q, q'); \quad \forall \alpha: 0 \leq \alpha \leq T \quad (9.7)$$

Because the equality in (9.7) holds for any sub-period length between 0 and T years, it is clear that the price for any sub-period of the *same* length must be the same. A more familiar way of looking at this subsidy-free price is to transform it into annual payments to capital (i.e., annual prices). Over the course of period 1, annual payments to capital p_j , made at the end of the j th year, relate to the total revenue collected at the beginning of period 1, as shown in (9.8a). Rearranging—in a similar manner to (7.24a) and (7.24b)—provides the annual payment to capital defined below in (9.8a) and (9.8b) as \bar{p}_1 , which from (9.7) must clearly remain constant throughout period 1.

$$\frac{1}{dB_N} \left(\frac{\rho_T}{1 + \rho_T} \right) K(q, q') = P_1q = \sum_{j=1}^T p_j q \left(\frac{1}{1+d} \right)^j = \sum_{j=1}^T \bar{p}_1 q \left(\frac{1}{1+d} \right)^j = B_T \bar{p}_1 q = \frac{1}{d} \left(\frac{\rho_T}{1 + \rho_T} \right) \bar{p}_1 q \quad (9.8a)$$

$$\bar{p}_1 = p_t(q) = \frac{K(q, q')}{qB_N}; \quad 0 \leq t \leq T \quad (9.8b)$$

This outcome shows that, of all the infinite sets of price paths which can satisfy the subsidy-free revenue constraint for period 1 given in (9.1), only a single price path provides a set of subsidy-free and anonymously equitable prices. Interestingly, the subsidy-free annual payments to capital throughout period 1 should be *uniform*, and simply annualising the subsidy-free revenues would have provided this result from the outset. But, as will be seen in the next sections (§9.1.4 and §9.1.5), this uniform price result—throughout the *entire* period—does not arise where either capacity expansion or capacity replacement are the optimal construction configurations. Nevertheless, the subsidy-free price paths do always involve *amortisation*, since the subsidy-free price is always the amortised opportunity cost of supply. It just so happens that for anticipatory construction, annualised revenue equals the amortised opportunity cost of supply throughout all of period 1.

Deriving the subsidy-free price path for period 2 under anticipatory construction follows a similar approach, and entry and self production possibilities are given in (9.9), (9.10a) and (9.10b) below. These intraperiod constraints are the same irrespective of which construction configuration is optimal, because the stand alone costs of incumbent firms or self producers during period 2 are always constrained by entrants that take a Type 2 entrant-like entry strategy (§8.1.2); one that serves entire market demand, q and q' . The difference is that a Type 2 entrant attempts to serve entire market demand from the *beginning* of period 2, whereas the potential intraperiod entrants (or self-producers), described by expressions (9.9)-(9.10b), attempt to serve entire market demand from any point onward *during* period 2.

$$P_{2\varepsilon}q + P'_{2\varepsilon}q' \leq \frac{1}{dB_N} K(q, q'); \quad \forall \varepsilon: \varepsilon \geq 0 \quad (9.9)$$

$$P_{2\delta}q + P'_{2\delta}q' \leq \frac{1}{dB_N} K(q, q') - \frac{S_{2\delta}(q, q')}{1 + \rho_\delta}; \quad \forall \delta: \delta > 0 \quad (9.10a)$$

$$\text{where: } S_{2\delta}(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\delta}{1 + \rho_N} \right) K(q, q') \quad (9.10b)$$

Combining these two sets of constraints with the subsidy-free revenue constraints, given in (9.2) and (9.3) above, results in subsidy-free prices for period 2. The annual payments to capital are once again constant (or uniform) and are shown in equations (9.11) and (9.12).

$$\bar{p}_2 \equiv p_t(q) = \frac{(1 + \rho_T)K(q) - \rho_T K(q, q')}{qB_N}; \quad T < t \leq N \quad (9.11)$$

$$\bar{p}'_2 \equiv p'_t(q') = \frac{(1 + \rho_T)[K(q, q') - K(q)]}{q'B_N}; \quad T < t \leq N \quad (9.12)$$

9.1.4 Subsidy-Free Prices under Capacity Expansion: the Influence of Optimal Investment

Like anticipatory construction, setting F equal to zero in the single product cross subsidy exclusion conditions relating to capacity expansion optimality provides the relevant subsidy-free revenue equations, and these are shown in (9.13)-(9.15).

$$P_1q \equiv P_\alpha q + \frac{P_{1\beta}q}{1 + \rho_\alpha} = \frac{1}{dB_N} \left(K(q) - \frac{K(q, q') - K(q')}{1 + \rho_T} \right) \quad (9.13)$$

$$P_2q \equiv P_{2\delta}q + \frac{P_{2\varepsilon}q}{1 + \rho_\delta} = \frac{1}{dB_N} [K(q, q') - K(q')] \quad (9.14)$$

$$P'_2q' \equiv P'_{2\delta}q' + \frac{P'_{2\varepsilon}q'}{1 + \rho_\delta} = \frac{1}{dB_N} K(q') \quad (9.15)$$

The other constraints, however, are somewhat more complex than under anticipatory construction. Under capacity expansion, competitors attempting to enter during period 1 now have two possibly dominant modes of entry. One is to construct capacity only sufficient for peak demand in period 1 (i.e., q), as does the incumbent optimally supplying demand from the beginning of period 1. However, an entrant that constructs sufficient capacity during period 1 to serve the entire market demand of period 2 (i.e., q and q'), can dominate other possible entry strategies, as long as the entrant completes construction less than $T_{AC \rightarrow CE}$ years before the *end* of period 1. $T_{AC \rightarrow CE}$, defined earlier in (8.3b), is the minimum first period length for capacity expansion to be the optimal construction configuration, assuming that capacity expansion is feasible at all (i.e., expression (8.49) holds).

What this means is that, even where capacity expansion is the optimal construction configuration for a monopolist supplying all market demand in both periods 1 and 2, anticipatory construction may be the optimal entry strategy for a later potential entrant. This demonstrates that, as discussed earlier (§6.1.4), what may be the optimal configuration prior to an investment, may not be so at a later point in time should the investment be replaced on a greenfields basis. Anticipatory construction will become the optimal entry strategy at such a point in time as it would have been the optimal construction configuration had there been no demand prior to that time. The related set of constraints is shown in (9.16a), subject to the sub-period length restriction in (9.16b), and the resale equation in (9.16c).

$$P_{1\beta}q \leq K(q, q') - \frac{S_1(q, q')}{1 + \rho_\beta}; \quad \forall \beta: 0 \leq \beta < T - \tau_E \quad (9.16a)$$

$$\text{for } \tau_E \equiv T - T_{AC \rightarrow CE} \quad (9.16b)$$

$$\text{where: } S_1(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\beta}{1 + \rho_N} \right) K(q, q') \quad (9.16c)$$

For potential entrants entering *within* τ_E (i.e., $T - T_{AC \rightarrow CE}$) years of the beginning of period 1, the dominating entry strategy (and thus the optimal investment decision on a greenfields basis) is to construct capacity only sufficient for the peak demand in period 1. This set of constraints, and the associated resale equation, are shown in expressions (9.17a) and (9.17b) respectively.

$$P_{1\beta}q \leq K(q) - \frac{S_1(q)}{1 + \rho_\beta} \quad \forall \beta: T - \tau_E \leq \beta \leq T \quad (9.17a)$$

$$\text{where: } S_1(q) \equiv \frac{1}{dB_N} \left[K(q, q') - K(q') - \left(\frac{1 + \rho_\beta}{1 + \rho_N} \right) K(q) \right] \quad (9.17b)$$

In parallel with both sets of constraints in (9.16) and (9.17), which model entry (or self-production) during period 1 with asset resale occurring at the end of period 1, is a set of constraints representing an entrant (or group of self producers) that serves consumers from the beginning of period 1, but sells up these assets before the end of period 1. Unlike the entry strategy associated with the set of constraints relating to entry *during* period 1, the dominant strategy for entrants (or self producers) serving demand from the *beginning* of period 1 is to construct assets with a capacity sufficient to supply market demand in period 1 only (i.e., q). However, the resale value depends on whether resale occurs before or after τ_E years. Consequently, the set of constraints is given in (9.18a), where the resale equation is (9.18c) if resale occurs after τ_E years—the condition for which is given in (9.18b)—or (9.18d) otherwise.

$$P_{1\alpha}q \leq K(q) - \frac{S_{1\alpha}(q)}{1 + \rho_\alpha} \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (9.18a)$$

$$\text{where for: } \rho_\alpha \geq \frac{(1 + \rho_T)[K(q, q') - K(q)] - K(q')}{K(q')} = \rho_{\tau_E} \equiv (1 + d)^{\tau_E} - 1 \quad (9.18b)$$

$$S_{1\alpha}(q, q') \equiv \frac{1}{dB_N} \left[K(q, q') - \left(\frac{1 + \rho_\alpha}{1 + \rho_T} \right) K(q') - \left(\frac{1 + \rho_\alpha}{1 + \rho_N} \right) K(q) \right] \quad (9.18c)$$

$$\text{otherwise: } S_{1\alpha}(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\alpha}{1 + \rho_N} \right) K(q, q') = \left(\frac{\rho_N - \rho_\alpha}{\rho_N} \right) K(q, q') \quad (9.18d)$$

Under capacity expansion optimality, expressions (9.16a) through (9.18d), when linked by (9.13), successfully protect the interests of all possible configurations of consumers coming and going during period 1, as well as any possible coalitions of those consumers. Combining all these constraints results in two pairs of sets of equalities, the first pair, (9.19a) and (9.19b), relating to subsidy-free prices after τ_E years, and the second pair, (9.20a) and (9.20b), to prices before τ_E years.

$$P_{1\beta}q = \frac{1}{dB_N} \left(\frac{\rho_\beta}{1 + \rho_\beta} \right) K(q, q'); \quad \forall \beta: 0 \leq \beta < T - \tau_E \quad (9.19a)$$

$$P_{1\alpha}q = \frac{1}{dB_N} \left(K(q) + \frac{K(q')}{1 + \rho_T} - \frac{K(q, q')}{1 + \rho_\alpha} \right); \quad \forall \alpha: \tau_E < \alpha \leq T \quad (9.19b)$$

$$P_{1\beta}q = \frac{1}{dB_N} \left(K(q) - \frac{K(q, q') - K(q')}{1 + \rho_\beta} \right); \quad \forall \beta: T - \tau_E \leq \beta \leq T \quad (9.20a)$$

$$P_{1\alpha}q = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right) K(q); \quad \forall \alpha: 0 < \alpha \leq \tau_E \quad (9.20b)$$

Transforming these constraints into prices results in (9.21), which indicates that, unlike anticipatory construction optimality, the subsidy-free price path under capacity expansion does *not* remain constant for all of period 1. The subsidy-free price path consists of *two* sub-periods of constant (or uniform) prices: the first relating to the amortised cost of the capacity $K(q)$; and the second to the amortised cost of the capacity $K(q, q')$. This second sub-period price is the amortised opportunity cost of supply after τ_E years, since it relates to the net intertemporal stand alone cost of self-production (based on greenfields optimality).

$$p_t(q) = \begin{cases} \frac{K(q)}{qB_N}; & 0 < t < \tau_E \\ \frac{K(q, q')}{qB_N}; & \tau_E \leq t \leq T \end{cases} \quad (9.21)$$

As explained above (§9.1.3), the intraperiod entry and self-production possibilities during period 2 are the same under any construction configuration. Combining these generally applicable constraints, from (9.9)-(9.10b), with the subsidy-free revenue equations specific to capacity expansion optimality, from (9.14) and (9.15), provides the subsidy-free prices for period 2, shown in (9.22) and (9.23).

$$\bar{p}_2 \equiv p_t(q) = \frac{K(q, q') - K(q')}{qB_N}; \quad T < t \leq N \quad (9.22)$$

$$\bar{p}'_2 \equiv p'_t(q') = \frac{K(q')}{q'B_N}; \quad T < t \leq N \quad (9.23)$$

9.1.5 Subsidy-Free Prices under Capacity Replacement

Deriving the subsidy-free prices relating to capacity replacement optimality follows a similar approach. The relevant subsidy-free revenue equations are shown in (9.24)-(9.26).

$$P_1 q \equiv P_{1\alpha} q + \frac{P_{1\beta} q}{1 + \rho_\alpha} = K(q) \quad (9.24)$$

$$P_2 q \equiv P_{2\delta} q + \frac{P_{2\varepsilon} q}{1 + \rho_\delta} = \frac{1}{dB_N} \left(\frac{1 + \rho_T}{1 + \rho_N} \right) K(q) \quad (9.25)$$

$$P'_2 q' \equiv P'_{2\delta} q' + \frac{P'_{2\varepsilon} q'}{1 + \rho_\delta} = \frac{1}{dB_N} \left[K(q, q') - \left(\frac{1 + \rho_T}{1 + \rho_N} \right) K(q) \right] \quad (9.26)$$

Like capacity expansion, the other constraints relate to two possibly dominant entry strategies. The first is to construct capacity during period 1 that is only sufficient for the demand of period 1, and the second is to construct capacity during period 1 that is sufficient to serve the entire market demand for period 2. This latter strategy can dominate all other possible entry strategies, as long as the entrant completes construction more than τ_R years after the beginning of period 1, where τ_R is defined below in (9.27b). As such, even where capacity replacement is the optimal construction configuration for a monopolist supplying all market demand in both periods 1 and 2, anticipatory construction may be the optimal entry strategy for a later potential entrant. The set of constraints relating to this entry strategy is shown in (9.27a), subject to the period length restriction in (9.27b), and the resale equation in (9.27c).

$$P_{1\beta} q \leq K(q, q') - \frac{S_1(q, q')}{1 + \rho_\beta}; \quad \forall \beta: 0 \leq \beta < T - \tau_R \quad (9.27a)$$

$$\text{for } \tau_R \equiv T - \log_{(1+d)} \left(1 + \frac{(\rho_N - \rho_T)K(q)}{(1 + \rho_N)[K(q, q') - K(q)]} \right) \quad (9.27b)$$

$$\text{where: } S_1(q, q') \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\beta}{1 + \rho_N} \right) K(q, q') \quad (9.27c)$$

For potential entrants entering within τ_R years of the beginning of period 1, the dominating entry strategy is to construct capacity only sufficient for the peak demand in period 1. This set of constraints, and the associated resale equation, are shown in expressions (9.28a) and (9.28b) respectively.

$$P_{1\beta} q \leq K(q) - \frac{S_1(q)}{1 + \rho_\beta}; \quad \forall \beta: T - \tau_R \leq \beta \leq T \quad (9.28a)$$

$$\text{where: } S_1(q) \equiv \frac{1}{dB_N} \left(\frac{\rho_T - \rho_\beta}{1 + \rho_N} \right) K(q) \quad (9.28b)$$

Like capacity expansion, in parallel with both sets of constraints in (9.27) and (9.28), which model entry during period 1 with asset resale occurring at the end of period 1, is a set of constraints representing an entrant or group of self producers that serves consumers from the beginning of period 1, but sells up those assets before the end of that period. The only dominant entry strategy is to construct

assets with a capacity sufficient to supply market demand in period 1 alone. However, the resale value depends on whether resale occurs before or after τ_R years. Consequently, the set of constraints is given in (9.29a), where the resale equation is (9.29c) if resale occurs after τ_R years—the condition for which is given in (9.29b)—or (9.29d) otherwise.

$$P_{1\alpha}q \leq K(q) - \frac{S_{1\alpha}(q)}{1 + \rho_\alpha} \quad \forall \alpha: 0 \leq \alpha \leq (T - \beta) \quad (9.29a)$$

$$\text{where for: } \rho_\alpha \geq \frac{\rho_T(1 + \rho_N)K(q, q') - \rho_N(1 + \rho_T)K(q)}{(1 + \rho_N)K(q, q') - (1 + \rho_T)K(q)} = \rho_{\tau_R} \equiv (1 + d)^{\tau_R} - 1 \quad (9.29b)$$

$$S_{1\alpha}(q) \equiv \frac{1}{dB_N} \left(\frac{\rho_T - \rho_\alpha}{1 + \rho_T} \right) K(q, q') \quad (9.29c)$$

$$\text{otherwise: } S_{1\alpha}(q) \equiv \frac{1}{dB_N} \left(1 - \frac{1 + \rho_\alpha}{1 + \rho_N} \right) K(q) = \left(\frac{\rho_N - \rho_\alpha}{\rho_N} \right) K(q) \quad (9.29d)$$

Under capacity replacement optimality, expressions (9.27a) through (9.29d), when linked by (9.24), successfully protect the interests of all possible configurations of consumers coming and going during period 1, as well as any possible coalitions of those consumers. Combining all these constraints results in two pairs of sets of equalities, the first pair, (9.30a) and (9.30b), relating to subsidy-free prices after τ_R years, and the second pair, (9.31a) and (9.31b), to prices before τ_R years.

$$P_{1\beta}q = \frac{1}{dB_N} \left(\frac{\rho_\beta}{1 + \rho_\beta} \right) K(q, q'); \quad \forall \beta: 0 \leq \beta < T - \tau_R \quad (9.30a)$$

$$P_{1\alpha}q = K(q) + \frac{1}{dB_N} \left(\frac{1}{1 + \rho_T} - \frac{1}{1 + \rho_\alpha} \right) K(q, q'); \quad \forall \alpha: \tau_R < \alpha \leq T \quad (9.30b)$$

$$P_{1\beta}q = \frac{1}{dB_N} \left(1 - \frac{1 + \rho_T}{(1 + \rho_\beta)(1 + \rho_N)} \right) K(q); \quad \forall \beta: T - \tau_R \leq \beta \leq T \quad (9.31a)$$

$$P_{1\alpha}q = \frac{1}{dB_N} \left(\frac{\rho_\alpha}{1 + \rho_\alpha} \right) K(q); \quad \forall \alpha: 0 < \alpha \leq \tau_R \quad (9.31b)$$

Transforming these constraints into prices results in (9.32), which indicates that—like capacity expansion—the subsidy-free price path consists of *two* sub-periods of uniform or constant prices, the first relating to the cost of capacity $K(q)$, and the second to $K(q, q')$. Again, each price is the amortised opportunity cost of supply relating to the net intertemporal stand alone cost of self-production on an optimal greenfields basis.

$$p_t(q) = \begin{cases} \frac{K(q)}{qB_N}; & 0 < t < \tau_R \\ \frac{K(q, q')}{qB_N}; & \tau_R \leq t \leq T \end{cases} \quad (9.32)$$

For period 2, combining the generally applicable constraints, from (9.9)-(9.10b), with the subsidy-free revenue equations specific to capacity replacement optimality, from (9.25) and (9.26), provides the subsidy-free prices for period 2, shown in (9.33) and (9.34).

$$\bar{p}_2 \equiv p_t(q) = \left(\frac{1 + \rho_T}{1 + \rho_N} \right) \frac{K(q)}{qB_N}; \quad T < t \leq N \quad (9.33)$$

$$\bar{p}'_2 \equiv p'_t(q') = \frac{K(q, q') - \left(\frac{1 + \rho_T}{1 + \rho_N} \right) K(q)}{q'B_N}; \quad T < t \leq N \quad (9.34)$$

9.1.6 The Transition Point for Increases in the Subsidy-Free Price

Under capacity replacement optimality, determining the point at which the subsidy-free price increases and at which the dominant entry strategy changes (i.e., τ_R from 9.27b), is not as straightforward as under capacity expansion optimality (i.e., 9.16b). Where capacity expansion is the optimal *construction* configuration, capacity expansion can also be the dominant *entry* strategy, and it will dominate at any time before that at which anticipatory construction becomes the optimal entry strategy. As explained above, this transition point is the same time at which anticipatory construction would have henceforth become the optimal construction configuration, assuming no prior demand—in other words, the optimal greenfields configuration.

By contrast, where capacity replacement is the optimal construction configuration, capacity replacement cannot itself be a dominant intraperiod entry strategy. This is because, under capacity replacement, an asset constructed to serve the market demand of period 1 alone (i.e., q) has no resale value at the end of period 1 (§8.3.2). Consequently, an incumbent monopolist optimally serving the entire market demand of both periods with a program of capacity replacement can always force a later entrant, using the same construction program, out of the market. This is because the later entrant must recover its costs relating to period 1 over a shorter period, and as such would have to offer a higher price.

Under capacity replacement optimality, potential entrants that propose to enter after τ_R years of the beginning of period 1, would have to offer a program of anticipatory construction. But before τ_R years, potential entrants must construct capacity sufficient to serve only demand q . This will lead to a program of capacity expansion for serving the entire market demand during period 2, either directly, should the entrant itself construct new capacity sufficient to meet the demand growth q' , or indirectly,

should this growth in demand be served by self production, or by another firm. It is interesting to note that a program of capacity expansion would be the outcome even if capacity expansion had not initially been a feasible construction configuration (i.e., should expression (8.49) not have held). In any event, potential intraperiod entrants can take no action that would result in a program of capacity replacement, as was suggested in the discussion of Baumol and Sidak's (1994a) sequencing of demands example earlier (§6.2.1).

The price/dominance transition point under capacity replacement is related to asset valuation, and the need to recover the entire capital outlay through accelerated depreciation (§7.3.3) as will be seen shortly (§9.2). The higher price in the latter part of period 1 is *not* due to the cost of the second period capacity being recovered *in advance*. It is needed to recover the entire cost of the *original* asset, which otherwise would become partially stranded. As noted above, the main characteristic of capacity replacement is that the resale value of the asset constructed to meet initial demand falls to zero at the time of that demand increases.

9.2 Economic Asset Valuation and Depreciation in the TGTP Model

9.2.1 Asset Valuation and Depreciation in the Model under Anticipatory Construction

In the TGTP model, there is no overall objective function; the goal is not to minimise or maximise a particular function, but simply to determine a range of subsidy-free bounds on prices. However, based on the assumptions made above (§9.1), the constraints on prices have resulted in a *single* price path relating to each of the possible optimal construction configurations. Therefore, it seems likely that there will be a single optimal depreciation (and asset valuation) path associated with the subsidy-free price path for a specific construction configuration, just as there was for the constant demand case (§7.3).

The potential resale value of the assets involved in the TGTP model form an intrinsic part of the set of constraints within which subsidy-free prices are determined. The resale values that appear in the constraint expressions are the values that the relevant assets would have to a subsequent firm or self-producer, taking account of the asset's remaining lifetime (i.e., service potential) and the cost of serving any current or future demand that would be unserved after the purchase of that asset. Since a resale value can always be directly derived from the model—because it is assumed that an “alternative use”; namely, resale to “alternative users” (§7.3.4)—is always feasible, the problem of circularity in valuation described earlier (§7.1.2) is not present. Asset valuations can be derived directly from the opportunity cost of supply, without the future stream of payments to capital being known.

Asset value is inextricably linked to the nature of current and future demand. For example, under both the single-good and two-good models it can be seen that when anticipatory construction is optimal, and demand is *finite* (i.e., demand ceases prior to the end of any asset's lifetime), the original asset is in fact perfectly fungible at the end of the first period (8.22b). Thus, the asset value does not depreciate.

This is because the original asset's best alternative use at the end of period 1 is by a firm wanting to supply all consumers in the second period (or all consumers in period 2 intending to self-produce). Since demand does not extend past the lifetime of, or exceed the capacity of, the original asset, the original asset still has just as much value as an entirely new asset of the same capacity.

This not the case, however, where demand is assumed to exist in perpetuity. As noted above (§9.1.2), the perpetual demand assumption simply requires that the current industry supplier (or self-producers) expect to be able to sell their assets to another supplier/self-producer that has similar expectations. This means that supplier/consumer willingness-to-pay and stand-alone cost, net of any resale value, would be based on an expectation of perpetual demand. If perpetual demand does not eventuate, then the entity that incorrectly maintained this expectation, but remains the owner of the assets at the time demand unexpectedly ceases, will of course make a loss, and the assets will be associated with some level of irrecoverable stranded cost.

Under anticipatory construction when it is assumed that there is perpetual demand, a single asset with an initial cost $K(q, q')$ serves entire intertemporal market demand. Hence, with zero fixed costs, the market demand is best served by intertemporal (natural) monopoly over the N year period of the asset's lifetime. The asset's value at any time during its lifetime can be found from the already-derived resale equations applicable to: (i) a firm or self-producer serving entire market demand from the beginning of period 1 and selling that asset prior to the end of period 1, as shown in (9.5b); and (ii) a firm or self-producer serving entire market demand from the beginning of period 2, and selling that asset prior to the end of period 2, as shown in (9.10b). Combining these equations to provide a single equation applicable over the asset's entire lifetime, results in (9.35), the asset value at the beginning of the i th year (V_i).

$$V_i(q, q') = \left(\frac{\rho_N - \rho_{i-1}}{\rho_N} \right) K(q, q') = \frac{B_{N-i+1}}{B_N} K(q, q'); \quad 1 \leq i < N + 1 \quad (9.35)$$

In words, the economic asset value is equal to the initial asset cost, multiplied by the ratio of: the uniform series present worth factor, for a period encompassing the number of years until the end of the asset's lifetime; to, the uniform series present worth factor for the entire period of the asset's lifetime. This result is similar to the resale equation under constant demand (7.16), and to the valuation of an asset where there are constant payments to capital (7.8).

Under anticipatory construction, the subsidy-free prices for each product remain constant within the relevant period. In period 1, the total annual payments to capital are therefore also constant, and are found by multiplying price, from (9.8b), by total demand (q). In period 2, the total annual payments to capital are found by multiplying the price from (9.11) by its associated demand (q), and summing this product with the product of price from (9.12) and its associated demand (q'). Consequently, in both periods, the annual total payments to capital remain constant, being $K(q, q')/B_N$. Since the payments to

capital are constant, the economic depreciation equation where there are constant payments to capital applies—(7.9). However, the same result can also be found from the more general expression of (7.3), given that I_{t+1} is zero over the asset’s lifetime. In either case, the total annual depreciation that needs to be recovered by an incumbent firm under anticipatory construction is as shown in (9.36).

$$D_i = \frac{K(q, q')}{B_N(1+d)^{N-i+1}} \quad (9.36)$$

Using (7.2) to combine the annual depreciation from (9.36), with the return on the asset value found from (9.35), results in a set of payments to capital which are consistent with the subsidy-free prices derived under anticipatory construction (§9.1.3). Consequently, it can be seen that the subsidy-free prices derived earlier are entirely consistent with BPW’s expressions for economic asset value and for optimal depreciation. And although there are an infinite number of depreciation streams that could potentially satisfy BPW’s conditions for optimal depreciation, the set of constraints developed for the purpose of deriving subsidy-free prices results in a single set of asset valuations, payments to capital, and consequently, depreciation payments.

On the other hand, as noted earlier (§7.2.2), BPW (p. 471) stipulate that “intertemporal price patterns must satisfy the rules of economic depreciation, with no contribution toward recoupment of fixed or sunk costs in periods of excess capacity”. Given BPW’s definition of sunk cost, discussed earlier (§3.5.2), this would imply that, under anticipatory construction, *intraproduct* payments to capital during period 1 would contain absolutely no depreciation component. Baumol (1971) compared this result to an intertemporal peak load pricing problem, and justifies his conclusion by stating that charging depreciation in “off peak” periods would contribute nothing toward increasing asset utilisation at times when unused capacity is available. However, as discussed throughout this thesis, spare capacity may be part of the optimal investment program, and the above analysis demonstrates that, if depreciation provision were not allowable under an optimal program of anticipatory construction, prices could not be subsidy-free.

If the depreciation provisions are set aside, then at the end of an original asset’s lifetime the accumulated depreciation will be equal to the initial capital outlay, as is seen from (7.5). Therefore, in the absence of inflation or changes of technology, the investment is self-sustaining, since the accumulated depreciation can be used to purchase a new (replacement) asset with the same service potential (i.e. capacity). And although the asset’s value decreases over its lifetime, the entire value of the investment as a whole, or the value of the monopoly firm’s entire operations, remains constant. This is because the value of the firm as a whole comprises not only the (decreasing) value of the physical asset, but also the (increasing) value of accumulated depreciation, as shown earlier (§7.3.2).

9.2.2 Asset Valuation and Depreciation in the Model under Capacity Expansion

Yet economic asset valuation paths and depreciation schedules follow very different paths from anticipatory construction under capacity expansion. For one thing, there are now two assets, one serving initial demand (q) with an initial cost $K(q)$, and another, completed at the end of the T th year, to serve the demand growth (q') with a cost of $K(q')$. The asset valuation for the original asset during period 1 is based on the resale value of that asset were it to be sold prior to the end of period 1. However, this resale value depends on whether the resale occurs before or after τ_E years (§9.1.4). Consequently, the equation for asset valuation, derived from (9.18c) and (9.18d), changes at the transition point τ_E where the subsidy-free price increases, and is shown in (9.37a) and (9.37b). In the second period, the potential resale value of the original asset to a subsequent firm or self producer provides its asset valuation, as shown in (9.38).

$$V_i(q) = \left(\frac{\rho_N - \rho_{i-1}}{\rho_N} \right) K(q); \quad 1 \leq i < \tau_E + 1 \quad (9.37a)$$

$$V_i(q) = \frac{1}{dB_N} \left[K(q, q') - \left(\frac{1 + \rho_{i-1}}{1 + \rho_T} \right) K(q') - \left(\frac{1 + \rho_{i-1}}{1 + \rho_N} \right) K(q) \right]; \quad \tau_E + 1 < i \leq T + 1 \quad (9.37b)$$

$$V_i(q) = \frac{1}{dB_N} \left[K(q, q') - K(q') - \left(\frac{1 + \rho_{i-1}}{1 + \rho_N} \right) K(q) \right]; \quad T + 1 \leq i \leq N + 1 \quad (9.38)$$

For the later asset constructed to meet demand growth in the second period, the asset valuation is again its potential resale value, as shown in (9.39).

$$V_i(q') = \left(\frac{\rho_N - \rho_{i-(T+1)}}{\rho_N} \right) K(q'); \quad T + 1 \leq i < N + 1 \quad (9.39)$$

The interesting point to note about the value of the original asset is that at some stage during period 2—in fact at $T_{CE \rightarrow CR}$ years—it will become *negative*, as can be seen by setting the LHS of (9.38) to zero, and rearranging to produce (8.48b). At first this may seem contradictory; if the asset's value is negative, then it does not seem likely that resale would be feasible. Nevertheless, even with negative asset values the subsidy-free prices, asset values, and depreciation provision are all still consistent with contestability theory's stipulations for economic asset value and optimal depreciation, as provided in (7.1)-(7.5). Even when the annual return component of (7.2) is negative, because the asset value is negative, the annual depreciation provision (determined from the annual drop in asset value) more than offsets this component, and the overall payment to capital is still positive. Moreover, the payments to capital are consistent with the subsidy-free prices determined above (§9.1.4). This result is a marked departure from the usual practice of “writing off the books” assets which do not have a positive value. However, because the value of the asset (assuming the obligation to serve demand in perpetuity)

continues to drop even further below zero, the net payments to capital still sustain the business as a whole and support an investment which was optimal at the time which it was made. It still continues to be optimal from the incumbent's perspective. Only from the point-of-view of a potential entrant—or the consumers themselves (considering how to supply themselves on a greenfields basis)—does the existing investment appear sub-optimal.

The justification for this seemingly anomalous result—discussed further below (§9.4.2)—is that the value of the incumbent's business activities *as a whole* still remains constant, and satisfies (7.25) earlier. A feasible resale price exists for an incumbent firm wishing to sell up its entire operations—in other words, the original asset plus the accumulated depreciation. (As discussed earlier (§7.3.2), this “accumulated depreciation” is not necessarily a fund of cash reserves. Depreciation may have been reinvested in the business through the purchase of replacement or new network assets, or used to make other investments not part of the core network business). The accumulated depreciation may exceed the initial capital outlay, but this is offset by the negative value of the physical asset. The transfer price for the entire firm would however be positive, and equal to the initial capital outlay $K(q)$. But, once the asset price has fallen below zero, an incumbent wishing to sit on its accumulated depreciation would need to compensate any successor firm to which it offloads only its physical assets.

It is also interesting to note that when the original asset reaches the end of its lifetime, and it needs to be replaced, the incumbent firm is secure from competition for any level of demand between q and entire market demand. A competitor cannot potentially enter the market at the end of the T th year unless it attempts to supply entire market demand (i.e., a Type 2 entrant). Consumers will however be indifferent between the incumbent and a Type 2 entrant/self-producer. Consequently, by accelerating depreciation in advance through raising its price after τ_E years, the incumbent's prices are sustainable.

Hence, the situation raised by Hay and Morris (1993, p. 95) that, when a monopolist's capital “dies” the market is “up for grabs”, does not eventuate. Hay and Morris raise the concern because, from a sustainability point of view, once entry-detering capital is fully depreciated, it seems that the basic asymmetry between incumbent and potential entrant disappears. In BPW's terms, the deterring “sunk cost” is no longer present, since, as noted earlier (§3.5.2), the fact that some investment or salvage is undertaken at a particular date automatically transforms older capital of the same type that would otherwise be “sunk” into capital that is liquid “*at the margin*” (BPW, p. 381). But as Hay and Morris point out, the fundamental asymmetry between incumbent and entrant actually still remains, meaning that the incumbent will always win the game of “getting in first”. Even though the original asset fully depreciates prior to the end of its asset lifetime, if the entrant invests at that time, it has to contend with the fact that the incumbent's plant still has a number of periods of life. This will imply losses for the entrant during that period. On the other hand, if the potential competitor waits until the original asset

really has reached the end of its useful lifetime, the incumbent can match the price of any entrant serving less than entire market demand.

The overall investment value associated with the capacity supplying the incremental demand also retains its constant value, $K(q')$. As noted earlier (§8.1.1), where fixed costs are zero, the capacity expansion case can be served equally efficiently by a monopolist, or by two firms. The asset valuation paths described by (9.37a)-(9.39) not only are consistent with subsidy-free prices, but with subsidy-free operations associated with each of the two assets. Because the overall investment value associated with each asset remains the same as the initial capital outlay for that asset, either a monopolist or two firms could operate the two assets as distinctly separate self-sustaining businesses, without the need for cross-subsidies between them.

A monopolist could of course utilise some of the accumulated depreciation provision associated with the original asset as a source of funds for its capital outlay on the asset serving incremental demand. In this case, the overall value of the monopolist's firm, comprising both assets, would not be equal to the total initial cost of the two assets, but rather the value would be equal to the total external funds required to maintain current service potential.

9.2.3 *Asset Valuation and Depreciation in the Model under Capacity Replacement*

Under capacity replacement, the asset valuation for the original asset during period 1 is based on the resale value of that asset were it to be sold prior to the end of period 1. However, as discussed above (§9.1.5), this resale value depends on whether the resale occurs before or after τ_R years. Consequently, the equation for asset valuation, derived from (9.29c) and (9.29d), changes at the transition point τ_R , where the subsidy-free price increases, and is shown in (9.40a) and (9.40b).

$$V_i(q) = \left(\frac{\rho_N - \rho_{i-1}}{\rho_N} \right) K(q); \quad 1 \leq i < \tau_R + 1 \quad (9.40a)$$

$$V_i(q) = \frac{1}{dB_N} \left(\frac{\rho_T - \rho_{i-1}}{1 + \rho_T} \right) K(q, q'); \quad \tau_R + 1 < i \leq T + 1 \quad (9.40b)$$

By the end of the first period, the value of the original asset will always have fallen to zero (assuming the asset is non-fungible at any other location). This explains why the payments to capital, and thus the subsidy-free price increases after τ_R years. The price increases so that depreciation can be *accelerated* to recover the entire initial capital outlay over T years, rather than N years.

In the second period, the potential resale value of the new asset serving entire market demand to a subsequent firm or self producer provides its asset valuation, as shown in (9.41). Effectively, the valuation of this later asset follows exactly the same path as does the single asset under anticipatory

construction, the only difference being that the asset in this case begins service at the end of the T th year, and it is at that time when the asset valuation equals the initial capital outlay. Hence, (9.41) is very similar to (9.35), except that the indexes are adjusted to account for the later construction date.

$$V_i(q, q') = \left(\frac{\rho_N - \rho_{i-(T+1)}}{\rho_N} \right) K(q, q'); \quad T+1 \leq i < T+N+1 \quad (9.41)$$

Somewhat similarly to capacity expansion, when capacity replacement is optimal and fixed costs are zero, the most efficient market structure is not limited to a monopoly. Two firms could also supply the market at the same cost. However, whereas the two firms would operate contemporaneously in the second period under capacity expansion, where capacity replacement is optimal, separate firms could serve each period in sequence as monopolies. Consequently, an *intertemporal* monopoly, is not the only “natural” market outcome. Again, the investment value associated with each asset, one in the first period and one in the second period, remains constant. Consequently, not only do no cross subsidies exist between prices, neither do any subsidies exist between the businesses.

Nevertheless, even though having two sequential monopolists would not reduce efficiency, from a sustainability point of view, the incumbent monopolist has a distinct advantage over a potential competitor, and is thus likely to remain in place. Similarly to capacity expansion, the incumbent fully depreciates the original asset before the end of its lifetime. However, in this case the asset has been fully depreciated by the end of the first period, whereas under capacity expansion the asset does not fully depreciate until sometime between the T th and N th year. Although the intention of the incumbent under capacity replacement is to stop use of the original asset at the end of the first period, should a competitor enter the market, it can threaten to recommission the fully paid off original asset, and supply q units of demand at a negligible price.

9.2.4 *Justification for Subsidy-Free Prices in the Context of Perfect Contestability*

Given the discussion in the previous subsections, the subsidy-free prices resulting from the analysis are consistent with perfect contestability, and they are also intuitively “fair” (or equitable). The Bertrand-Nash assumption is upheld, there is perfect freedom of entry and exit for incumbents and competitors alike, and prices derived by assuming that an intertemporal level-playing field exists between all participants. Nevertheless, the passage of time is not ignored, since future costs are discounted and asset resale is allowed between sequential sub-periods. Moreover, the implicit first mover advantage of first period entities is recognised.

The subsidy-free price during period 1 will always be either $K(q)/qB_N$ or $K(q, q')/qB_N$, as shown in (9.8b), (9.21) and (9.32). Under anticipatory construction, it will be the higher of the two prices, since spare capacity is installed from the outset. It is justifiable for consumers in period 1 to pay this price, since, had they delayed their consumption, they could benefit in the greater demand to occur later without

increasing the required level of capacity. Hence, first period consumers are rightly required to pay the cost of “bringing forward” construction—a concept similar to Turvey’s (1969) concept of long run marginal cost (§4.1.3), and to the flip side of option value (§5.3.4)—and as such must contribute to the opportunity cost of the resultant spare capacity. If it appears that anticipatory construction is likely to be the optimal construction configuration, but it is not known exactly when demand will increase, the appropriate price until demand increases can still be determined with certainty. The subsidy-free price, prior to demand growth, is simply the amortised total cost of construction divided through by the current number of demand units. Consequently, as Marcel Boiteux suggests, spare capacity “has its own income” (§6.1.1).

Under capacity expansion, the *initial* price is the lower of the two possible payments to capital (i.e., $K(q)/qB_N$). However, an incumbent firm is completely justified in increasing the price during period 1 at year τ_E , as shown in (9.21). This is because, from that point in time forward, the opportunity cost of supply is based on building new capacity sufficient to serve *all* future market demand, rather than just the demand occurring in the first period. Furthermore, under capacity expansion, the set of subsidy-free prices allows both assets to be run as separate standalone businesses with no cross-subsidies between them (§9.2.2). Similarly, under capacity replacement, an incumbent firm can increase the price at year τ_R without incurring an intertemporal cross-subsidy (§9.2.3), since if the incumbent did not, it would be unable to fully recover the cost of the asset serving period 1 demand, and could not make a normal profit. Just as *supranormal* profits imply a cross subsidy, so do *subnormal* profits. Paraphrasing Ralph Turvey (§7.3.3), under either capacity expansion or capacity replacement the expectation of demand for the 5MVA tomorrow can be said to raise the cost of the 10MVA today (§9.2.5). In the words of Alfred Kahn which open this Chapter, there is a clear cost “causation” involved in Boiteux’s anticipatory construction case, and in Turvey’s capacity expansion and capacity replacement cases. The subsidy-free price always rises to the price that would be optimal on a greenfields basis (§6.3.5), which is (usually) the same as that which would be optimal for a potential entrant. This is not really paying for assets in advance, simply recovering the costs caused by the actions of current consumers, in light of the expected actions of future consumers (§7.3.3).

The price for utilising the q units of capacity during period 1 will always be higher than the price for utilising the same level of capacity in period 2. The period 2 price will always lie somewhere between $[K(q,q') - K(q')]/qB_N$ and $K(q)/qB_N$, depending on the year of demand growth, as shown in (9.11) by letting T tend to zero, in (9.22), and in (9.33) by letting T increase to N . This makes sense, because these consumers potentially benefit both the original period 1 consumers, as well as the incremental period 2 consumers. They allow the intertemporal cost of the original capacity to be recovered over a greater period of time and, except under capacity expansion, allow the costs of incremental capacity (which has a lower average cost) to be recovered over a greater number of consumers.

The relative price associated with the incremental units (q') will depend on the nature of the cost function K , and the level of demand growth. However, the price associated with q' will always lie between $[K(q, q') - K(q)]/q'B_N$ and $K(q')/q'B_N$, as shown in (9.12) by reducing T to zero, in (9.23), and in (9.34) by increasing T to N . Only under capacity expansion, where the incremental consumers are supplied by their own separate capacity, will they pay the full amortised cost of that capacity, $K(q')/q'B_N$.

If, given the specific characteristics of K , capacity expansion is not a feasible construction configuration, then the prices will still lie between the limits discussed above. However, depending on the year of demand growth, the price for \bar{p}_2 will never fall to the lower limit, and \bar{p}'_2 will never be as high as its upper limit. This is demonstrated through a simple example provided in the following subsection.

9.2.5 *A Simple Example of Subsidy-Free Price Paths for a Zone Substation*

The discussion in the previous subsection is best understood through a simple example. The cost function K , is considered to be the simple transformer cost function presented earlier (3.2). Total demand for capacity is considered to be 10MVA in the first period, and 15MVA in the second period. The two goods involved—namely the demand for capacity at different locations in a power distribution network—are thus for 10MVA, beginning in the first period, and 5MVA at a different location, beginning in the second period. Demand for both goods lasts in perpetuity. All assets have a lifetime of 50 years, and the discount rate is 8%.

Substituting the relevant parameters into the RHS of (8.49) indicates that capacity expansion can be feasible if asset lifetime exceeds 23.3 years. Since asset lifetime is 50 years, capacity expansion can be feasible. This means the relevant construction optimality expressions are the transition years from anticipatory construction to capacity expansion (8.3b), and from capacity expansion to capacity replacement (8.48b). The expression for transition between anticipatory construction and capacity replacement (8.47b) can be ignored, since it is dominated by the other transition equations where capacity expansion is feasible, as can be seen by examining expressions (8.44)-(8.46).

Performing these substitutions indicates that $T_{AC \rightarrow CE}$ equals 14.275 years, and $T_{CE \rightarrow CR}$ is just under 41 years. This means that: anticipatory construction is optimal if the second period—the onset (i.e., the year) of demand growth from 10MVA to 15MVA—begins within 14.275 years of the beginning of the first period; capacity replacement is optimal if the second period begins after 41 years; and capacity expansion is optimal otherwise.

Figure 9.1 shows the subsidy-free prices depending on when the second period begins (i.e., the year of demand growth). The price p' is the annual per unit payment to capital for the second good of 5MVA, the demand for which only begins after the end of the T th year. The value is taken from (9.12),

(9.23) and (9.34), whether anticipatory construction, capacity expansion or capacity replacement are optimal, respectively. The actual payment to capital (and thus the price path) throughout period 2 will be constant. Figure 9.1 is not a plot of the price path, but of the subsidy-free prices depending on when demand growth occurs. As discussed in the previous subsection, this annual price lies between $[K(q,q') - K(q)]/q'B_N$ and $K(q')/q'B_N$. Similarly, p_2 is the (constant) annual per unit payment to capital for the initial 10MVA, but in the second period only. It is drawn from (9.11), (9.22) and (9.33) for the three construction configurations respectively, and its value lies between $[K(q,q') - K(q')]/qB_N$ and $K(q)/qB_N$.

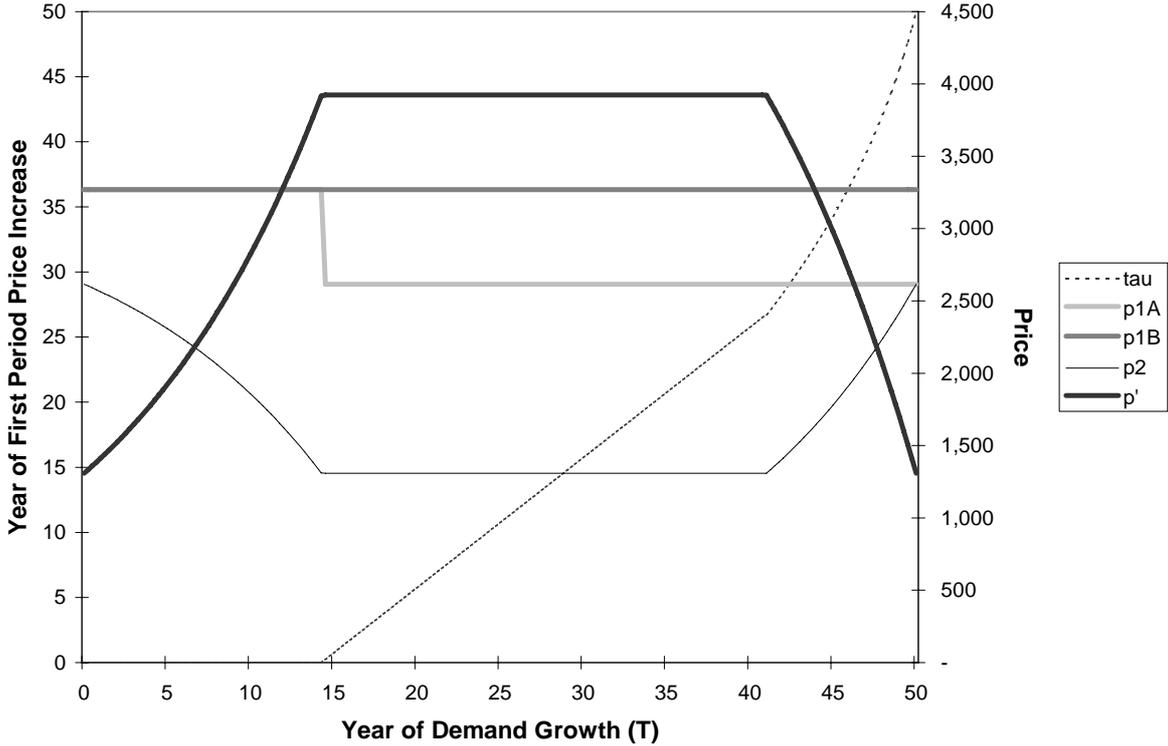


Figure 9.1: First and Second Period Subsidy-Free Prices for Transformer Capacity as a Function of the Year of Demand Growth (T)

First period prices are a little more involved, but are always either $K(q)/qB_N$ or $K(q,q')/qB_N$. Under anticipatory construction, the price for the first good remains constant throughout the entire first period, and is indicated in Figure 9.1 by the label p_{1B} , where T is less than 14.275 years. Where the year of demand growth is greater than this value, either capacity expansion or capacity replacement is optimal, and there is a price transition point during period 1. The curve labelled “tau” in Figure 9.1 shows the first period transition year. For T between 14.275 years and 41 years, the curve is derived from (9.16b) for τ_E , and for T greater than 41 years, the curve is derived from (9.27b) for τ_R . The curves p_{1A} and p_{1B} show the first period prices before and after the first period transition year respectively. The annual payments to capital before the transition point are constant, as they are after the transition point up until the end of period 1.

One example each of actual price and asset valuation paths are presented in Figures 9.2a-9.4b. In these plots the x-axis is no longer the year of demand growth, but the passage of time. For instance, Figure 9.2a shows the schedule of annual per unit payments to capital over a 50 year time interval for a given T , in this case, 10 years. In this plot, the price $p(q)$ is the price of the first good, and the price $p(q')$ is the price of the second good which has no value until after 10 years has passed (i.e., after the first period has elapsed). The three constant price values on this plot are simply the same values as can be found from Figure 9.1 for T equal to 10 years. The corresponding asset valuation path, this time over 100 years, is shown in Figure 9.2b. Since *total* payments to capital are constant, the valuation path follows the well-known economic depreciation path under constant payments to capital (§7.2.3).

Figures 9.3a and 9.3b show the corresponding price and valuation paths where T is 30 years; hence capacity expansion is optimal. The price increases after τ_E years—which in this case is 15.725 years—to reflect that anticipatory construction is now the optimal greenfields asset configuration. In this case the incumbent firm has two assets after 30 years, and Figure 9.3b shows the individual asset and combined asset valuation. Note that depreciation accelerates after τ_E years; a consequence of the payments to capital increasing at that time. After T years, the value of the original asset rapidly declines and goes *negative*. Nevertheless, the total asset value is positive, and the value of the firm as a whole will equal the combined external funding amount. Where T is 45 years, capacity replacement is optimal. Payments to capital again increase in the first period, this time after τ_R years—which in this case is 34.3 years—as shown in Figure 9.4a. Figure 9.4b indicates that depreciation accelerates after this date so that the entire initial capital outlay can be recovered before the onset of demand growth.

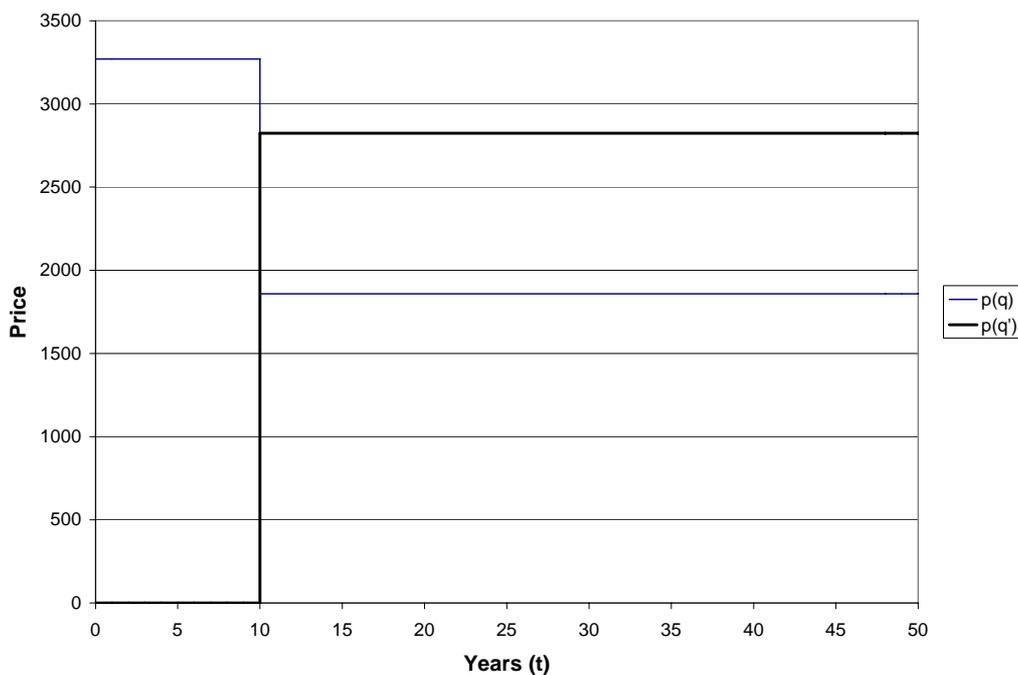


Figure 9.2a: Annual Subsidy-Free Prices under Anticipatory Construction ($T = 10$ years)

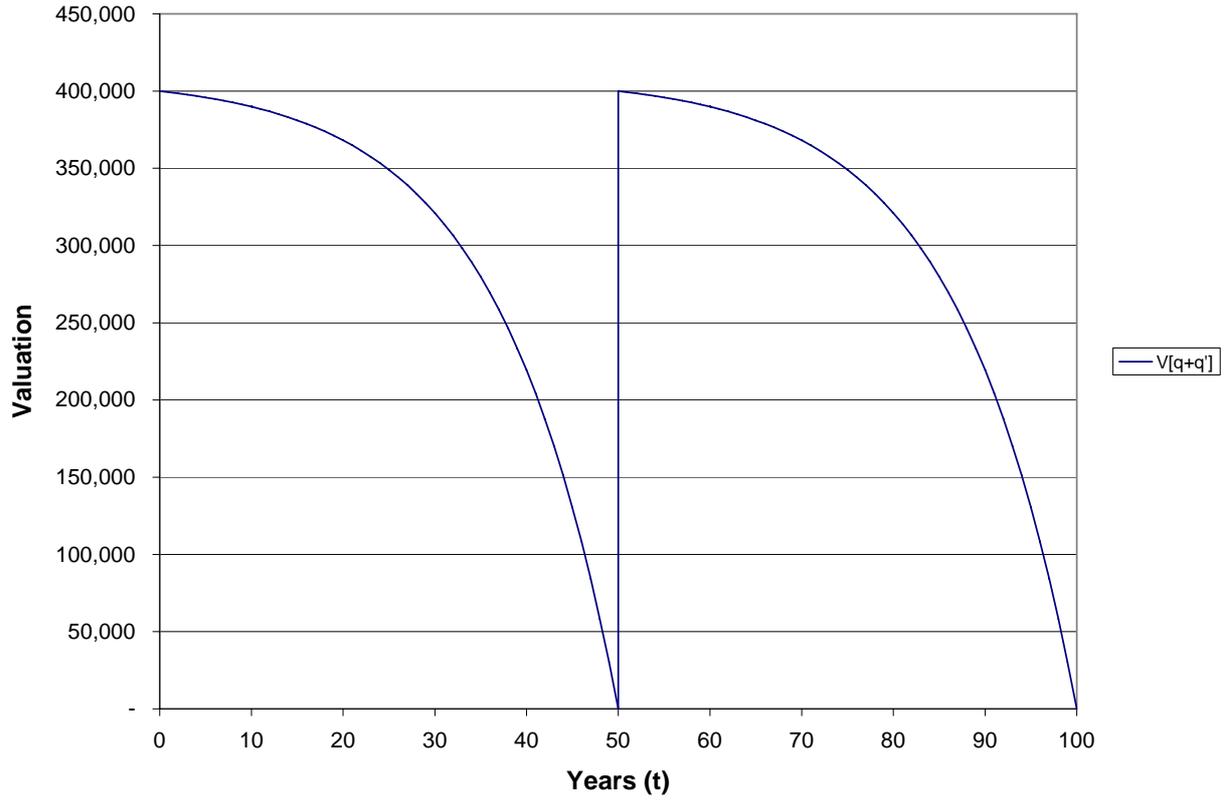


Figure 9.2b: Economic Asset Value under Anticipatory Construction ($T = 10$ years)

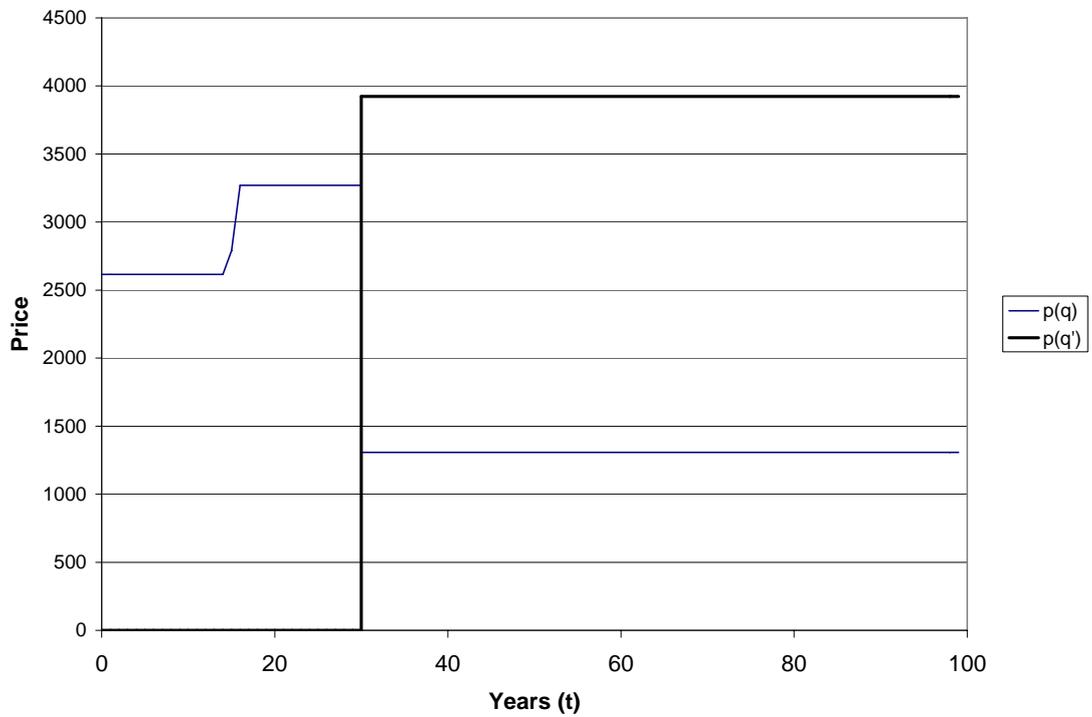


Figure 9.3a: Annual Subsidy-Free Prices under Capacity Expansion ($T = 30$ years)

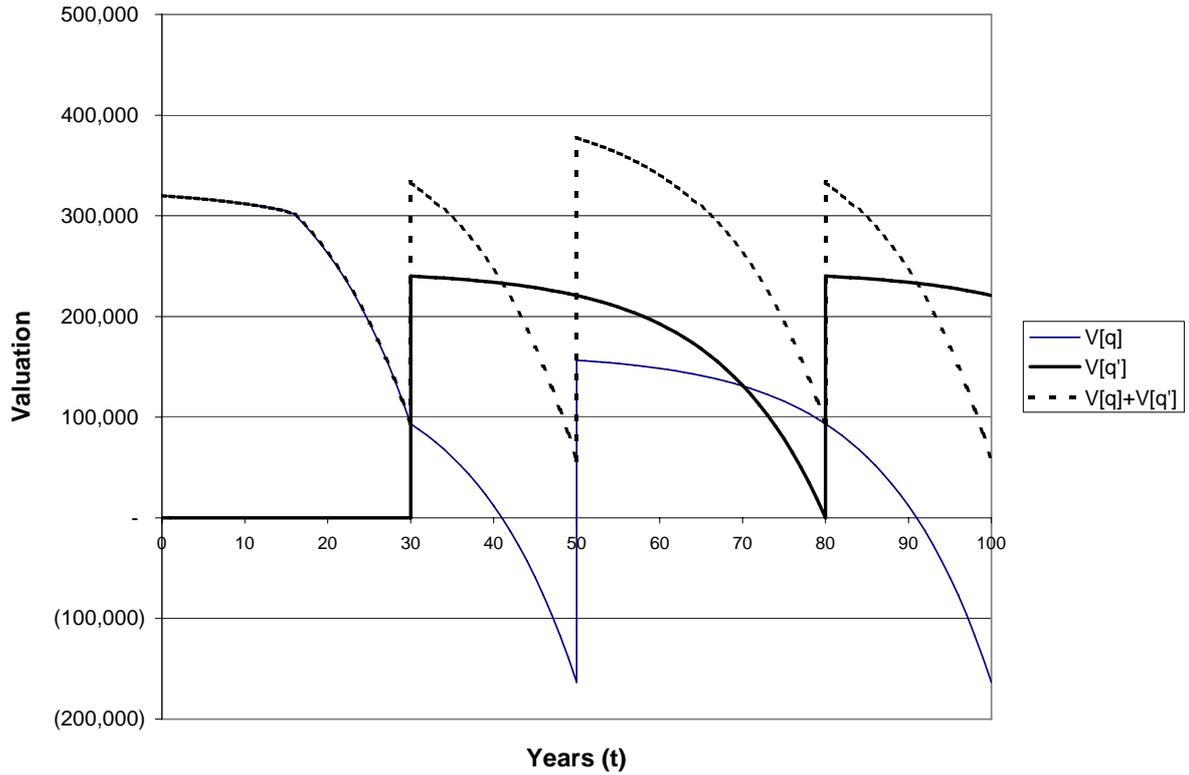


Figure 9.3b: Economic Asset Value under Capacity Expansion ($T = 30$ years)

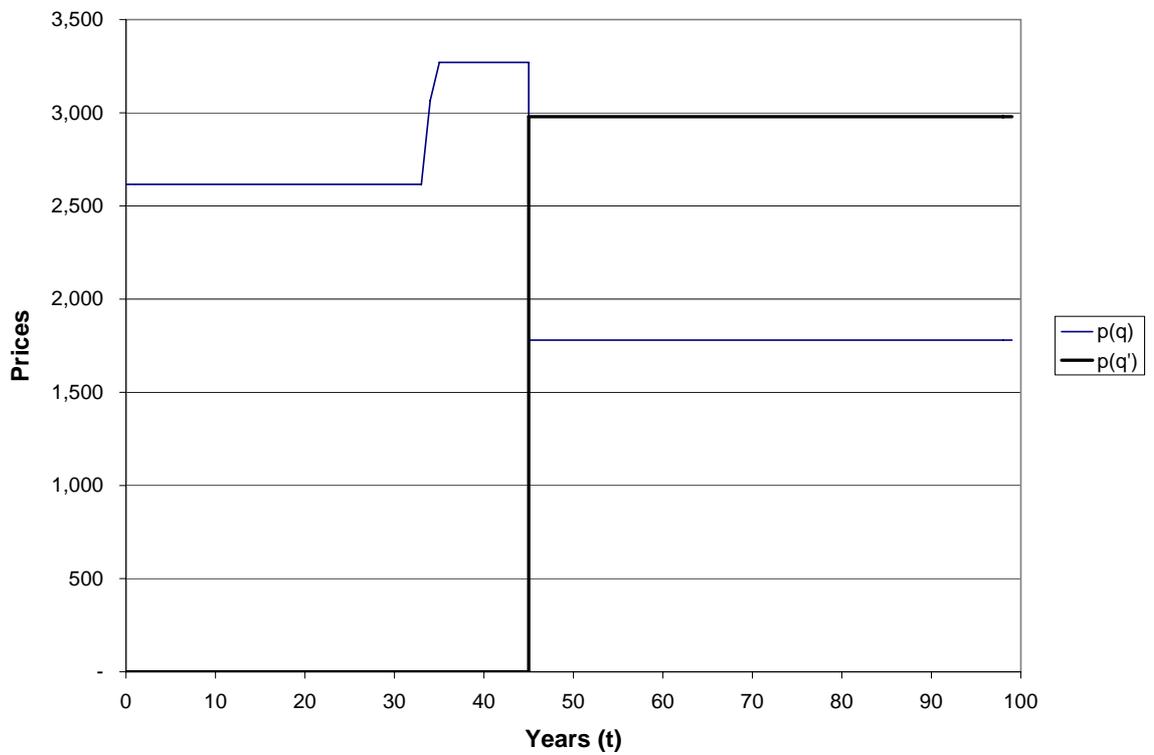


Figure 9.4a: Annual Subsidy-Free Prices under Capacity Replacement ($T = 45$ years)

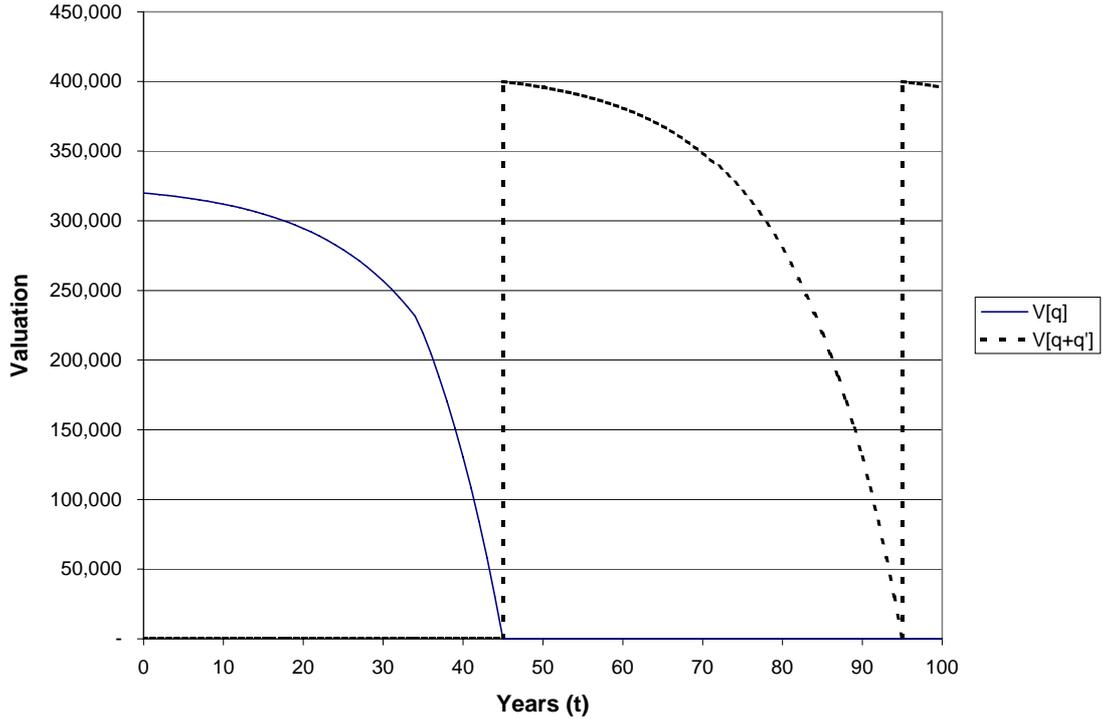


Figure 9.4b: Economic Asset Value under Capacity Replacement ($T = 45$ years)

9.3 The Efficiency and “Fairness” of Line Charges in New Zealand

9.3.1 Price Paths for Subsidy-Free, LRIC, and ODV Pricing under Anticipatory Construction

The same approach as that taken in the previous Section can be applied to the zone substation example presented earlier (§6.2.2-§6.2.4). All the usual simplifying assumptions, including no change in technology and no inflation, are made. Where load increases from between 15-22.5MVA to 22.5MVA within 16 years (or 15.06 years exactly), anticipatory construction will be the optimal construction configuration, and this involves a 2x20MVA substation design. Consider as an example the case where demand growth increases above 22.5MVA at the end of the 11th year after substation completion. The total cost of the 2x20MVA substation is \$1,204,000, or $K_{AC1}(2x20MVA)$, and from (6.2), the present value cost of the substation in perpetuity (TC_{AC}) is \$1,230,230. The difference between the initial construction cost, and the present value of current and ongoing capital costs, is very small because the substation lifetime is a lengthy 50 years. From (9.8b), it can be inferred that the subsidy-free annual payments to capital are constant, and are simply the total construction cost divided through by B_N in each year—in other words $K_{AC1}(2x20MVA)/B_N$ —which means the subsidy-free annual revenue associated with the substation should be \$98,418.

For contrast, the long run marginal (or incremental) cost-based prices, and the ODV-based cost prices—those based on a return on and return of capital based on its underlying optimised deprival value (§7.4.2)—are also calculated. Long run incremental cost is here determined using Turvey’s (1969) approach to calculating marginal cost, which is based on the cost of bringing forward investment by one

period (§4.1.3). This provides a very different result from taking an approach which considers capital costs to be “sunk” *ex post*, which would mean that the marginal costs of capacity would be zero (§4.1.4). Turvey’s approach provides the following equation (9.42) for the LRIC of capacity in each year.

$$\begin{aligned} \text{LRIC}_{AC} &= (1+d) \left[\text{TC}_{AC} - \frac{\text{TC}_{AC}}{1+d} \right] \\ &= (1+d) \left[\frac{1}{dB_N} K(2 \times 20\text{MVA}) + \frac{K(2 \times 20\text{MVA})}{dB_N(1+d)} \right] = \frac{K(2 \times 20\text{MVA})}{B_N} \end{aligned} \quad (9.42)$$

Hence, it can be seen that, at least in the case of anticipatory construction, setting prices to the long-run incremental cost of the substation capacity, would provide the *same* outcome as determining prices on a subsidy-free basis. Such a result was also the case where there was no demand growth and capacity remained constant (7.2.3). The long-run incremental cost in (9.42) recognises the opportunity cost of tying up resources in the substation assets, the costs of replacing those assets at some later point in the future, and the opportunity cost of capital (i.e., d), and as such is also consistent with Schramm’s (1991) definition of long-run incremental cost (§4.1.5), as well as Baumol and Sidak’s (1994a) list of the costs to be considered in prices (§6.3.1). The depreciation schedule derived from the subsidy-free or LRIC-based prices will also follow the path outlined in (7.9).

ODV-based prices are, however, markedly different from the subsidy-free or LRIC-based prices. In its most general sense, the ODV approach of optimising the asset base and subsequently depreciating the assets, could be consistent with the subsidy-free prices above. The optimisation step of the ODV methodology is fully consistent with considering the optimal greenfields construction configuration for the current and future set of demands from the perspective of a potential entrant, or a group of self-producing consumers. However, in practice, straight-line depreciation is used. This means that the annual payments to capital (PC_i) recovered from any asset of initial cost K , and lifetime N , are simply as provided below in (9.43), based on (7.2) with $r \equiv d$.

$$PC_i = dV_i + D_i = dV_i + \frac{K}{N} = K \left[d \left(1 - \frac{i-1}{N} \right) + \frac{1}{N} \right] = K \left(df_i + \frac{1}{N} \right) \quad (9.43)$$

where: f_i is the straight line depreciation factor equal to $(N - i + 1)/N$. Because the return on asset component of the ODV-based price is indexed to a depreciating asset value, and the return of asset component is constant, the payments to capital decrease over the asset’s lifetime. Assets valued using straight line depreciation do meet the economic asset value criterion provided in (7.1) for any year during the asset’s lifetime, subject to an assumption of zero economic profit. Hence, substituting (9.43) into (7.1), and defining $r \equiv d$, provides the economic value of an asset under straight-line depreciation (for $1 \leq i \leq N$).

$$V_i = \sum_{j=i}^{\infty} (rV_j + \bar{D}) \left(\frac{1}{1+d} \right)^{j-i+1} - \sum_{k=1}^{\infty} I_{kN+1} \left(\frac{1}{1+d} \right)^{kN-i+1} = K \frac{1}{d} \left(r\phi_i + \frac{1}{N} \right) - K\Gamma_i = Kf_i = V_i \quad (9.44a)$$

$$\text{where: } \phi_i \equiv d \left[B_{N+1-i} - \frac{1}{N} (\Delta_i - B_{N+1-i}) + \Gamma_i B_N \delta \right] = f_i + \delta - \frac{B_{N+1-i}}{B_N} = f_i - \frac{1}{dN} + \Gamma_i \quad (9.44b)$$

The terms B and Γ are defined in (6.1) and (7.7b) respectively, and δ and Δ as follows in (9.45) and (9.46a) below, where Δ represents the summation of a finite geometric series where the exponent and the multiplying factor are not the same, but both uniformly increase.

$$\delta \equiv \frac{1}{d} \left(\frac{1}{B_N} - \frac{1}{N} \right) \quad (9.45)$$

$$\Delta_i \equiv \sum_{j=i}^X j \left(\frac{1}{1+d} \right)^{j-i+1} = (1+d)^{i-1} (A_N - A_{i-1}); \text{ for } 1 \leq i \leq N \quad (9.46a)$$

$$A_X \equiv \sum_{j=0}^X j \left(\frac{1}{1+d} \right)^j = \frac{(1+d) \left[1 - (X+1) \left(\frac{1}{1+d} \right)^X + T \left(\frac{1}{1+d} \right)^{X+1} \right]}{d^2} = B_X \left(1 + X + \frac{1}{d} \right) - \frac{X}{d} \quad (9.46b)$$

Hence, capital cost outlay is recovered where ODV-based prices are applied under anticipatory construction. This of course assumes that anticipatory construction is recognised as the optimal construction configuration. A difficulty arises under the Ministry of Economic Development's approach to implementing the ODV method (Energy Markets Regulation Unit, 2000c), since the planning horizon considered for network optimisation is limited to ten, and for some assets, five years (§7.4.4). Although this 2x20MVA substation example is just an example, the implication is clear. Where planning horizons are restricted, this may lead to sub-optimal investments.

In this simple substation example, because demand is not considered to increase until after the 11th year, the optimal construction configuration under a limited ten year planning horizon would be to match capacity to *initial* demand. Hence, when demand actually increases in the 12th year, this will, by default, result in a program of capacity expansion, rather than anticipatory construction, over time. This has two possible effects. Firstly, for new greenfields investments, there may be an incentive to invest sub-optimally. Secondly, already "sunk" investments, that were correctly undertaken using the optimal investment, might be deemed suboptimal (given the shorter planning horizon). This would mean that "allowable" returns on those assets would be lower than needed to make a zero economic profit.

Figure 9.5, shows the different price (i.e., annual payments to capital) paths under subsidy-free pricing, LRIC-based pricing, as well as ODV-based pricing under both the optimal construction configuration of anticipatory construction, and the sub-optimal configuration of capacity expansion. All price paths allow full recovery of capital outlays. However, the ODV-based price path where capacity

expansion is the actual investment program, will recover greater revenues than the other options, since the construction configuration is not least cost.

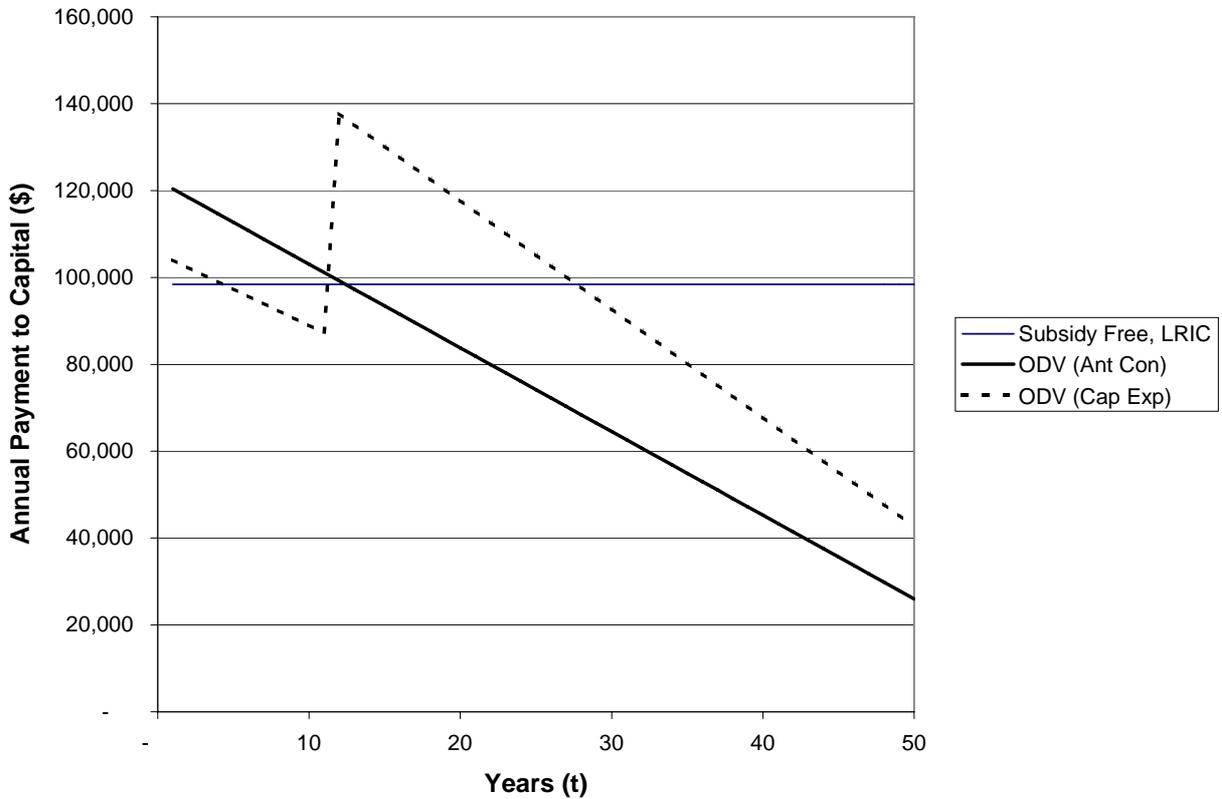


Figure 9.5: Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing (Anticipatory Construction Optimality)

9.3.2 Price Paths under Capacity Expansion and Capacity Replacement

Consider next an example where the demand growth increases above 22.5MVA at the end of the 20th year after substation completion (i.e., T equals 20), rather than the 11th year. This now means that capacity expansion becomes the optimal construction configuration. In this case, as discussed earlier (§6.2.4), capacity expansion requires an initial zone substation of 2x15MVA to be expanded to a 3x15MVA configuration. The initial capital cost is \$1,040,667, or $K_{CE1}(2x15MVA)$, and the expansion cost in the 21st year is \$520,333, or $K_{CE2}(1x15MVA)$. The present value cost of this investment program in perpetuity (TC_{CE}) is \$1,077,019.

From (9.21)-(9.23), the total annual subsidy-free payments to capital are thus $K(2x15MVA)/B_N$, before τ_E years, and $K(2x20MVA)/B_N$ afterward. These values correspond to \$85,067 per annum initially, rising to \$98,418 per annum—the annual level under anticipatory construction. Hence, the payments to capital never rise to a level of $K(3x15MVA)/B_N$, which would be the price level required to recover the cost of capital if a 3x15MVA transformer configuration were initially optimal. The price transition point can be found from (9.16b). Where capacity is optimally expanded to meet demand

growth at 20 years, this means that τ_E equals 20 years less 15.06 years—the transition point of optimality between anticipatory construction and capacity expansion—which is 4.94 years. Hence, under capacity expansion, capital cost recovery occurs *in advance* through accelerated depreciation in the period between τ_E and T years. However, this accumulated accelerated depreciation compensates the firm for the under-recovery of capital which occurs after capacity is expanded.

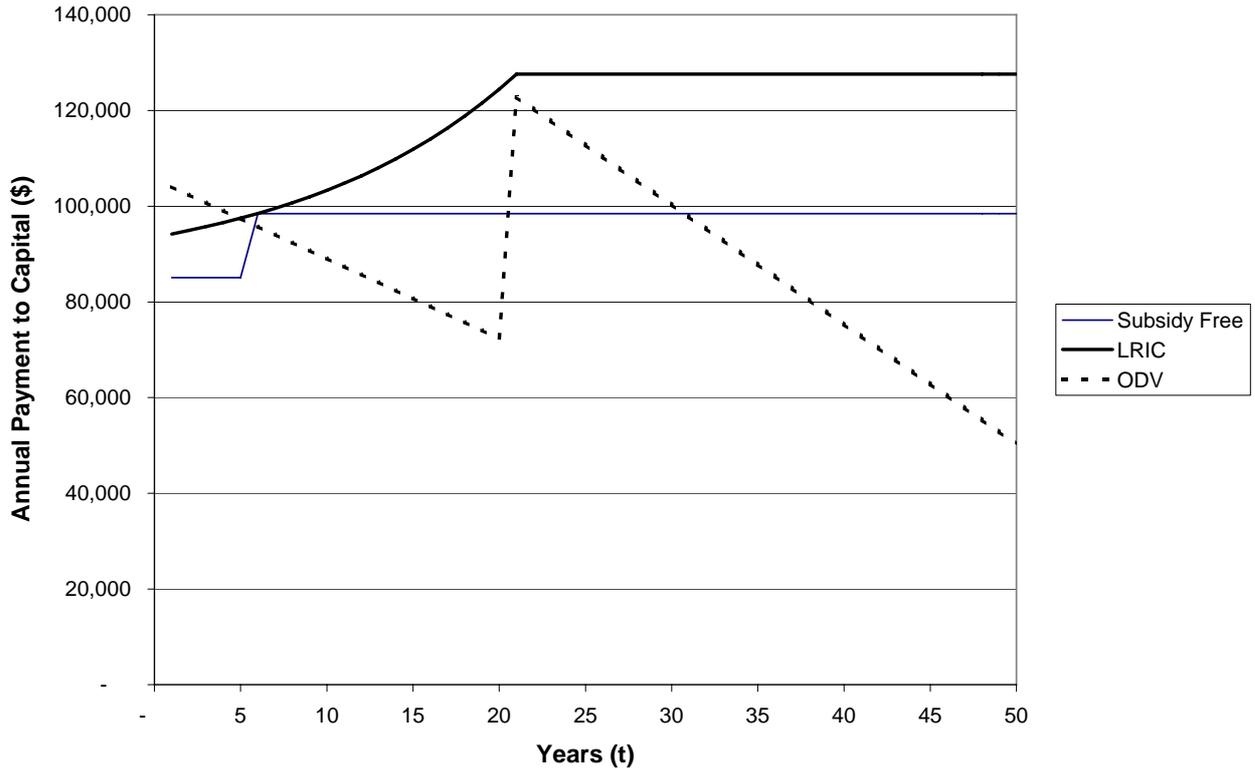
The annual LRIC-based price under capacity expansion can be found as follows from (9.47a) and (9.47b). Unlike anticipatory construction—where LRIC-based prices are the same as subsidy-free prices and allow a zero economic profit to be made—under capacity expansion, the LRIC prices exceed the subsidy-free prices, except at one point, τ_E years, where the two prices are equivalent. Consequently, under strict LRIC-pricing, the incumbent firm constructing the zone substation would receive a substantial positive economic profit. The LRIC-prices rise gradually toward the time of capacity expansion, “signalling” in advance the need for additional capacity. This differs from the reason that subsidy-free prices increase, which is related to the *currently* and *hypothetically* optimal greenfields configuration, rather than to *actual* capital investment in *future*.

$$\text{LRIC}_{\text{CE}_i} = \frac{(1+d)}{dB_N} \left[\begin{array}{l} K_{\text{CE}_1}(2 \times 15\text{MVA}) + \frac{K_{\text{CE}_2}(1 \times 15\text{MVA})}{1 + \rho_{T-i+1}} \\ - \frac{1}{1+d} \left(K_{\text{CE}_1}(2 \times 15\text{MVA}) + \frac{K_{\text{CE}_2}(1 \times 15\text{MVA})}{1 + \rho_{T-i+1}} \right) \end{array} \right]; \text{ for } i \leq T \quad (9.47a)$$

$$\text{LRIC}_{\text{CE}_i} = \frac{(1+d)}{dB_N} \left[K(3 \times 15\text{MVA}) - \frac{K(3 \times 15\text{MVA})}{1+d} \right] = \frac{K(3 \times 15\text{MVA})}{B_N}; \text{ for } i > T \quad (9.47b)$$

The ODV-based price path is derived from (9.43) by depreciating the initial investment, $K_{\text{CE}_1}(2 \times 15\text{MVA})$, and the expansionary investment, $K_{\text{CE}_2}(1 \times 15\text{MVA})$. In this case, the planning horizon makes no difference to the “allowable” ODV-based price path. The investment is not sub-optimal, and the present value of revenue exactly covers the present value of capital costs. This ODV-based price path is contrasted with the LRIC-based price path, and the subsidy-free price, in Figure 9.6.

Finally, consider an example where capacity replacement would be the optimal construction configuration, as would be the case if T —the year of demand growth—is 45 years. Again, as discussed before (§6.2.4), capacity replacement requires an initial zone substation of 2x15MVA to be modified by replacing the transformers in a 2x20MVA configuration. The initial capital cost is \$1,040,667, or $K_{\text{CR}_1}(2 \times 15\text{MVA})$, the same as under capacity expansion, but the modification or replacement cost of installing the 2x20MVA transformers in the 46th year is \$990,000, or K_{CR_2} . The present value cost of this investment program in perpetuity (TC_{CR}) is \$1,038,479.



**Figure 9.6: Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing
(Capacity Expansion Optimality)**

From (9.32)-(9.34), the total annual subsidy-free payments to capital are thus $K(2 \times 15\text{MVA})/B_N$, before τ_R years, and $K(2 \times 20\text{MVA})/B_N$ afterward. These values—just as under capacity expansion—correspond to \$85,067 per annum initially, rising to \$98,418 per annum. The price transition point can be found from (9.27b). Where capacity is optimally expanded to meet demand growth at 45 years, this means that τ_R equals 32.5 years. Hence, the accelerated depreciation schedule after this date under capacity replacement allows the capital cost of equipment replaced after 45 years—namely the 2x15MVA transformers—to be fully recovered.

The annual LRIC-based prices under capacity replacement reflect that the transformer-specific investments made initially do not need be replaced at the time of demand growth, unlike the transformers themselves. Under capacity replacement, LRIC-based pricing *under-recovers* capital expenditures on a present value basis, and again, the LRIC and the subsidy-free prices are equivalent at the transition point τ_R . The LRIC-based price rises markedly toward the date when the transformers must be replaced, as can be seen from Figure 9.7, which contrasts these prices against the subsidy-free prices as well as the ODV-based prices. While the LRIC-based prices under-recover total costs, the ODV-based and subsidy-free based prices both make zero economic profits on the least cost construction configuration.

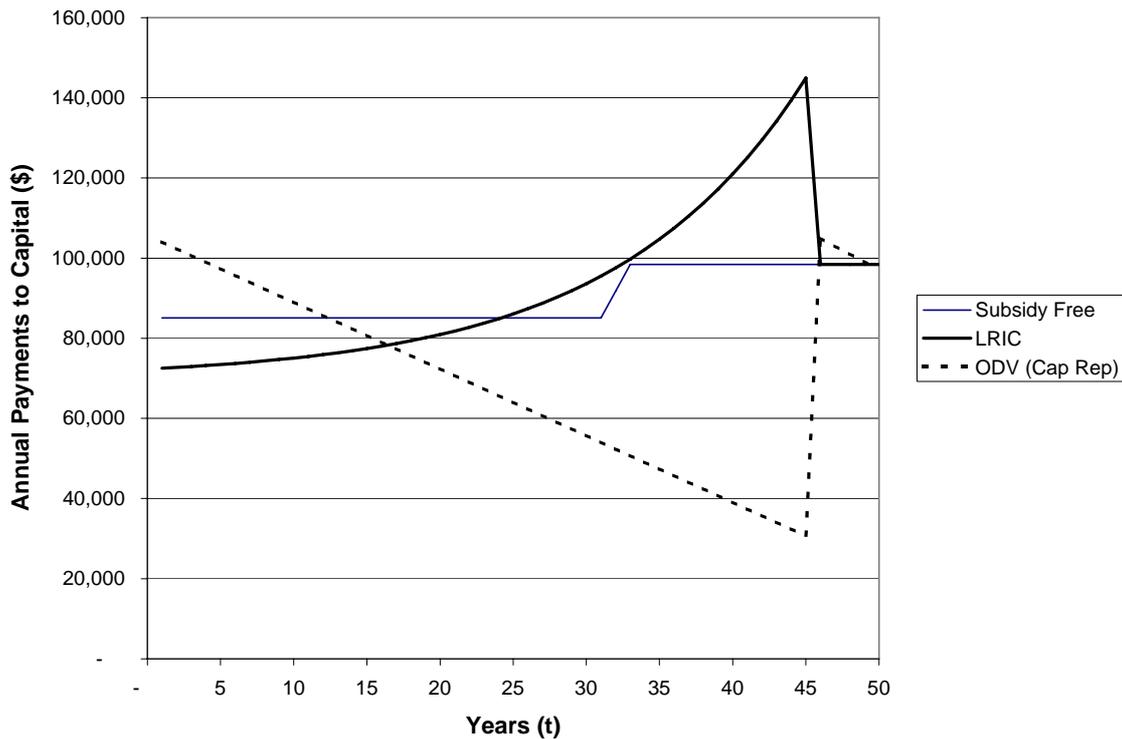


Figure 9.7: Zone Substation Price Paths under Subsidy-Free, LRIC and ODV Based Pricing (Capacity Replacement Optimality)

9.3.3 Evidence of Monopoly Rents

The results in the previous sub-sections might provide some indicative results regarding the prospect that New Zealand’s ELBs are receiving monopoly rents from their network assets, given that many of them appear to be implementing ODV-based pricing. As seen earlier (§7.5), the ODV methodology, and particularly its use as a basis for pricing, has been subjected to considerable criticism. The key concern is that the approach may not protect consumers from monopoly pricing, with even the Government suggesting that consumers are being “fleeced” (§2.4.6).

The first question considered here is whether the ODV optimisation process effectively nets out any sub-optimal investment, or conversely whether—given the short planning period—optimal investments are not allowed to be adequately recovered. Clearly, restrictions on the planning horizon—which is a feature of the *implementation* of the ODV approach, rather than an inherent principle of the methodology itself—can impact the optimality of investment. Given that in such a case a sub-optimal investment is considered to be optimal, ELBs in such a situation can fully recover their capital outlays. The negative consequences of the sub-optimal investment are thus faced by higher prices to consumers.

The second question is—assuming that the ODVs of the ELBs are in fact optimal and also appropriately reflect up-to-date replacement costs (which are major assumptions)—whether a return on and return of the ODV asset base is fair from *both* a consumer as well as an ELB perspective. The

analysis in the previous sections indicates that where anticipatory construction is the optimal construction configuration, the costs of spare capacity should be borne by current consumers, and not just the future consumers for whom that spare capacity is constructed. However, assuming that the planning horizon under ODV-based pricing correctly identifies that that spare capacity is optimal, prices will allow the incumbent ELB to recover its capital outlay. On the other hand, consumers in earlier years pay significantly higher capital cost contributions than consumers in later years, which as discussed earlier (§6.4.3), has no basis under intertemporal subsidy-free pricing.

Under capacity expansion, intertemporal subsidy-free prices allow the costs of later asset replacement of original assets to be recovered “in advance”. While this may appear “unfair”, the price increase involved is significantly lower than the increase in prices that would occur under ODV-based pricing in the year of demand growth, and is designed to accumulate the funds necessary to offset losses made in future years in order to make a zero economic profit overall. (This contrasts with LRIC-pricing which not only recovers the costs of assets in advance, but makes a significant positive economic profit).

Given that the average age of network assets in the ELBs is about 23 years,⁴ it seems likely that many ELBs have been earning returns far below a level that reflects the opportunity cost of subsidy-free supply. Figure 9.8 shows the ratio between the actual payments to capital received during either the 2000/2001 or 1999/2000 financial year—depending on the availability of disclosure data—and the minimum possible subsidy-free payments to capital, for most of the ELBs. “Minimum possible subsidy-free payments to capital” are derived from each ELB’s ORC (§7.4.2), divided through by B_N . This is the *minimum* possible total payment, because the actual subsidy-free total payment would be determined from summing the subsidy-free payments to capital from each and every subnetwork—particularly on a zone substation basis—throughout the entire network. Where those subnetworks are constructed under anticipatory construction, or when capacity expansion or replacement is optimal, and the age of the assets is less than τ_E and τ_R years respectively, then ORC/B_N will approximate the subsidy-free payment to capital, assuming that the optimisation process has been performed appropriately. In other cases, ORC/B_N will be lower than the subsidy-free prices, since after the price transition points under capacity expansion and capacity replacement, depreciation should be *accelerated*, and the total allowable payments to capital could be higher without breaching subsidy-free requirements.

Figure 9.8 shows this ratio based on two different WACC values. A typically acceptable maximum for post-tax nominal WACC is around 8%. However, for a nominal WACC to be used requires that any revaluation gains—due to the appreciation of asset value from inflation and other factors affecting replacement costs—be treated as income. Hence, the subsidy-free payments to capital derived here, for a single year, should more appropriately be based on a real WACC, given that an

⁴ Based on 1999 Disclosure Data for ten of the larger ELBs.

optimal greenfields network design is based predominantly on current costs. Assuming a real WACC of around 6% corresponds to an appreciation in annual replacement costs of just under 2%.

Observing Figure 9.8 indicates that for a real WACC of 6%, only 9 of the 26 ELBs examined here are earning a sufficient return from an intertemporally subsidy-free viewpoint. (For a nominal WACC, only two ELBs are earning sufficient returns). Of course, a single year’s results ignoring revaluation gains cannot be used directly to infer that there is a lack of monopoly rents in New Zealand’s power distribution sector—all it can do is suggest whether the overall payments to capital are “intertemporally subsidy-free” in that year. A full analysis would require scrutiny of how revenues and costs are allocated to the monopoly activities of each ELB. In addition, the ORCs gleaned from ELB disclosure information may not reflect an optimal network design. If so, the “fair” subsidy-free payments to capital might be overestimates. On the other hand, it is assumed the ORCs are based on up-to-date modern equivalent asset values. Also, the “fair” payments to capital derived here are minima. Where capacity expansion or replacement is optimal for many subnetworks, then the sum of all the subsidy-free revenues could be much higher. Hence, considerably more analysis is required before a final conclusion could be drawn. Nevertheless, this thesis has presented a new framework for addressing the issue of efficient and “fair” pricing for power distribution services in New Zealand—one that explicitly considers the intertemporal price preferences of consumers.

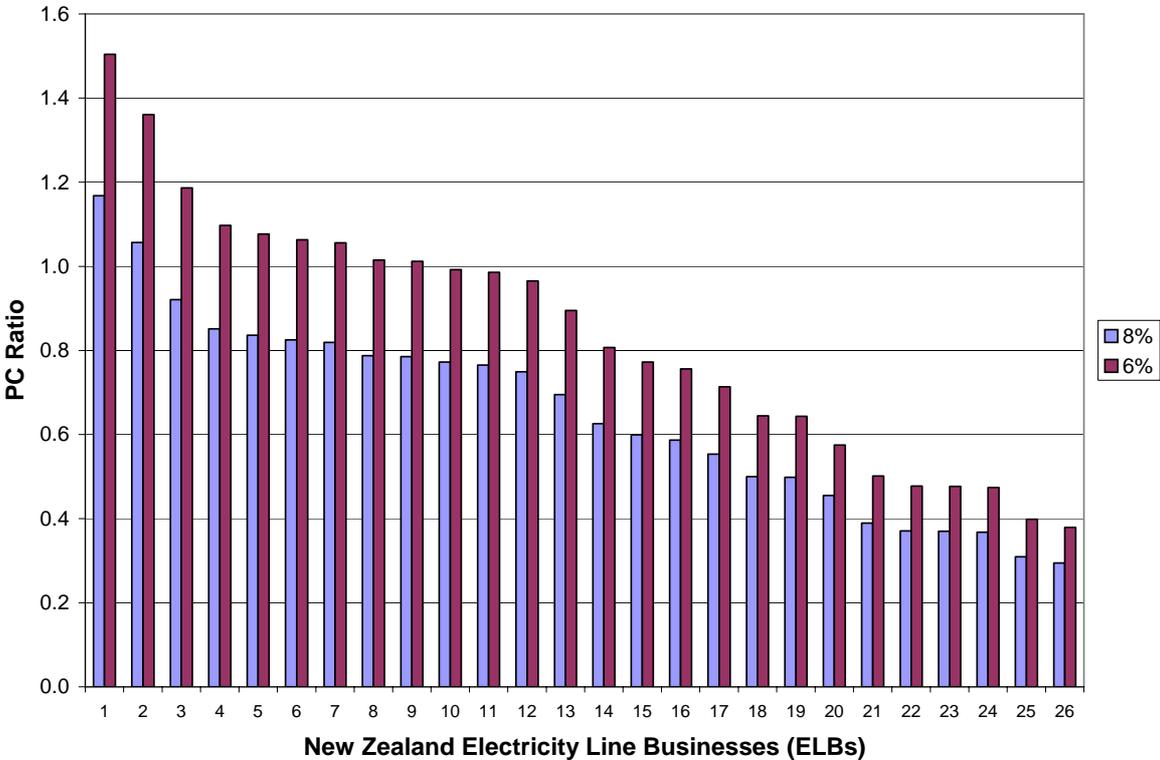


Figure 9.8: Ratio of Actual Payments to Capital to Minimum Economic Payments to Capital for New Zealand ELBs (under Different WACCs)

9.4 Limitations of the TGTP Model

9.4.1 Implications for Regulatory Policy

This thesis has posed the question: what are the characteristics of efficient and “fair” line charges for power distribution network services? This question has been posed in the context of New Zealand’s light-handed regulatory regime for electricity lines businesses. Although the results of the TGTP model suggest that the ELB price levels arising under the information disclosure regime do not necessarily have the characteristics of intertemporal subsidy-free prices, the model itself is not intended to explain why such is the case.

Instead, the main contribution of this thesis, and of the TGTP model itself, is to highlight that a key limitation of the standard formulation of the constrained market pricing approach—derived from within a contestable markets framework—is the way in which the time dimension is treated. The TGTP model has explicitly introduced intertemporal interdependence into constrained market pricing by extending BPW’s single-good/two-period model of intertemporal unsustainability to a TGTP model. While subsidy-free prices are generally derived ignoring the effect of current and future demand (§4.3.6)—an approach that stems back to Faulhaber (1975)—the TGTP model demonstrates that taking account of demand can have a dramatic impact on the bounds on subsidy-free prices, and can in some cases drive these bounds to an equality. In such cases, the depreciating economic value of an incumbent monopolist’s assets can then be directly derived from the future stream of intertemporal subsidy-free prices.

Even assuming the TGTP model were sufficiently expanded and refined to the extent that policymakers could have confidence in implementing policies which would provide incentives for line charges to move toward the prices arising from such a model, the appropriate policies might not necessarily be through more heavy-handed regulation, or at least through greater regulation alone. In New Zealand’s case, overall ELB price levels in many cases appear to have been lower than one might expect from a monopolist operating under a relatively light-handed regulatory regime of information disclosure (§9.3.3). Hence, while a tighter regulatory regime could perhaps be warranted to address issues with those ELBs at the higher end of the pricing spectrum, at the other end, policies addressing firm-specific governance issues might be more pertinent. However, these are questions that this thesis has not attempted to address. Rather, as outlined at the outset (§1.2.2), the focus has been on determining some of the characteristics of “benchmark” efficient and fair line charges, taking intertemporal interdependence into account.

Assuming that changes in regulatory policy are in fact the appropriate means to address prices that diverge from these benchmark levels, the extent to which regulators might use any model for guidance in designing a real-world regulatory regime for ELBs—whether in New Zealand or elsewhere—is limited by the simplicity of, and assumptions inherent in, that model as well as by the

limitations of the wider paradigm within which it has been developed. In the TGTP model's case, the paradigm is contestability theory.

As noted earlier (§2.1.7), Baumol and his colleagues argued that the paradigm of perfect competition, while self-contained and internally consistent, requires that prices be set to benchmark levels based on marginal cost—a prescription seen as leading to financial disaster (BPW, pp. 503-504). The model of perfect contestability was in part developed to address some of the shortcomings of perfect competition from a regulatory perspective. However, while BPW advocated replacing perfect competition with perfect contestability as the appropriate benchmark for regulatory intervention, the paradigm still only provides *guidance* to regulators with respect to both price and entry, and not a prescriptive solution. Furthermore, for subsidy-free prices derived from within a contestable markets paradigm, Baumol and Sidak (1994a, p. 43) also stressed that bounds on subsidy-free prices are determined neither to explain nor to predict actual market behaviour, but rather to provide guidance to regulators where the real world falls short of the perfectly contestable ideal (§5.1.1).

Apart from the treatment of the time dimension, the other key limitation of the traditional constrained market pricing approach is that it does not take into account uncertainty and risk. Similarly, the TGTP model sheds little light on the impact of uncertainty, because perfect information is a fundamental axiom of contestable markets theory. For instance, BPW's text on contestability limits its discussion of uncertainty to that of forecasting future prices, within the context of the Bertrand-Nash assumption (BPW, Ch. 13). Given that demand plays such a key role in determining intertemporal subsidy-free prices, and given that uncertainty in demand is an inherent factor of the electricity industry, extending the TGTP model to incorporate the effects of uncertainty would be an important area for future research.⁵

Some possible issues to be considered regarding the impact of uncertainty are considered below. However, before turning to these, some other possible limitations of the assumptions underpinning the TGTP model are also worth highlighting.

⁵ Unlike contestability theory, the real options theory of investment, extensively articulated in Dixit and Pindyck (1994), explicitly takes uncertainty into account in evaluating investment decisions involving irreversible assets. However, as Small and Ergas (1999) note, the real options literature mainly considers unregulated firms. Small and Ergas cite Teisberg (1993)—who considers the interaction of regulatory risk and market uncertainty on the size of electric utility investments—as a notable exception, and also point to Chapter 9 of Dixit and Pindyck's text, in which regulatory constraints are incorporated into a continuous time model of real option values. One of the few studies involving the application of real options theory specifically to investments in electric power lines under uncertainty is by Teplitz-Sembitzky (i.e. Martzoukos and Teplitz-Sembitzky, 1992)—cited earlier for his work on deterministic electricity pricing and investment (e.g., §4.3).

9.4.2 Negative Economic Asset Values

As described above (§9.2.2), asset values in the TGTP model may fall below zero. Where capacity expansion is optimal, the economic asset value of the original asset, costing $K(q)$, always drops below zero at $T_{CE \rightarrow CR}$ years. Such a result only occurs under capacity expansion optimality, and does not occur where either capacity replacement or anticipatory construction is optimal. Nevertheless, one might argue that such a result is problematic, potentially providing incentives for firms (under capacity expansion) to withdraw service from consumers connected to an asset that has a negative value, or deterring firms from re-investing in such assets when they need replacing.

Under capacity expansion optimality, the possible incentives for the regulated firm to withdraw service at the time which the asset value becomes negative are different from those at the time when the asset needs to be replaced. The value of this original asset becomes negative because, in order to meet demand in perpetuity, the incumbent firm must replace the original asset at a replacement cost of $K(q)$ at the end of the N th year. The economic asset value takes explicit account of this implicit obligation to maintain supply in perpetuity. Nevertheless, the net present value of payments to capital between the time when the asset value falls below zero and the date of asset replacement (i.e., \hat{V}_i , where $i > T_{CE \rightarrow CR}$) is always positive, as shown in (9.48) below. Similarly, payments to capital remain positive over the entire lifetime of the asset. Although payments to capital comprise a negative return on assets component, this is more than offset by the larger positive depreciation component. Consequently, the incumbent firm does not have an incentive to withdraw service at such time as the economic asset value becomes zero.

$$\hat{V}_i = \sum_{j=i}^N \bar{p}_2 q \left(\frac{1}{1+d} \right)^j = \sum_{j=i}^N \frac{K(q, q') - K(q')}{B_N} \left(\frac{1}{1+d} \right)^j = \left(\frac{\rho_{N-i}}{1 + \rho_{N-i}} \right) \frac{K(q, q') - K(q')}{dB_N} > 0 \quad (9.48)$$

Up until $T_{CE \rightarrow CR}$ years there remains a feasible asset transfer (or market) price at which a coalition of self-producing consumers (or an external firm) would be indifferent between: (a) purchasing the original capacity from the incumbent, and constructing the incremental capacity itself; and (b) constructing a single new asset. Nevertheless, after $T_{CE \rightarrow CR}$ years, the incumbent would be unable to sell the original asset to consumers (or to a hypothetical entrant). The monopolist would have to compensate any purchaser for taking the asset off its hands, since the original asset can only be accommodated as part of a construction configuration serving entire market demand by incurring costs higher than that associated with the currently optimal greenfields asset configuration. However, if a purchaser did not have an obligation to maintain supply past the end of the original asset's lifetime it would be prepared to purchase the asset without compensation, at a positive resale value of \hat{V}_i .

The negative asset value also arises because—looking backward, rather than forward—the original asset has become fully depreciated at $T_{CE \rightarrow CR}$ years. In other words, the incumbent firm has fully made a return *of* its initial capital outlay (as well as a return *on* that capital). However, the TGTP model indicates that an incumbent monopolist is economically justified in charging a positive price (i.e. payment to capital) for such an asset after $T_{CE \rightarrow CR}$ years, even though the asset would have—in an accounting sense—already been “written off” its books. Such pricing is not an abuse of monopoly power, but—as is discussed below—is in fact required to compensate the monopolist for the implicit obligation to maintain supply in perpetuity by replacing assets as necessary. Consumers are no worse off than they would be under any other possible set of prices, and the prices are thus subsidy-free in an intertemporal sense. This result is firmly grounded in the contestable markets paradigm, as the conditions for economic asset value and optimal depreciation—presented in (7.1)-(7.5)—are all satisfied.

In sum, a regulated firm can still have an incentive to maintain supply even after some of its assets have economic values that fall below zero. The negative asset values that arise under capacity expansion are consistent with: (i) the net present value of the future stream of payments to capital and reinvestment costs required to maintain the same level of service in perpetuity; (ii) the amount by which the incumbent firm would have to compensate a purchaser obliged to maintain the same level of service in perpetuity (assuming that all demand is served); and (iii) charging a positive price for a fully depreciated asset.

9.4.3 *The Implicit Obligation to Supply*

This leaves the objection that, at the time the original asset needs to be replaced (at N years), the present value of future payments to capital (over the lifetime of the replacement asset) would be less than the cost of replacing that asset. As such, the incumbent firm might have little incentive to undertake the necessary replacement. Under capacity expansion this situation occurs because, after N years, the cost of replacing the original asset is $K(q)$, whereas the present value of payments to capital over the next N years is $K(q, q')$ minus $K(q')$ —which is a lower amount. It might therefore seem reasonable to add constraints into the TGTP model that require the present value of future payments to capital (over the lifetime of any replacement asset) to be greater than the replacement cost of that asset. (Adding such constraints to the TGTP model would lead to an over-constrained model, as \bar{p}_2 would reduce to $K(q)/qB_N$, and \bar{p}'_2 to $K(q')/q'B_N$, which in combination, are prices that are inconsistent with the “Type 2 entrant” constraint).

However, at this point, it is worth re-emphasising that intertemporal subsidy-free prices in the TGTP model are from a consumer perspective. Were consumers able to supply the goods (i.e., network services) to themselves, then a negative asset value would pose no problem, as asset replacement could be funded from accumulated depreciation, as shown in (7.25). Ensuring that a third party provider (i.e. a

regulated firm) has the incentives to supply services at price levels which are subsidy-free is a separate issue, and—like marginal cost prices under perfect competition—does not affect the benchmark nature of the intertemporal subsidy-free prices. Adding such additional constraints into the TGTP model would not be consistent with intertemporal subsidy-free pricing.

Nevertheless, the issue of the regulated firm’s incentives to supply (or to replace assets) is an important one, and has some parallels with the concern that marginal cost pricing is impractical in practice, because revenue does not necessarily cover cost. However, under capacity expansion in the TGTP model, the revenue/cost equation ensures that revenues do cover costs overall, although—as BPW (p. 470) note—optimal behaviour may call for losses in some years. The incentive problems arise because of a mismatch in the timing of the stream of revenues and the stream of costs—all the “losses” occur grouped together after $T_{CE \rightarrow CR}$ years.

However, unlike a prospective purchaser, which after $T_{CE \rightarrow CR}$ years would need to be compensated for maintaining supply in perpetuity, the incumbent firm has *already* been compensated in advance, as a result of the price increase at τ_E years.⁶ The intertemporal nature of this compensation suggests that, without some external mechanism—such as the threat of entry, or the direct intervention of a regulator—it is unlikely that an incumbent monopolist engaging in optimal capacity expansion would set intertemporal subsidy-free prices of its own volition.⁷ While there is an implicit obligation to serve built into the TGTP model, this obligation is notionally counterbalanced by an implicit obligation upon consumers to accept that some capital recovery of the replacement asset must occur *prior* to reinvestment taking place. Because intertemporal coalitions of consumers are in general a hypothetical construct, in practice this obligation (or “regulatory contract”) would need to be borne on consumers’ behalf by a third party arbitrator, such as a regulator.

An implicit regulatory contract is not an unreasonable assumption for the TGTP model under capacity expansion. One of the very purposes of regulation is to address market failure, such as in

⁶ This is a somewhat similar situation to the obligations associated with decommissioning a nuclear power station. At the time of decommissioning, the power station owner has no incentives to incur decommission costs, as the costs will likely outweigh the benefits. Nevertheless, because the regulator has allowed these costs to be recovered in advance through comparatively higher prices, the owner has the obligation to proceed with the decommissioning.

⁷ Consequently, one might infer that New Zealand’s light-handed regulatory regime would be unlikely to provide such incentives. Value-maximising firms might be considered unlikely to voluntarily reduce prices to levels consistent with the second period of the TGTP model under capacity expansion. Under the information disclosure regime, ELBs have not been explicitly constrained from accelerating depreciation, or from raising prices to recover the costs of future replacement. However, as many ELBs appear to have consistently disclosed relatively low rates of return under the *Electricity (Information Disclosure) Regulations* (§9.2.2), this might suggest that factors other than the regulatory regime—for instance, the governance arrangements of the businesses—are constraining some ELBs from operating as value or profit maximisers.

instances where the market mechanism will not result in an otherwise optimal investment program because firms would not be sufficiently compensated over time for undertaking such a program.

In the TGTP model, the regulator—acting on behalf of consumers—may permit prices to increase (through accelerated depreciation) in advance of a future investment. The later reduction in value is entirely offset by the earlier increase in prices, and this increase would not have been permitted by the regulator unless the regulated firm had agreed to set lower prices in the future. As discussed above (§5.3.2), Sidak and Spulber (1998) warn about the damage that can be caused to a firm if regulators ignore the investment-backed expectations of that firm. Conversely, consumers can suffer if regulators ignore the payment-backed obligations of a firm.⁸

Under capacity expansion, the TGTP model effectively provides a fund of accumulated depreciation which can be used to exactly compensate the firm for any shortfall in the present value of payments to capital associated with replacing an asset. Accumulating and maintaining this fund of accumulated depreciation from τ_E years onward would not inhibit the regulated monopolist from providing appropriate levels of dividends to its owners (whether its owners be consumers or external shareholders). This is because dividends are considered to be paid from return *on* capital, and not return of capital (Hay and Morris, 1993, p. 424).⁹

However, if the regulator considers that the incumbent firm might prefer to dip into this “fund” prior to replacement occurring (perhaps to provide comparatively larger dividends to its shareholders), the regulator could—as but one possible mechanism—require the incumbent to maintain prices at the same levels after τ_E years, and itself directly levy consumers for the difference (i.e., $[K(q,q') - K(q)]/qB_N$). This levy could be placed into an accumulated depreciation fund and used by the regulator to compensate the incumbent at a later date.¹⁰

⁸ Firms often seek advance payment for providing services. While at the time of delivery the net present value of providing those services will be negative, this does not negate the obligation on the firm to undertake delivery.

⁹ Hay and Morris (1993, pp. 378-379 and 424-426) explain that, with respect to a firm’s sources and uses of funds, retained earnings come from the return on capital, and not the return of capital. Together, depreciation and retained earnings provide internal cash flow that, in combination with each other and with external funds, can be used for investment expenditure and increases in net current assets. In the TGTP model, accumulated depreciation rather than retained earnings is assumed to be available for funding reinvestment in the network.

¹⁰ The regulator might take the view that, how the regulated firm uses this fund is at the firm’s discretion, but the possible unavailability of that fund at the time it is needed to be invested in replacement assets does not obviate the firm’s obligation to supply. The firm has been compensated (in advance) for that obligation.

While there are likely to be practical implementation issues associated with various mechanisms for ensuring that the regulated monopolist is adequately compensated for its obligation to supply, the key point is that these issues are separable from the derivation of the intertemporal subsidy-free prices themselves. Even where investment incentives are a valid issue, the benchmark prices in the model do not necessarily change. Nevertheless, in implementing policies to move prices toward those benchmarks, policymakers will need to ensure that the overall governance environment (of which regulation is but one aspect) retains adequate incentives on regulated firms to replace existing assets and to undertake ongoing new investment in an optimal manner.

9.4.4 Modelling Only Two Services

A further consideration is that regulators are generally more concerned with overall price levels than with individual tariffs (§5.1.2), and are usually content to leave the regulated firm to discover Ramsey-efficient tariff structures for itself, subject to an overall revenue or (average) price cap (and perhaps an obligation to supply). Consequently, as Sidak and Spulber (1998, p. 314) explain, “the firm’s rates should be established so that, on average, it earns zero economic profits on its regulated services as a whole” across the full aggregation of regulated services that the incumbent is required to offer. Sidak and Spulber emphasise that “*firms* earn profits; individual products or services do not”, and point out that service obligations will result in the incumbent earning a negative contribution to its overall profitability from some services. As such, the incumbent will also have to earn returns on other services, that viewed in isolation would appear to yield a positive economic profit.

Positive and negative economic profits will therefore occur for individual services both spatially and temporally. Providing various combinations of services throughout any power distribution network, in order to meet increased demand, is likely to require a mix of optimal construction configurations. At various locations throughout any network, different circumstances may require incrementally expanding capacity, building spare capacity in anticipation of future demand, and replacing existing assets before the end of their useful lifetime. For an individual service at a particular point in time, both accelerated depreciation and returns on spare capacity might appear to result in positive economic profits. Conversely, services in other parts of the network involving incremental investments might appear to be earning sub-normal returns. However, this is on an annual basis, and not over the lifetime of the investment program (as opposed to the lifetime of the assets concerned).

A limitation of the TGTP model is that only two services are considered, whereas a real-world ELB has a multiplicity of services. The problem of sub-normal or positive returns for individual services is much starker for a firm that supplies only two services. Similarly, the notion of an accumulated depreciation fund (held by either the incumbent itself or the regulator) may appear somewhat unrealistic for a two-good firm, but the concept is more reasonable for a firm providing multiple services. The more services there are in a network—over time and across different locations—the more likely it is that

positive and negative economic profits will offset each other. Accumulated depreciation and returns on spare capacity in one part of the network become a source of funds to offset apparent losses in another.

9.4.5 *Uncertainty and Risk*

Including only two services in the TGTP model also has implications for the model's ability to be extended to address the issue of uncertainty and risk. Systematic risks can be offset in the model through an increase in the WACC—i.e. the discount rate, d (e.g., Dixit and Pindyck 1994, p. 346). As Hay and Morris (1993, p. 219) explain, part of the WACC is a “reward” for risk taking, and can be treated as a legitimate cost of doing business. *Ceteris paribus*, a higher d would: (i) decrease total costs, as can be seen from (8.41)-(8.43); (ii) make capacity expansion relatively more attractive than the other configurations—or, if capacity expansion is infeasible, make capacity replacement relatively more likely to be optimal than anticipatory construction—as can be seen from Figure 6.2; and (iii) under capacity expansion or replacement, cause the period 1 price increase to occur later, as can be seen from (9.16b) and (9.27b). However, one of the most significant risks facing electricity lines businesses—or any network industry involved in investments in irreversible assets—stems from uncertain demand, which is an unsystematic risk.

The TGTP model assumes perfect information regarding future demand. Were it known in advance that demand is not perpetual, but would cease after i years (where $T < i < N$), the regulated firm would need to recover a greater depreciation component in its payment to capital in order for the zero economic profit constraint to be satisfied. This is consistent with intertemporal subsidy-free pricing, because a coalition of consumers would also be faced with recovering the costs of the asset(s) over a shorter period, which implies a higher price.

In reality, the regulated firm may have a difficult time convincing a regulator of the need to accelerate depreciation, increase prices and/or invest in additional spare capacity now, to offset a potential shortfall or excess in demand at a later date. Uncertainty in the magnitude and timing of demand growth directly affects the distribution network investment decision (e.g., Lesser and Feinstein 1999). While intertemporal subsidy-free prices could be based on what currently would be the least-cost construction program, based on *stochastic* data, the uncertainty may result in the incumbent monopolist not necessarily making exactly a normal economic profit, and sub-optimal under or over investment may arise.

With increasing returns to scale, uncertainty in demand can require optimal “excess” capacity to be incorporated in every asset (e.g., Boiteux 1949; Turvey 1969; Kleindorfer and Fernando 1993; §6.1.1)—even under capacity expansion—increasing total costs and requiring depreciation to be further accelerated in order to provide a normal profit. On the other hand, the long lifetimes and lumpy nature of distribution network investments mean that there is always some risk of asset stranding (§3.5.1 and

§6.3.3), because the timing and level of demand for network capacity cannot be forecasted with complete accuracy.

The risk of asset stranding—particularly that associated with larger consumers—may be able to be mitigated to some extent via the terms of connection contracts. However, this leaves the risk associated with the rest of the consumer base to be managed. An estimate of how significant asset stranding is in New Zealand’s electricity distribution industry can be made from the difference between ELB depreciated replacement cost and ODV values, as the difference reflects the reduction in valuation due to optimisation (§7.4.2) and to the economic value (EV) “write down” (§7.4.3). In combination, these two reductions might be expected to reflect the impact of changes in demand and technology, as well as the threat of bypass supply. For the 2001 disclosure year, on average this difference was about 2.1%, about 0.25% of which was due to the EV adjustment (data from Parsons Brinckerhoff Associates, 2002, Appendix 2). Interestingly, the variation in the difference appears to decrease markedly with ELB size.¹¹ However, these adjustments to an ELB’s asset base could also partially reflect inefficiencies or “gold-plating” prior to the advent of reform, so 2.1% might be an upper bound to the level of possible asset stranding for an efficient-sized ELB facing incentives for efficient operation and investment.

Demand risk—like the impact of positive or negative annual returns—becomes much more significant the less is the number of services provided by the firm (and this is to some extent related to firm size). As such, a two-good model—such as the TGTP model—dramatically exacerbates the magnitude of any loss in value caused by asset stranding, and is therefore of limited use in quantifying the cost of risk due to demand uncertainty. In larger networks with comparatively high consumer density (§3.4.2), diversifying the risk of uncertain demand is more practicable. Greater consumer density implies more interconnections between feeders supplied from different zone substations. Also, larger businesses are more likely to have the financial resources to be able to invest in asset management systems that allow rapid or even real time operation of switches between feeders, allowing demand to more closely match capacity. Hence, for electricity demand served by any particular zone substation, short to medium term shortfalls or surpluses in capacity can be managed by allocating demand across neighbouring distribution feeders and zone substations.

In early 2001, there were 29 ELBs in New Zealand, with the largest four businesses in combination serving 62% of the total number of consumer connections. At the other end of the spectrum, the smallest ELB served 4,258 consumer connections—only 0.25% of the total. The largest ELB had more than 120 zone substations, whereas a number of the smaller businesses had as few as three

¹¹ For instance: for the 14 ELBs with ODVs up to NZ\$70m, this difference ranges from 0% to 7.5%; for the 9 ELBs with ODVs from NZ\$70m-\$200m, the difference ranges from 0.3% to 4.1%; for the 3 ELBs with ODVs from NZ\$200-\$500m, the range is 0.7%-2.9%; and the two ELBs with ODVs above NZ\$500m both exhibit a difference of 1.9%.

or four. (One ELB had no zone substations at all, being connected to Transpower’s transmission grid at 11kV). Consumer density ranged by an order of magnitude, from 3.4 connections per km of line to 35.3 connections per km of line in the densest network (PricewaterhouseCoopers, 2002, pp. 6-8).

This raises an interesting regulatory policy question—whether the higher demand risk (and possibly other risks) inherent in smaller less dense networks should be explicitly taken into account when considering the issue of efficiency. The question can be posed thus: should smaller ELBs with relatively low consumer density be compensated—by being permitted to accelerate depreciation relatively more than other ELBs (or perhaps to be permitted a higher WACC)—to account for their potentially higher risk of asset stranding? Wyatt *et al.* (1989) concluded that having more than 11 ELBs in the distribution industry does not appear to be a cost-minimising industry structure (§3.4.3). Consequently, Wyatt *et al.*’s result at least suggests that explicitly taking into account the higher risks borne by ELBs as a result of scale—although not necessarily low customer density—might be counterproductive, as this could dampen incentives for further industry consolidation and improved productive and dynamic efficiencies.

Clearly, a full assessment of the characteristics of efficient and “fair” line charges in a New Zealand context will require extending the work presented in this thesis, by integrating the impact of risk and uncertainty into the model of intertemporal subsidy-free prices. However, this thesis has attempted to show how intertemporal effects can begin to be explicitly taken into account when determining efficient, “fair” and subsidy-free benchmark prices.

CHAPTER X

CONCLUSIONS: TOWARD EFFICIENT AND “FAIR” LINE CHARGES

Applying the concepts of least-cost suppliers, contestability, and sustainability for multiproduct firms is problematic. The concepts represent idealised cases: US economists, Sanford Berg and John Tschirhart (1995)

An improved expectation of jam to-morrow can be said to raise the cost of bread today: UK advocate of marginal cost pricing, Ralph Turvey (1969)

In the language of economists, deliberate overequipment is an “arbitrage activity” and not a “production activity”. It has its own income: French pioneer of marginal cost pricing, Marcel Boiteux (1949)

Berg and Tschirhart (1995) suggest that several issues remain at the forefront of current policy debates over the appropriate regulation of utilities: (i) how to determine which industry structures (and government policies) best promote new services and production processes; (ii) how to determine when a natural monopoly actually exists given changing demands and technologies; (iii) how to select price configurations that will both recover costs and encourage efficient consumption choices; and (iv) how to ensure that those prices are sustainable—that is, not susceptible to the threat of entry from competitors, or to self-production by consumers. While the focus of this thesis has clearly been on the *third* item—in the context of power distribution in New Zealand—Berg and Tschirhart’s list of policy issues neatly captures the overall scope of the inquiry presented in the previous nine Chapters. But, as their quote which opens this Chapter suggests, the topic is not a trivial one, and the model presented in this thesis is without a doubt highly idealised, and could be extended substantially in both breadth and depth.

This Chapter synthesises the examination of the relevant cost concepts, and pricing and depreciation principles, undertaken throughout this dissertation, in the context of the overall question: *what are the characteristics of efficient and “fair” prices (i.e., line charges) for distribution services?* In particular, given that the achievement of efficient and “fair” prices was a desired outcome of New Zealand’s power sector reforms, this Chapter closes by briefly assessing whether the line charges currently set by electricity line businesses (ELBs) in New Zealand have these characteristics. If such is not the case, this challenges whether the overall reform objective of *economic efficiency*—in both its *static* and *dynamic* forms—is being attained. Throughout this Chapter, some shortcomings of the analysis, and possible directions for related research inquiry in the future, are identified.

10.1 The Question of Natural Monopoly

New Zealand’s legal separation of the business of electricity retailing from the business of electricity distribution (§2.4.5)—namely the conveyance of electrical energy through a physical power

distribution network, and the connection of consumers to that network—would seem to provide a golden opportunity for assessing the efficiencies of distribution network services in isolation. This thesis has taken the position that the unbundled provision of power distribution network services is a *multiproduct* activity in both space and time, and that the product which consumers purchase is an “option” on *connection capacity* at a particular location, measured in kVA (§3.1.1-§3.1.2). This option lasts for the entire time period which those consumers choose to be connected to the network, and not just at the times when they draw electrical energy from the network.¹

Internationally, and no less in New Zealand (§3.2.2), it has often been asserted that electricity distribution is a “naturally monopolistic” activity, in the absence of a clear definition of what is meant by natural monopoly. Besides, some network competition has been observed in New Zealand (§3.3.2), and was not unknown in the United States until proscribed by regulation (§3.2.1). While the standard definition of a natural monopoly appears straightforward—namely, that a single firm can produce total market demand at a lower cost than any combination of two or more firms (§2.1.5)—two problematic issues exist: first, how to describe the dimensions of the market in question, in both space and time; and second, how to assess whether a single firm would in fact incur the lowest capital and operational costs in serving that market. With respect to the dimensions of the market, it is clear that, in practice, a large number of firms do contribute to the provision of distribution network services in New Zealand, both horizontally and vertically (§3.4.1). Moreover, many of the functions associated with the provision of distribution services—particularly construction, maintenance, and billing—are no longer provided by the owners of the distribution network assets themselves, but are contracted out to third party firms, and appear to have contestable or even competitive characteristics (§3.2.3).

The standard definition of natural monopoly, which relates to the number of *firms*, fails to acknowledge that the power distribution sector involves a complex set of interrelated functions, has cost complementarities with other elements of the electricity supply industry (§3.1.3), and may share economies with similar functions in other network industries (§3.3.4). Mergers relating to network ownership within the power distribution sector, or between different utility industries, as well as divestment of potentially-competitive or contestable network functions, have all been undertaken in order to improve efficiencies (or increase profits), mindful of trade-offs with increased transaction costs. Yet simply because a market temporarily arranges itself in a particular manner does not necessarily imply that

¹ However, in part due to the losses associated with electrical energy conveyance, there are strong cost complementarities which exist between network and retailing services. Hence, while the separation of line and energy functions allows network services to be analysed in isolation, it does not allow any possible inefficiencies introduced by that separation to be estimated. The impact on the optimality of distribution investment and operation due to unbundling is a subject which is not examined in this thesis, but which merits attention. In addition, the issue of the impact of diversity on distribution costs has also been set aside, and future work needs to extend the analyses presented in this dissertation accordingly.

that arrangement is optimal; many of the changes in organisational structure and much of the competitive market behaviour which occurred upon the initial deregulation of New Zealand's power distribution sector could have simply been a process of "market discovery" (§3.3.2). Moreover, increased costs are more likely to be substantial from the duplication of assets rather than the proliferation of firms. While a single distribution network might be required in order to minimise capital expenditure, in theory New Zealand's open access regime permits bypass or subnetwork competition for network services, not just competition for energy sales (§3.3.1). In practice, this will depend on the real-world barriers to entry (§3.4.5), and more analytical work is required to identify these in a New Zealand context. Hence, competition does not necessarily imply the "wasteful" duplication of assets. And, even though "asset duplication" is itself sometimes referred to in a pejorative sense, both redundancy (§3.1.2) and spare capacity (§6.1.1) can be intrinsic aspects of an optimal distribution network design over time (§3.1.3).

Given these complications, this thesis concludes that asking whether electricity distributors are natural monopolies is not really a pertinent question in the first place. The concept of natural monopoly is simply a proxy for assessing whether the services in the relevant market are provided at least cost. Hence, this somewhat shifts the focus of the second issue—the appropriate cost tests for establishing the existence of natural monopoly. Clearly, many possible kinds of network economies are associated with distribution network investment and operation (§3.4.2). In particular, strong economies of scope can be realised due to the economies of scale inherent in the costs of joint capacity at zone substations (§3.6). Yet, even though making the provision of electricity services "contestable wherever possible" is a currently held objective of the power sector reform process (§2.4.7), there appears to be little recognition in New Zealand of the appropriate test for natural monopoly which has arisen from contestability theory (§3.4.3), namely the subadditivity test (§3.4.4). However, rather than using the subadditivity test to determine whether the distribution sector's cost structure is a natural monopoly, such a test is more appropriately undertaken to answer the broader question: what is the least cost sector organisation, number of firms, ownership structure, and nature of commercial relations? In other words: is the distribution sector economically efficient, given its interrelationships with other markets? Unfortunately, the subadditivity test is difficult to apply in practice. Consequently, more research is required into this question, as well as into the level of efficiency and fairness improvements realised by the new ownership arrangements (particularly any relative efficiency and fairness benefits or costs of consumer ownership; §6.3.1), industry structure, and commercial relationships.

At any rate, the presence of natural monopoly—however it is established—does not in itself indicate the presence of an insufficient return on investment on the one hand, or the exercise of monopoly power on the other. Hence, the question is not a necessary aspect of the search for a benchmark of efficient and fair prices, and in itself does not establish whether or not the overall goal of economic efficiency is being achieved. Moreover, the exercise of "monopoly power" does not necessarily require the firm in question to be a monopoly; therefore, a better term is perhaps the concept of "market power".

10.2 Efficient Pricing and the Time Dimension

The debates over efficient pricing of utility services are wide-ranging, and many of the roots of the debate stem from within the electricity supply industry itself (§4.1-§4.3). The two key factors which stand out as complicating, and in fact shaping, the nature of these debates are the impact of *time*, and of *uncertainty*. This thesis has focused almost exclusively on the first of these two factors, and clearly the derivation of subsidy-free prices presented in this dissertation needs to be extended to handle the effects of uncertainty and risk (§9.4.5).

Part of the problem seems to have been that traditional *short run* marginal cost (SRMC) or *long run* marginal cost (LRMC) pricing (§4.1.3-§4.1.4) of bundled electricity supply did not treat electrical energy and network connection capacity as distinct products. Yet even for such bundled products, two-part tariffs—where different price components reflect the distinct costs of capacity and of short run expenditures—have been shown to Pareto dominate comparable linear pricing approaches under the competitive market model (§4.2.4). Unbundling electricity retailing and electricity distribution into two distinct types of products, and pricing them separately, could effectively result in such a two-part tariff.² The costs of electrical energy can be priced on a short run basis, while the costs of network capacity can be recovered acknowledging that investments in durable assets are innately long run in nature.

While short run marginal cost pricing can effectively signal the *opportunity costs* (§4.1.5) of the moment-to-moment costs of generating power, and usually allow generation capacity costs to be recovered (§4.2.5), SRMC-based pricing applied to electricity distribution will fail to reflect the opportunity costs of network capacity (§5.3.2). Unless electricity distribution is accepted to be economically efficient in the guise of a loss-making public service, revenue reconciliation and incentives for investment cannot be ignored (§4.2.1). Pricing network capacity on SRMC alone would thus be “extraordinarily naïve” (§4.2.2), and fallaciously ignores the “investment-backed expectations” of electricity distributors (§5.3.2). Yet SRMC-pricing is still reaffirmed by some as the efficient approach to pricing electricity distribution (§4.1.1); possibly this a reflection of the pervasive influence of the conventional peak load pricing model (§4.3.4)—in the presence of constant returns to scale—which upholds that off-peak users should not contribute to the costs of capacity (§4.1.2).

Within the framework of contestability theory, Baumol and Sidak (1994a) managed to step aside from this protracted short run versus long run marginal cost pricing debate by concluding that the problem with the competitive market model is that it calls for prices to be *equated* to some measure of cost (§5.1.1), given its underlying assumption of constant returns to scale (§2.1.7). Nevertheless, the

² Nevertheless, in New Zealand’s case, electricity retailers are allowed to rebundle these distinct price components together, and are expected to offer at least one pricing option involving a tariff with a low fixed charge component (§4.4.2).

stand alone cost (SAC) and *incremental cost* (IC) bounds on subsidy-free prices in their *constrained market pricing* approach are clearly long run in nature (§6.1.2). For instance, the overall total revenue/cost equality—the zero economic profit constraint—is determined on a present value basis (§6.1.3). As such, the approach is inherently *forward-looking* rather than *backward-looking* (§6.3.1). Another important characteristic of the approach is that it explicitly requires the subsidy-free bounds on prices to incorporate the full opportunity costs of supply (§6.3.1).

A key concern with the application of constrained market pricing as a regulatory tool for price control is the potentially large gap that arises between incremental costs and stand alone costs in practice, allowing firms great pricing freedom (§5.1.3). Yet if the bounds produced by the approach were economically correct, such an outcome would be more of a political issue. However, this thesis has demonstrated that the concern relates not only to the way constrained market pricing has been implemented in practice, but also to the way the method has been formulated. While the bounds typically provide necessary economic constraints on subsidy-free prices, they rarely provide sufficient constraints, and this is the reason for the broad range between the bounds.

Implementation of constrained market pricing has differed widely in practice, depending on how stand alone costs, and particularly incremental costs, have been calculated (§5.3.3). A difficulty is that both types of costs are hypothetical (§4.3.6). Faulhaber’s (1975) original approach required that IC be derived by subtracting the complementary SAC value from the zero economic profit constraint (§4.3.2), rather than determined directly, and that a combinatorial test be applied for groups of products, not just for single products (§5.2.3-§5.2.4). In practice, both of these requirements are sometimes ignored. Where incremental costs are determined solely for individual products in their own right, costs physically *common* (or *joint*) to more than one product may be left out of incremental costs, meaning that the lower bound cost surface lies below where it would otherwise would be. But even if incremental cost constraints are evaluated on a combinatorial basis, a typical conclusion is that physically common costs do not contribute to incremental costs in any case (§5.2.1), a position shown to be fallacious where there are increasing returns to scale, even in the absence of intertemporal effects (§5.2.2). Moreover, the widespread idea that where *spare capacity* exists, incremental costs are low, or even zero (§5.1.3), is a similar fallacy to the short run notion that efficient pricing does not contribute to the “historic” costs of “sunk” capacity (§4.1.4). If stand alone costs are derived from erroneously low IC values, then SAC bounds will be correspondingly excessive. Where bounds determined in this manner are used for regulatory purposes, exercise of market power will most likely remain unchecked.

Apart from specific problems arising from incorrect implementation of the method, the standard formulation of the constrained market pricing approach itself has shortcomings, in regard to the way the *time dimension* is treated. While Baumol and Sidak correctly state that fixed costs do not contribute to incremental costs where time is not a parameter (§5.2.4), they suggest that the same result holds even

where the initial demand for two products is sequenced (§5.3.1). The underlying supposition is that, where fixed costs have already been incurred—or when spare capacity exists—meeting demand growth incurs no additional costs at the time of increased demand. While this may well be true, a fallacy arises by viewing additional costs and incremental costs as equivalent (§5.3.2). Perhaps “incremental cost” is a poor term to use to describe what actually is the cost of producing a particular set of products in the *absence* of the other products (§5.3.3). Even where past fixed or capital outlays are “historic”, they can still tie up resources today and into the future.

Furthermore, constrained market pricing is predicated on the principle that demand-side effects can be ignored, and that firms are free to take consumer demand characteristics into account themselves when setting prices (§4.3.6).³ As such firms can efficiently price discriminate, perhaps using the Ramsey pricing approach (§4.2.2). But if investments involve *non-fungible* capacity exhibiting increasing returns to scale, then investment decisions become *intertemporally interdependent* (§3.5.1), and are likely to be sub-optimal if the demand of future consumers is ignored. If an optimal investment program requires tying up resources in *anticipation* of future demand (§6.1.1), then current consumers should make some contribution to the opportunity costs incurred, because their current demand is itself a factor which dictates the optimal construction configuration (§6.2). Just as there is an opportunity cost involved in meeting consumption today, there is an *option value* associated with postponing that demand until tomorrow (§5.3.4). Neglecting this intertemporal interdependence of costs associated with serving both current and future demand means that, while being a long run approach, constrained market pricing—as articulated by Baumol and Sidak—remains a *static* method (§6.1.2). *Dynamic* considerations are introduced by implementing the approach in tandem with the imposition of incentive regulation (§6.4.1), rather than through any element of constrained market pricing itself. Yet, in doing so, the focus is more on encouraging innovation than ensuring optimal investment.

In constrained market pricing, the benchmark for SAC and IC bounds are the costs from the point of view of a hypothetical entrant. Consequently, the entrant’s perspective on how all or part of the current and future market can be optimally served on a *greenfields* basis is crucial (§6.1.4). Significantly, the optimal greenfields asset configuration can differ from the asset configuration which results from the successive application of the optimal investment rule (§4.1.4) by the incumbent firm (§6.2). Hence, while constrained market pricing is clearly forward-looking in outlook, SAC and IC could still be

³ Demand is treated very simply in the models presented in this thesis. The implicit assumption is that consumer demand for network connection capacity is inelastic, and statically and intertemporally independent (although the costs involved in meeting those demands can be intertemporally interdependent). Demand is not infinitely inelastic however, because it is capped by the rational willingness-to-pay bounds (§4.3.6 and §5.1.1) of net intertemporal SAC and net intertemporal IC (§7.3.1). But even comparing just the fairly crude notions of finite and perpetual demand indicates what a profound effect demand can have on intertemporal stand alone cost and intertemporal incremental cost (§8.3.3).

indexed to either, neither, or both of *replacement costs* and *historic costs*, depending on whether the costs incurred by a hypothetical entrant would be the same as, or different from, those which have been and/or will be incurred by the incumbent firm (§6.3.3). In some cases, changes in demand or technology (§6.3.2) may be so marked that the optimal greenfields asset configuration might include no assets which are part of an incumbent's existing asset base, or which replace those assets at some future date. Consequently, the question whether prices (or asset values) depend on historic or replacement costs is inappropriate, if it takes the perspective of the incumbent firm rather than that of an entrant. As Boiteux suggested, prices should be based on what *would* be the network design, were the current asset configuration optimally matched to current and future demand (§6.3.5). (While acknowledging that the optimal design might include spare capacity in anticipation of future demand, Boiteux appears to have neglected asset indivisibilities). In the presence of uncertainty such a pricing approach does not “guarantee” the incumbent firm a return on and return of its investment costs—however, a firm in a competitive market is not guaranteed a return either (§6.3.4).

10.3 Intertemporally “Fair” Pricing and Economic Depreciation

“Fairness” in this thesis has been narrowly defined within the framework of economic efficiency itself (§4.3.1). This requires asking: do prices reflect the full opportunity costs of the resources being utilised? To simplify matters, it has been assumed that private and social opportunity costs are the same, thus allowing this question to be rephrased as: does a firm receive adequate compensation in the form of a return *on* capital employed, and a return *of* capital employed? If firms receive no less than this level in absolute terms, then their interests are protected, and if they receive no more than this level, then consumer interests—as a whole—are also protected. Consequently, subsidy-free pricing requires that firms receive a zero economic profit (§4.3.2), and that the absolute price levels faced by consumers—in an aggregate sense—are “fair”.

In a perfectly contestable (or competitive) market, “fair” prices are seen as those emerging from the workings of the market mechanism itself. But when firms are price setters rather than price takers—and can exercise some monopoly or market power—then policymakers have often tended to address the issue of fair prices by benchmarking allowable revenues to a return on and return of asset value (§7.1.1).⁴ Regulated prices—which are the same as *payments to capital* where operating (or variable) costs are ignored—thus comprise a return component, which recovers the opportunity cost of capital, and a *depreciation* component, which recovers the capital outlay itself (§7.2.1) and—through accumulation—can maintain the firm's overall value, even as individual asset values depreciate (§7.3.2). In a contestable market, economic asset value is usually derived from the discounted stream of future net revenues from

⁴ This thesis has not addressed the debates surrounding the appropriate calculation of the opportunity cost of capital (i.e., the WACC).

the products associated with the asset in question (§7.1.3), and economic depreciation is simply the change in economic asset value over the period of interest. But, in a regulated market, prices are themselves generally derived at least in part from underlying asset value, hence an unfortunate element of “circularity” creeps into the concepts of valuation and depreciation, even where incentive regulation is imposed in place of traditional rate-of-return regulation (§7.1.2). Such a problem has been highlighted with respect to New Zealand’s approach to electricity distributor valuation and line charges (§7.5.2).

Usually regulators specify the allowable depreciation (or valuation) method, and this determines how asset value—and thus prices—change over time. Interestingly, it can be shown that there are an infinite number of different streams of depreciation provisions which can provide a firm with a zero economic profit, and this even holds for such common methods as straight line depreciation (§7.2.2). This might suggest that the choice of the depreciation or valuation method is unimportant to the issue of fair prices, and that firms could be permitted to “front-load” or “back-load” their depreciation schedules as they see fit (§7.3.3). On the other hand, Baumol (1971) suggested that at any time in which there is “unused” capacity, then prices should include *no* contribution to capacity costs (§7.2.1)—in other words, the depreciation payment should be zero—a conclusion which appears to harken back to the outcome of the conventional peak load pricing problem, and which has also crept into the policy prescriptions derived from contestability theory (§7.2.2).

However, subsidy-free pricing not only requires that absolute price levels be fair, but that relative price levels are as well (§5.1.2). Even within a static framework, the concept which best protects consumer interests is that of *anonymous equity* (§4.3.3-§4.3.4), which recognises that consumer interests are better protected with reference to a benchmark of their own costs of *self-production*—derived from the formation of game-theoretic consumer coalitions (§4.3.5)—rather than the costs of a hypothetical entrant. (Since subsidy-free prices are anonymously equitable in the presence of declining average incremental costs, the two concepts have been used somewhat interchangeably throughout this thesis). However, any firm may serve multiple consumers not just over space, but over *time* as well, and given the apparent misapplication of the time dimension in constrained market pricing, the models derived in this thesis have primarily focused on the somewhat neglected question of the relative efficiency and fairness of price levels in a temporal—rather than a locational—sense. Although this might appear similar to examining absolute (but static) price levels over time, it differs, because any ramifications arising from the intertemporal interdependence of costs have been explicitly taken into account, allowing *intertemporally anonymously equitable* prices to be derived (§6.4.2). A firm’s “fair” annual receipts are thus not only guided by what provides it with a zero economic profit in present value terms, but by what provides that fair return and an anonymously equitable price path in aggregate terms.

While there are many price paths which can ensure that a firm makes a fair return on its investment, consumers that enter and leave the market over an asset’s lifetime will not be indifferent to

significant fluctuations in prices, even if such prices still provide the supplier with a zero economic profit (§6.4.3).⁵ As Baumol (1971) once observed, the pattern of future prices should minimise the distortion of consumer choice over time. Yet, although Baumol and Sidak claim that the SAC-based price ceilings under constrained market pricing “protect consumers”, if subsidy-free prices are based on costs of a hypothetical entrant—one that is effectively indifferent to the time path of receipts—then consumers’ interests are only protected if the timeframe of their demand happens to coincide with the lifetime of the assets comprising the investment. Hence, protecting consumer interests—as a whole over time—is not sufficient to ensure intertemporal anonymous equity. A more appropriate benchmark for determining the bounds on subsidy-free prices would seem to be the costs of hypothetical self-production coalitions formed intertemporally of current and future consumers (§6.4.2). And within the perfect contestability paradigm, self-producing consumers are assumed to have access to perfect information regarding their own costs of self-production and the nature of future demand (§5.1.1). While there are an infinite number of streams of depreciation payments which will ensure that an incumbent recovers its initial capital outlay, it seems likely that there will only be a single depreciation path associated with each of the *intertemporal stand alone costs* of supply, or the *intertemporal incremental costs* of supply.

For the case of no demand growth, a feasible asset resale price can exist between the self-producing consumers that initially invested in an asset, and a later group of self-producing consumers, even if that asset is deemed non-fungible (§7.3.1). An asset which has no alternative *use*, may have some value to (subsequent) alternative *users*, in its existing use (§7.3.4). The effect that this has is to reduce the intertemporal stand alone cost of self-production from the initial consumers’ point of view. Had demand for the services provided by an asset ceased prior to the end of the asset’s lifetime, and no subsequent demand for the utilisation of that asset emerge, then intertemporal stand alone costs of self-production would have been higher. The upper bound on intertemporal subsidy-free prices would therefore also be higher, since prices would need to recover total costs over a timeframe which is shorter than asset lifetime. But where resale between successive consumer groups is feasible, this provides a more restrictive *net* intertemporal stand alone cost bound to the subsidy-free prices. By complementarity (§4.3.2), this allows a higher and more restrictive net intertemporal incremental cost bound to be derived. Interestingly, where non-transferable (i.e., non-resellable) fixed costs are negligible, the net intertemporal stand alone costs and net intertemporal incremental costs are *equal*. And if only a single subsidy-free price path exists, then the problem of valuation circularity is resolved. Economic asset value and the associated depreciation schedule can be directly derived from the intertemporal subsidy-free prices. For the case of constant demand, the economic depreciation payments are heavily back-loaded and can be shown to be equivalent to the long run incremental cost of capacity (§7.2.3). Notably, this result has

⁵ Given that the lifetimes of power distribution assets are up to 60 years, such coming and going of consumers at a particular location in a distribution network seems highly likely (§3.1.2).

important implications for the application of constrained market pricing in practice, since it implies that the concerns regarding great pricing freedom (§5.1.3) may be rendered null and void, as long as intertemporal factors are appropriately taken into account.

Again, like its relevance for pricing, the question of whether historic costs or replacement costs should be included in asset valuation, is revealed to be imprecise. The answer, as Turvey suggested, seems to be not necessarily either (§7.3.3). Since asset value can be derived from the subsidy-free prices, historic costs or replacement costs will be reflected in value only to the extent that such costs are reflected in the hypothetical optimal asset configuration of self-producing current and future consumers.

10.4 The Characteristics of Efficient, “Fair” (and Sustainable) Line Charges

As Berg and Tschirhart (1995) indicate, regulatory restrictions are typically considered: “for a single-product firm—intertemporally, or for a multiproduct firm at a given point in time”. This thesis has attempted to redress this shortcoming by extending the single-good/two-period model—used by Baumol and colleagues (i.e., BPW) to analyse intertemporal unsustainability (§8.1)—to a two-good/two-period (TGTP) model. Although BPW did not use their single-good model to evaluate cross subsidies, limiting their extensive discussion of cross subsidies to the “timeless world”, it is shown that the extended BPW model has some interesting insights to provide on the range between intertemporal stand alone costs and intertemporal incremental costs in the presence of demand growth, as well as some insights relating to the issue of intertemporal unsustainability itself.

By extending BPW’s model from one good to two (sequenced) goods (§8.2)—both consumed in perpetuity (§8.3)—it was demonstrated that modelling a *single* product market is itself the main factor which contributes to the intertemporal unsustainability inherent in their model. The problem of unsustainability mainly arises because the incumbent firm is unable to distinguish between the initial consumers and the subsequent consumers, and therefore they must face the same prices, even if their costs of supply differ (§8.4.1). In a network, however, consumers are distinguishable by their location in that network, and can be charged different prices. This cannot be considered price discrimination if the costs of providing network connection at those locations are different. Such costs will differ—even where the demand characteristics at two locations are identical (except for the fact that they are sequenced)—as long as the joint costs of non-fungible capacity are subject to intertemporal economies of scale, and where the optimal construction configuration furnishes an intertemporal interdependence of costs. This occurs because time has its own opportunity cost; a result evidenced in the apparent equivalence of many concepts, especially long run incremental cost, option value, and economic depreciation (§7.1.1). Furthermore, while capacity expansion is an inherently less sustainable construction configuration than the other two possibilities of anticipatory construction or capacity expansion, the “finite demand assumption” in BPW’s model makes capacity expansion—and thus unsustainability—more likely to be optimal than might be the case for real network industries (§8.4.2).

Unsustainability can also be caused by treating the coalition-building power of current and future consumers as being *symmetric*, which makes finding a “core solution” to the game theoretic formulation of the sustainability problem more difficult than it should be (§8.4.4). Explicit recognition of the time dimension gives earlier consumers a fundamental “first mover advantage” over later consumers, and gives them the deciding vote over which coalition serves their interests best. Nevertheless, this does not disadvantage subsequent consumers, since they are actually better off than they would otherwise be—namely, if the market were to remain unserved simply because a “symmetric” equilibrium cannot be found (§8.5.1). But it is also apparent that contestability theory provides a more robust analytical framework for examining the efficiency of industry structure than it does for predicting the outcome of strategic market behaviour, and there are many factors which could explain the lack of observed widespread unsustainability in real world utility industries (§8.4.5 and §9.1.1).

Sunk costs are therefore not necessarily the primary “villain of the piece”, at least as far as being responsible for unsustainability goes (§8.4.3). On the other hand, fixed costs do improve sustainability, as expected. Moreover, it is clear that where demands are sequenced, fixed costs do contribute to the incremental costs of the product for which demand commences first (§8.1.4).⁶ Hence, although common fixed costs are not entirely “the responsibility of the service that happened to be provided first” (§5.3.1), that initial service should contribute at least a return *on* the funds employed in the fixed cost outlay (§8.5.2). This seems wholly reasonable, since it reflects the opportunity cost of the resources involved.

The question remains though, if sunk costs are permitted in the TGTP model, and costs are not considered to be symmetric, then how can the model be considered to provide a pricing benchmark consistent with a paradigm of perfect contestability? Nevertheless, the important issue is not whether strategic interaction in the market *will* actually result in a sustainable, efficient and fair pricing equilibrium, but what that outcome itself *should* be. With respect to sunk costs, since an equilibrium solution is always assumed feasible, then costs will never be irrecoverable, even if they are irreversible. And the model demonstrates that irreversible costs associated with non-fungibility do not act as a barrier to entry, because non-fungibility is only considered to prevent asset transfer to alternative *uses*, and not to alternative *users* (§7.3.4). Hypothetical entrants, or self-producing consumer coalitions, are able to restrict revenues (or prices) to the opportunity costs of supply. Hence, neither irreversibility nor non-fungibility appear to preclude a contestable outcome, and irrecoverability is assumed away by the zero economic profit assumption. With respect to symmetry, it is difficult to imagine a model where costs are truly symmetric if the time dimension is adequately taken into account. Even if underlying cost

⁶ In the TGTP model (Chapters VIII and IX), fixed costs do not relate to capacity, but to expenditures such as one-off startup costs which do not relate to a license or physical asset which could potentially be resold. Fixed *capacity* costs also contribute to incremental costs, but it is assumed that the associated capacity can be resold to subsequent consumers; as such these costs are assumed to be subsumed into the capacity cost function.

functions are identical, as soon as demand sequencing occurs, symmetry cannot be maintained. But again, this does not appear to impact what appears to be a contestable outcome from the model, even if strictly-speaking this outcome cannot be considered “perfect”. Nevertheless, while the Bertrand-Nash assumption (§8.4.5 and §8.5.1) can either be relaxed or retained, the TGTP model needs to be further extended to examine cases without perfect information, and incorporate the effects of *uncertainty* (§9.4.5).

Allowing intraperiod entry or self-production is clearly the key for converting subsidy-free revenues to intertemporal subsidy-free prices, and for deriving the associated economic depreciation schedule (§9.1-§9.2). Based on the results of this simple two-good/two-period model (§9.2.4-§9.2.5), subsidy-free prices are clearly forward-looking, in the sense that they relate to the hypothetical costs incurred by current and future consumers, optimally meeting their own demand through self-production. Yet subsidy-free prices are not *so* forward-looking that they fallaciously ignore the *current* opportunity costs of capacity, even if that capacity had been “sunk” in the past. And while subsidy-free prices can reflect *both* historic and/or replacement costs, they will not necessarily reflect either (if referenced back to existing assets). Furthermore, subsidy-free prices cannot really be said to recover the costs of capacity “in advance”—in other words, pay for “tomorrow’s costs” (§7.5.4)—or to “signal” the need for future investment (§4.1.4), since subsidy-free prices simply reflect the opportunity costs inherent in the *currently* optimal asset configuration on a *greenfields* basis. Nevertheless, this configuration may itself reflect both current *and* future demand.

Where spare capacity would be optimally built today—in anticipation of future demand, and as the result of declining average incremental costs—then today’s prices should cover the amortised opportunity cost of the total capacity required to meet both current and future demand. Where capacity does not require expansion or replacement until some later date, then initially prices only need to cover the amortised opportunity cost of the capacity optimally required to meet current demand alone, but they should rise to the opportunity cost of total capacity at that time when it would become optimal for consumers (supplying themselves on a greenfields basis) to construct capacity sufficient to meet both current and anticipated demand. These results reaffirm Marcel Boiteux’s position that, where capacity is optimally constructed in *anticipation* of future demand, it should have its own income (§6.1.1). It also reaffirms Ralph Turvey’s view that the expectation of lower costs in the future raises today’s costs and, by association, today’s prices (§7.3.3), providing—in relevant circumstances—justification for *accelerated* depreciation (§9.2.2-§9.2.3). In general, subsidy-free prices are those which rise (or fall) to the current level of the opportunity cost (i.e., value) of assets to both current and future consumers, since this reflects the willingness of today’s consumers to pay for those assets (§9.2.4). This willingness-to-pay is rationally capped by the net intertemporal stand alone cost of self-production, since this nets out of the initial asset cost the discounted resale value of those assets to future consumers, based in turn on the willingness-to-pay of those future consumers. Consequently, incumbent firms cannot

capture the total consumer surplus between the demand and supply curves, since willingness-to-pay will be constrained by the least cost alternative supply source, which is the consumers' own cost of self-production, net of asset resale value. On the other hand, where there is perfect information, firms can still make a zero economic profit, so their interests are protected as well. Firms can raise prices through accelerated depreciation when the amortised costs of consumer self-production would exceed their own amortised costs.

10.5 The Efficiency and “Fairness” of Line Charges in New Zealand

While the notion that electricity distribution is a natural monopoly has provided the theoretical underpinnings for the reforms of New Zealand's power distribution sector (§2.1.5), the extent of natural monopoly has not been established, nor its characteristics well defined (§3.2.2). The concept is often used simply to refer to any situation in which there appears to be an absence of competition or contestability. If not used correctly, or if the natural monopoly concept is simply not relevant to particular activities in the distribution sector, then use of the term should be abandoned. Instead, the focus should be directed toward the real question, which is whether the reforms have led to the Government's desired outcomes: economic efficiency in all its forms, and prices that are fair and subsidy-free. As such, assessments of market power and barriers to entry may be more pertinent than those for “natural monopoly”.

In any case, by taking the approach that the “industry should regulate itself” through a light-handed regulatory regime based primarily on information disclosure (§2.1.1), the Government has—at least up until August 2001—allowed the industry to potentially find its own “naturally” least cost ownership structure, organisational structure, and commercial relationships, particularly in terms of the number of firms, the functions which each of those interrelating firms have, and the nature and conditions of contractual agreements. To some extent, such a regime has been put in place recognising that a more heavy-handed approach would be limited by the asymmetry of information between regulators and the regulated, while taking steps to remedy that asymmetry. Interestingly, electricity distribution in New Zealand appears to exhibit some of the key characteristics of a contestable market, at least where it comes to connecting *new* consumers (§3.3.3). Distribution line charges under the light-handed regulatory regime are relatively “sticky”—a weak form of the Bertrand-Nash assumption (§2.1.7)—and the open access nature of the regime notionally allows entrants to connect to any part of an incumbent's own network, thereby forgoing the need to duplicate the incumbent's “sunk” costs (§3.3.1). Over time, an information disclosure regime could produce a wealth of data with which hypotheses about the efficiency and contestability of the power distribution market can be tested. There is, however, little of such analysis being performed, and this thesis does not add to the body of knowledge which could be gleaned from such econometric work.

Nevertheless, in light of the results found from the two-good/two-period model presented in this thesis, some preliminary observations can be made regarding the efficiency and fairness of power distribution line charges in New Zealand. Initially, a cursory examination of line charge methodologies (§4.4.3) has indicated that electricity line businesses (ELBs) generally apply some form of the economically-discredited fully distributed cost allocation method (§4.2.3). In fact, such an approach was recommended by the Ministry of Commerce itself (§4.4.2), even though Government agencies are clearly aware of the principles and debates underlying allocatively efficient pricing (§4.4.1). Moreover, while it has been claimed that, in the sense of relative price levels, cross-subsidies have been eliminated, such a conclusion has been solely based on an examination of the prices themselves, and not on the underlying costs (§4.4.4).

As to absolute price levels, many ELBs have set their annual revenue requirements indexed to the optimised deprival valuation (ODV) of their network (§7.4). The ODV approach is to some extent consistent with contestable outcomes—at least in theory (§7.4.1)—particularly given its optimisation requirement based on a quasi-greenfields network (§7.4.2), and its economic value cap on revenues, based on all alternative means of supply (§7.4.3). Nevertheless, the implementation of the method in a New Zealand context is inconsistent with the intertemporal subsidy-free prices derived from the TGTP model for two reasons. First, given the long lifetimes of network assets, the ODV approach—as implemented in New Zealand—permits a short planning horizon (i.e., less than ten years), and attributes spare capacity with a negative connotation (§7.5.3). Second, requiring that assets be depreciated using straight line depreciation is not consistent with intertemporal subsidy-free pricing (§9.3.1-§9.3.2).

This thesis has not examined the extent to which the notional optimisation performed in deriving the ODVs adequately reflects a truly optimal greenfields network (subject to realistic real world constraints such as local government regulations, and geographical features). However, if it is assumed that the optimisations have been performed correctly (and the modern equivalent asset values used in the valuation are correct), then there appears to be little evidence to suggest that—at least on average—ELBs are unduly exercising any market power, assuming that they could do so if they desired (§9.3.3). And while this thesis has not assessed the validity of the claim that ELBs realised major windfall profits as a result of the change from historic book valuation methods to the ODV approach (§7.5.1), such a criticism is not relevant to whether current line charges adequately reflect today's opportunity costs of supply. Should compensation for consumers be warranted with respect to past revaluation gains, then this could potentially be settled outside the price mechanism without affecting ELB investment decisions or consumer consumption decisions going forward. So, in conclusion, under New Zealand's light-handed information disclosure regime for electricity line businesses, line charges and depreciation schedules do not appear to have exhibited the characteristics of efficient and "fair" prices embodied in the principles of intertemporal subsidy-free pricing.

BIBLIOGRAPHY AND REFERENCES

New Zealand Electricity Line Business Data and Associated Unpublished Reports

- Alpine Energy (1999). *Report on the Optimised Deprival Valuation of Alpine Energy Limited's Distribution Assets*. Report prepared by KPMG for Alpine Energy Ltd., May, Timaru.
- Alpine Energy (2001a). *Alpine Energy Ltd. Lines Business Financial Statements for the Year Ended 31 March 2001*, Alpine Energy Ltd., Timaru.
- Alpine Energy (2001b). *Asset Management Plan 2001-2011, Second Review June 2001*. Alpine Energy Ltd., Timaru.
- Alpine Energy (2001c). *Line Charges Effective 1 June 2001*. Alpine Energy Ltd., Timaru.
- Buller Electricity (2000). *Form for the Derivation of Financial Performance Measures from Financial Statements*. Buller Electricity Ltd., Buller.
- Centralines (1999). *Summary ODRC Asset Register 31 January 1999*. Centralines Ltd., Waipukurau.
- Centralines (2000). *Form for the Derivation of Financial Performance Measures from Financial Statements*. Centralines Ltd., Waipukurau.
- Centralpower (2000). *Information Disclosure – Procedure for the Derivation of Line Charges*. Centralpower Ltd., Palmerston North.
- Counties Power (1998). *Summary ODV for Counties Power Ltd. as at March 1998*. Counties Power Ltd., Pukekohe.
- Counties Power (2001a). *Disclosure of Valuation Report 31 March 2001*. Counties Power Ltd., Pukekohe.
- Counties Power (2001b). *Counties Power Ltd. – Disclosure of Financial Statements for the Year Ended 31 March 2001*. Counties Power Ltd., Pukekohe.
- Dunedin Electricity (1997). *Information Disclosure by Dunedin Electricity Ltd. – Summary of ODRC Asset Register at 1 January 1997*. Dunedin Electricity Ltd., Dunedin.
- Dunedin Electricity (2001a). *Use-of-System Pricing Methodology*. Dunedin Electricity Ltd., Dunedin.
- Dunedin Electricity (2001b). *Information Disclosure by Dunedin Electricity Ltd. for the Year Ended 31 March 2001*. Dunedin Electricity Ltd., Dunedin.
- Dunedin Electricity (2001c). *Information Disclosure by Dunedin Electricity Ltd. – Summary of ODRC Asset Register at 31 March 2001*. Dunedin Electricity Ltd., Dunedin.
- Eastland Network (2001a). *Schedule of New Charges, 1 April 2001*. Eastland Network Ltd., Gisborne.
- Eastland Network (2001b). *Information for Disclosure 2000/2001*. Eastland Network Ltd., Gisborne.
- Electricity Ashburton (2001a). *Network Line Charge Allocation Methodology*. Electricity Ashburton Ltd., Ashburton.
- Electricity Ashburton (2001b). *Disclosure 2001*. Electricity Ashburton Ltd., Ashburton.
- Electricity Ashburton (2001c). *Line Business ODV Valuation as at 31 March 2001*. Electricity Ashburton Ltd., 13 August, Ashburton.
- Electricity Invercargill (1998). *Valuation of the Network Business System Assets of Electricity Invercargill Limited*. ODV Valuation Undertaken for Electricity Invercargill Ltd. by Ernst and Young and Worley Consultants Ltd., 22 June 1998, Invercargill.

Electricity Invercargill (2000a). *Powernet Line Pricing Methodology for the Electricity Invercargill Ltd. Network*. Electricity Invercargill Ltd., Invercargill.

Electricity Invercargill (2000b). *Financial Statements as Required by the Electricity (Information Disclosure) Regulations 1999 for the Year Ended 31 March 2000*. Electricity Invercargill Ltd., Invercargill.

Hawke's Bay Network (2001). *2001 Information Disclosure*. Hawke's Bay Network Ltd., Napier.

Hawke's Bay Power (1997). *ODV Valuation 31 March 1997*. Report prepared by Coopers and Lybrand for Hawke's Bay Power Ltd., Napier.

Horowhenua Energy (2000). *Horowhenua Energy Ltd. Lines Business – Disclosure 2000*. Electralines Ltd., Levin.

Mainpower (2001). *Information Disclosure for the Year Ended 31 March 2001*, Mainpower New Zealand Ltd.

Marlborough Lines (2000). *Financial Statements Prepared in Accord with the Electricity (Information Disclosure) Regulations 1999 and Amendment Regulations 2000*. Marlborough Lines Ltd., Blenheim.

Nelson Electricity (2000). *Statement of Financial Performance for the 12 Months Ended 31 March 2000*. Nelson Electricity Ltd., Nelson.

Nelson Electricity (2001). *Pricing Methodology Disclosure*. Nelson Electricity Ltd., Nelson.

Network Tasman (2001). *Information for Disclosure for the Year Ended 31 March 2001*. Network Tasman Ltd., Nelson.

Network Waitaki (1999). *Allocation of Assets to Network Component Groups*. Network Waitaki, June, Timaru.

Network Waitaki (2000a). *Report on the Optimised Deprival Valuation of Network Waitaki Limited's Electricity Distribution Business*. Report prepared by KPMG for Network Waitaki Ltd., March, Timaru.

Network Waitaki (2000b). *Information for Disclosure*. Network Waitaki Ltd., July, Timaru.

Network Waitaki (2000c). *Line Charge Methodology*. Network Waitaki Ltd., September, Timaru.

Northpower (1998). *Network Asset Valuation as at 31 March 1998 – Results Summary*. From Report prepared by Ernst and Young for Northpower Ltd., Kaikohe.

Orion (1997). *Network Assets Valuation as at 31 March 1997 Including Optimisation and Economic Adjustment to Obtain the Optimised Deprival Value (ODV)*. Report prepared by Ernst and Young for Orion New Zealand Ltd., Christchurch.

Orion (2001a). *Valuation Report – ODV of System Fixed Assets as at 31 March 2001*. Orion Ltd., June, Christchurch.

Orion (2001b). *Information for Disclosure*. Orion Ltd., August, Christchurch.

Otago Power (1998). *Valuation of the Network Business Assets of Otago Power Limited*. Optimised Deprival Valuation Undertaken for Otago Power Ltd. by Ernst and Young and Worley Consultants Ltd., 20 July 1998, Balclutha.

Otago Power (2001). *Otago Power Limited Line Business Financial Statements for the Year Ended 31 March 2001*. Otago Power Ltd., Balclutha.

Powerco (2001). *Certification of Valuation Report of Line Owners*. Powerco Ltd., Palmerston North.

Scanpower (1998). *Valuation of Scanpower Ltd. Line Business as at 31 March 1998*. Scanpower Ltd., Dannevirke.

Scanpower (2001). *Optimised Deprival Valuation of the System Fixed Assets of Scanpower Ltd. as at 31 March 2001*. Report prepared by E-DEC Ltd. for Scanpower Ltd., Dannevirke.

Tasman Energy (1998). *Tasman Energy Network ODV, 31 March 1998*. Tasman Energy Ltd., Nelson.

The Power Company (1999). *Report on the Optimised Deprival Valuation of The Power Company's Distribution Business*. Report prepared by KPMG for The Power Company Ltd., August, Invercargill.

The Power Company (2001). *The Power Company Limited Line Business Statement of Accounts for the Year Ended 31 March 2001*. The Power Company Ltd., Invercargill.

TransAlta (1997). *ODV Valuation Electricity Network at 31 March 1997*. TransAlta New Zealand Ltd., August, Wellington.

United Networks (2001). *United Networks Limited – Financial Statements, Performance Measures & Statistics Disclosure for the Year Ended 31 March 2001*. United Networks Ltd., Takapuna.

Vector (1999). *Vector Optimised Deprival Valuation for the Year Ending 31 March 1999*. Valuation Report, 17 May, Vector Ltd., Auckland.

Vector (2001a). *Vector Optimised Deprival Valuation for the Year Ending 31 March 2001*. Valuation Report, 18 July, Vector Ltd., Auckland.

Vector (2001b). *Vector Ltd. Electricity Lines Business Statement of Financial Performance for the Year Ended 31 March 2001*. Vector Ltd., Auckland.

Waipa Networks (2001a). *Certification of Valuation Report of Line Owners*. Waipa Networks Ltd., Te Awamutu.

Waipa Networks (2001b). *Waipa Networks Limited Line Business - Form for the Derivation of Financial Performance Measures from Financial Statements*. Waipa Networks Ltd., Te Awamutu.

WEL Energy (1998). *Summary ODRC Asset Register 1998*. WEL Energy Ltd., Hamilton.

WEL Energy (2000). *Form for the Derivation of Financial Performance Measures from Financial Statements*. WEL Energy Ltd., Te Awamutu.

Westpower (1999). *ODV Valuation of Line Business as at 31 March 1999*. Westpower Ltd., August, Greymouth.

Westpower (2000). *Westpower Limited Information Disclosure*. Westpower Ltd., Greymouth.

New Zealand Acts of Parliament, Bills and Regulations

Commerce Act 1986.

Electric Power Boards Amendment Act 1990.

Electricity Act 1992.

Electricity Industry Bill 2000.

Electricity Reform Act 1998.

Electricity (Information Disclosure) Regulations 1994.

Electricity (Information Disclosure) Regulations 1999.

Electricity (Information Disclosure) Regulations 1999 Consolidated with the Electricity (Information Disclosure) Amendment Regulations 2000.

Energy Companies Act 1992.

Resource Management Act 1991.

State Owned Enterprises Act 1986.

New Zealand Government Publications, Reports, Media Releases, Letters and Speeches

- Arthur Young (1989). *Retail Electricity Pricing: a Review of Historical Trends and Future Influences*. Report prepared for the Market Analysis Unit of the Ministry of Energy, Report 89/1004, Wellington.
- Bradford M (1999). *Electricity Line Company Controls Introduced into Parliament*. Media Release from the Minister of Energy, 25 May, Wellington.
- Butcher D., (1990a). *Aim of Reform of Electricity Distribution is Lower Prices*. Media Release from the Minister of Commerce and Minister of Energy, 25 May, Wellington.
- Butcher D., (1990b). *Electricity Reforms an Incentive to Job Growth*. Media Release from the Minister of Commerce and Minister of Energy, 25 July, Wellington.
- Cabinet Policy Committee (1989). *Electricity Distribution Restructuring: the Ownership Issue*. New Zealand Cabinet Policy Committee, October, Wellington.
- Commerce Commission (2001). *Price Control Study of Airfield Activities at Auckland, Wellington, and Christchurch International Airports*. Draft Report of the New Zealand Commerce Commission, 3 July, Wellington.
- Energy Markets Policy Group (1998a). *Electricity Reform Package: Questions and Answers*. Resources and Networks Branch, Ministry of Commerce, April, Wellington.
- Energy Markets Policy Group (1998b). *A Better Deal for Electricity Consumers: an Outline of the New Zealand Government's Electricity Reform Package*. Resources and Networks Branch, Ministry of Commerce, April, Wellington.
- Energy Markets Policy Group (1998c). *Electricity Industry Reform: Discussion Paper on the Operation of the Specific Thresholds for Price Control for Electricity Line Businesses*. Resources and Networks Branch, Ministry of Commerce, December, Wellington.
- Energy Markets Policy Group (1999). *Purpose of the Price Control Regime*. Resources and Networks Branch, Ministry of Commerce, March, Wellington.
- Energy Markets Policy Group (2001a). *Chronology of New Zealand Electricity Reform*. Resources and Networks Branch, Ministry of Economic Development, February, Wellington.
- Energy Markets Policy Group (2001b). *New Zealand's Electricity Sector*. Resources and Networks Branch, Ministry of Economic Development, February, Wellington.
- Energy Markets Regulation Unit (1998). *Electricity Information Disclosure Statistics 1998*, Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Commerce, November, Wellington.
- Energy Markets Regulation Unit (1999). *Handbook for Optimised Deprival Valuation of System Fixed Assets of Electricity Line Businesses*, 3rd. edn., Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Commerce, April, Wellington.
- Energy Markets Regulation Unit (2000a). *Discussion Paper on the Requirement for Economic Valuations Under the Electricity ODV Handbook*. Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Economic Development, April, Wellington.
- Energy Markets Regulation Unit (2000b). *Electricity Information Disclosure Handbook – Handbook for Complying with the Electricity (Information Disclosure) Regulations 1999 as amended by the Electricity (Information Disclosure) Amendment Regulations 2000*. Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Economic Development, June, Wellington.

-
- Energy Markets Regulation Unit (2000c). *Handbook for Optimised Deprival Valuation of System Fixed Assets of Electricity Line Businesses*. 4th. edn., Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Economic Development, October, Wellington.
- Energy Markets Regulation Unit (2000d). *Electricity Information Disclosure Statistics 2000*, Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Economic Development, December, Wellington.
- Energy Markets Regulation Unit (2001). *Electricity (Information Disclosure Regulations) Newsletter* No. 19, Energy Markets Information and Services Group, Resources and Networks Branch, Ministry of Economic Development, May, Wellington.
- Energy Policy Group (1994a). *Electricity Disclosure Guidelines - Guidelines on Business Procedures to Assist Electricity Distributors Comply with the Information Disclosure Regime*, Energy and Resources Division, Ministry of Commerce, June, Wellington.
- Energy Policy Group (1994b). *Handbook for Optimised Deprival Valuation of Electricity Line Businesses*, Energy and Resources Division, Ministry of Commerce, June, Wellington.
- Energy Policy Group (1994c). *Electricity Information Disclosure – Adjustments to Financial Statements for Performance Measures*, Energy and Resources Division, Ministry of Commerce, August, Wellington.
- Energy Policy Group (1994d). *Guideline for Renewal Accounting of Infrastructure Assets for Performance Measurement Derivation*, Energy and Resources Division, Ministry of Commerce, August, Wellington.
- Energy Policy Group (1994e). *Questions and Answers on Financial Performance Measures, Optimised Deprival Valuations and Avoidance of Double Counting of Asset-Related Expenses*, Energy and Resources Division, Ministry of Commerce, August, Wellington.
- Energy Policy Group (1995). *Light-Handed Regulation of New Zealand's Electricity and Gas Industries*. Energy and Resources Division, Ministry of Commerce, October, Wellington.
- Ernst and Young (1990). *Retail Electricity Tariffs – The Impact of Commercialisation and Regulatory Changes*, Report to the Minister of Commerce and Energy, May, Wellington.
- Ernst and Young (1994). *Rationale for Financial Performance Measures in the Electricity Information Disclosure Regime – Including Use of Optimised Deprival Values, and Avoidance of Double Counting of Asset Related Expenses*, Report to Energy Policy Group, Energy and Resources Division, Ministry of Commerce, for briefing the Electricity Supply Association of New Zealand (ESANZ), August, Wellington.
- Gale, S. (1989) *A Transaction Cost Approach to Energy Contracts and Conservation*. Ministry of Energy Report RD8824, Wellington.
- Harris G., Gale S., Allan R., Lucas C., Marinkovich M. (1993). *Promoting the Market for Energy Efficiency*, Report prepared for the Officials Committee on Energy Policy, Wellington.
- Hodgson, P. (2000a). *A Fair Deal for Electricity Consumers*. Media Statement, Minister of Energy, 3 October, Wellington.
- Hodgson, P. (2000b). *Tougher Rules for Electricity Line Valuations*. Media Statement, Minister of Energy, 3 October, Wellington.
- Jefferies, P.A. (1997). *Vulnerable Groups: A Fertile Field for the Work of National Institutions for the Promotion and Protection of Human Rights*. Speech by the Chief Commissioner, Human Rights Commission, New Zealand, Fourth International Workshop of National Institutions for the Promotion and Protection of Human Rights, 27 November, Merida, Mexico.
- Kidd D. (1994). *Speech to the Annual General Meeting of the Electricity Supply Association of New Zealand*, Minister of Energy, 8 September, Wellington.
-

-
- Kidd D. (1995). *Speech to the Green Tiger Conference*, Minister of Energy, 21 March, Wellington.
- Lloyd K.A. (1986). *Contestability in Context*. Discussion Paper G86/5, Reserve Bank of New Zealand, October, Wellington.
- Luxton J. (1991a). *Speech to the Power Industry Reform Conference*, Minister of Energy, 27 August, Wellington.
- Luxton J. (1991b). *Speech to the Electricity Reform Conference*, Minister of Energy, 2 December, Wellington.
- Luxton J. (1992a). *Energy Policy Framework Announced*, Media Release by the Minister of Energy, 30 June 1992, Wellington.
- Luxton J. (1992b). *Letter to P. Cebalo, General Manager of the Auckland Electric Power Board*, Office of the Minister of Energy, 23 December 1992, Wellington.
- Ministry of Commerce (1991). *Electricity Industry Reform Information Disclosure Regime – Provisional Guidelines for Electricity Line and Electricity Business Procedures*, Ministry of Commerce, Wellington.
- Ministry of Commerce (1998). *Auckland Power Supply Failure 1998*. Report of the Ministerial Inquiry into the Auckland Power Supply Failure, Ministry of Commerce, July, Wellington.
- Ministry of Commerce (1999). *Review of the Optimisation Rules for Optimised Deprivation Valuation of New Zealand Electricity Networks*. Discussion Paper prepared by Parsons Brinkerhoff Associates for the Ministry of Commerce, December, Wellington.
- Ministry of Commerce (2000). *Ministerial Inquiry into the Electricity Industry: Issues Paper*. Office of the Ministerial Inquiry into the Electricity Industry, Ministry of Commerce, February, Wellington.
- Ministry of Commerce, The Treasury (1995). *Regulation of Access to Vertically-Integrate Natural Monopolies – a Discussion Paper*. Ministry of Commerce, and The Treasury, Wellington.
- Ministry of Economic Development (2000a). *Inquiry into the Electricity Industry*. Report to the Minister of Energy, June, Wellington.
- Ministry of Economic Development (2000b). *Energy Policy Framework*. Energy Resources and Safety Division, October, Wellington.
- New Zealand Government (1998). *Statement of Economic Policy: Market Power in the Electricity Sector*. 23 December, Wellington.
- New Zealand Government (2000). *Government Policy Statement – Further Development of New Zealand's Electricity Industry*. 7 December, Wellington.
- Peters W., Birch B., Bradford M. (1998a). *Summary: a Better Deal for Electricity Consumers*. Media Release by the Treasurer, the Minister of Finance and the Minister of Energy, 7 April, Wellington.
- Peters W., Bradford M. (1998b). *Electricity Distribution and Retail Reforms*. Media Release by the Treasurer and the Minister of Energy, 7 April, Wellington.
- Peterson D., Harper L., Sanson M. (1992). *Sustainable Energy Management in New Zealand: Improvements Required in Government Policy*, Report to the House of Representatives, Office of the Parliamentary Commission for the Environment, March, Wellington.
- Rural Supply Working Party (1990). *Electricity Distribution Industry Reform: Social and Rural Impacts*. Report of the Rural Supply Working Party, Energy and Resources Division, Ministry of Commerce, Wellington.
- Saha G., Sell P. (1990). *Retail Electricity Tariffs – the Impact of Commercialisation and Regulatory Changes*. Report prepared by Ernst and Young for the Minister of Commerce and Energy, Wellington.
- TPEB (1991). *The Separation of Transpower*. A Report to the Minister of State Owned Enterprises, Transpower Establishment Board, Wellington.

Treasury (1984). *Economic Management*, The New Zealand Treasury, Wellington.

Williamson B., Mumssen Y. (2000). *Economic Regulation of Network Industries*. Report prepared by National Economic Research Associates (NERA) for the New Zealand Treasury, Treasury Working Paper 00/5, Wellington.

Wyatt N., Brown M., Carataga P., Duncan A., Giles D. (1989). *Performance Measures and Economies of Scale in the New Zealand Electricity Distribution Industry*. Policy Division, Ministry of Energy, Wellington.

Non-Government Reports, Newspaper Articles, Unpublished Papers, Monographs, Dissertations, Letters and Submissions

Bergara M.E., Spiller P.T. (1996). *The Introduction of Direct Access in New Zealand's Electricity Market*, Report PWP-043, Program on Workable Energy Regulation, University of California Energy Institute, Berkeley, CA.

Bertram G. (1991b). Power industry reform, sustainability, and energy efficiency: some conflicts. Paper presented at *Power Industry Reform Conference*, August, Wellington.

Bertram G. (1992). Pricing as an integral part of electricity reform. Paper presented at *Electricity Reform Conference*, October, Wellington.

Bertram G., Dempster I., Gale S., Terry S. (1992). *Hydro New Zealand – Providing for the Progressive Pricing of Electricity*, Report Prepared for the Electricity Reform Coalition, Wellington.

Bertram G., Terry S. (2000). *Lining Up the Charges: Electricity Line Charges and ODV*. Report prepared by Simon Terry Associates Ltd., Wellington.

Bollard A.E. (1991). *Economic Liberalisation in New Zealand: It was the Best of Times, It was the Worst of Times*. Working Paper 91/5, New Zealand Institute of Economic Research, Wellington.

Burnell S., Evans L., Yao S. (1999). *The Optimal Interconnection Contract under Partial Bypass in Oligopolistic Network Industries*. New Zealand Institute for the Study of Competition and Regulation, Wellington.

Deane R. (1989). Reflections on privatisation. Paper presented at *IBM Utilities Industry Executive Conference*, September 1989, Bangkok.

Economides N. (1997). *The Tragic Inefficiency of the M-ECPR*. Stern School of Business, New York University, New York, NY.

EIA (2000). *The Changing Structure of the Electric Power Industry 2000: an Update*. Energy Information Administration, United States Department of Energy, Washington, DC.

Electricity Week (1995). ESANZ expresses relief at ECNZ split, but wanted more. *Electricity Week* 16(9), 1-2.

Ergas H., Small J.P. (2000). *Real Options and Economic Depreciation*. Discussion Paper, Centre for Research in Network Economics and Communications, University of Auckland, Auckland.

ESANZ (1993). *Code of Practice: Contestable Energy Trading*. Electricity Supply Association of New Zealand Ltd. (ESANZ), Wellington.

Evans L. (1998). *The Critical Importance of Information: Incentive Regulation and its Application in Electricity*. New Zealand Institute for the Study of Competition and Regulation, Wellington.

Evans L., Quigley N. (1998). *Common Elements in the Governance of Deregulated Electricity Markets, Telecommunications Markets and Payments Systems*, New Zealand Institute for the Study of Competition and Regulation, Wellington.

Evans L., Quigley N., Zhang J. (2000). *An Essay on the Concept of Dynamic Efficiency and its Implications for Assessments of the Benefits from Regulation and Price Control*, New Zealand Institute for the Study of Competition and Regulation, Wellington.

-
- Farley P. (1991). Ownership of electric power companies. Paper presented at the *Electricity Reform Conference*, Wellington.
- Fernyhough J. (1990). *Letter from the Chairman of ECNZ to Richard Prebble, Minister for State Owned Enterprises*, 14 September, Wellington.
- Gale S., Strong N. (1999). *Electricity Lines Businesses, Performance Indicators*. Report for the Ministry of Commerce, New Zealand Institute of Economic Research, March, Wellington.
- Gallagher J., Lewis C. (1988) *Energy Management In New Zealand - What Is Optimal Intervention?* New Zealand Institute of Economic Research, Contract No. 53, April, Wellington.
- Gardner T., Gilson L. (1994). Predictable patterns in the transition from protected monopoly to market competition. *Proceedings of the 10th CEPSI Conference*, Christchurch, 239-251.
- Gilbert E. (1997). *Cross-Subsidies Between Lines and Energy Businesses by Electricity Retailing Firms in New Zealand*. Unpublished BCom (Hons) Dissertation, The University of Auckland, Auckland.
- Giles D., Wyatt N. (1989). *Economies of Scale in the New Zealand Electricity Distribution Industry*, Discussion Paper No. 8904, Department of Economics, University of Canterbury, Christchurch.
- Gunn C. (1995a). Energy efficiency vs economic efficiency? New Zealand electricity sector reform in the context of the national energy policy objective. Paper presented at *Joint Conference of the New Zealand Association of Economists, and the Law and Economics Association of New Zealand*, Lincoln University, 29 August, Christchurch.
- Gunn C. (1995b). 'Understanding' or 'The Willing Suspension of Disbelief' – Neo-Austrian Microeconomics vs the Mainstream. Unpublished Dip. Com. paper, Department of Economics, The University of Auckland, Auckland.
- Gunn C. (1996). *Optimal Energy Company Behaviour in New Zealand's Deregulated Electricity Market: Issues and Comments*. Working Papers in Economics No. 156, Department of Economics, The University of Auckland.
- Gunn C. (2002). On intertemporal subsidy-free prices and economic depreciation: constrained market pricing revisited. Under review for publication in the *Journal of Regulatory Economics*.
- Guthrie G., Small J.P., Wright J. (2000). *Pricing Access: Forward versus Backward Looking Cost Rules*. Discussion Paper, Centre for Research in Network Economics and Communications, The University of Auckland, Auckland.
- Harvard Business School (1998). *Clear Communications Ltd. vs. Telecom Corporation of New Zealand*. Harvard Business School N9-798-085, Harvard Business School, Boston, MA.
- Hazledine T. (1992). *Industrial Organization: An Introductory Survey*. Working Papers in Economics No. 107, Department of Economics, The University of Auckland, Auckland.
- Hazledine T. (1994). *The 'Public Interest' in Competition Policy: Due Process or Economic Rationalism?* Working Papers in Economics No. 142, Department of Economics, The University of Auckland, Auckland.
- Heald D. (1994). *Cost Allocation and Cross Subsidies*, European Commission, Brussels.
- Heffernan D. (1993). Power company wealth creation. Paper presented to *AESIEAP Chief Executives' Conference*, September, Maramarua.
- Hunt S. (1994). The problems of transmission pricing. *Power Program on Workable Regulation – Electricity and the Changing Structure of the Electricity Industry*, Oakland, CA.
- Hunter B., Matheson A. (1994). Researching and monitoring New Zealand's competitive energy market. *Proceedings of the 10th CEPSI Conference*, Christchurch, 573-583.

-
- IEA (1989). *Energy Policies and Programmes of IEA Countries: 1988 Review*, International Energy Agency, Organisation for Economic Co-operation and Development (OECD), Paris.
- IEA (1991). *Utility Pricing and Access: Competition for Monopolies*, International Energy Agency, Organisation for Economic Co-operation and Development (OECD), Paris.
- IEA (1999). *Energy Use and Efficiency in New Zealand in an International Perspective: Comparison of Trends through 1995*. Unpublished Report by the International Energy Agency and the Lawrence Berkeley National Laboratory, Berkeley, CA.
- IEA (2001a). *Competition in Electricity Markets*, International Energy Agency (IEA), Organisation for Economic Co-operation and Development (OECD), February, Paris.
- IEA (2001b). IEA commends New Zealand energy policy, but points to mixed objectives in electricity reform. *IEA/Press (01)16*, International Energy Agency (IEA), Organisation for Economic Co-operation and Development (OECD), 11 July, Paris.
- IEA (2001c). *Energy Policies of IEA Countries – New Zealand 2001 Review*, International Energy Agency (IEA), Organisation for Economic Co-operation and Development (OECD), July, Paris.
- Irwin T. (2000). *Fostering Innovation in the Electricity Industry*. Discussion Paper prepared by Law and Economics Consulting Group (LECG) for Orion New Zealand Ltd., 10 March, Wellington.
- Jackson K. (1990). *Electricity Provision and the Concept of Service in New Zealand: an Historical Example of Pricing Policies*. Discussion Paper No. 71, Department of Economics, University of Auckland, Auckland.
- Jagger C. (1996). *Access Issues in Electricity Distribution and Supply*. Unpublished LLB(Hons) dissertation, Faculty of Law, The University of Auckland, Auckland.
- Kahn A. (2001). *Statement of Alfred E. Kahn On Behalf Of Auckland International Airport Ltd.*, 10 August, 2001, Ithaca, NY.
- Kask S. (1988a) *A Conceptual Framework for Energy Policy Analysis*. New Zealand Energy Research and Development Committee, Report No. 147, Auckland.
- Kask S. (1988b) *Organisational Structure and Pricing: Preliminary Findings from a Review of Retail Electricity Pricing in New Zealand*. New Zealand Energy Research and Development Committee, Report No. 148, Auckland.
- Kask S., Saha G. (1986) *The New Zealand Electricity Industry: a Structural Analysis of Competition, Contestability and Efficiency*. Discussion Paper No. 24, Department of Economics, University of Auckland.
- Kennedy School of Government Case Program (1998). *Price Cap: The UK's Efforts to Regulate Regional Distribution Companies*, Kennedy School of Government, Harvard University, Cambridge, MA.
- Kerr R. (1999). Electricity pricing: competition and regulation. Paper presented at *AIC Worldwide New Zealand National Power Conference*, Wellington.
- Leay B. (1995). Benefits of electricity reforms now clearly emerging. *Letter to Chairpersons and CEOs of New Zealand Power Companies*, Executive Director of the Electricity Supply Association of New Zealand (ESANZ), 10 March, Wellington.
- Leyland B. (2000). *Submission to the Ministerial Inquiry into the Electricity Industry*. 3 March, Auckland.
- Lough R. (1994). Transition to a competitive market and its impact on energy efficiency. Paper presented at *3rd International Energy Efficiency and DSM Conference*.
- Lowrey M.N., Kaufmann L. (1998). *Price Cap Regulation of Electricity Distribution*. Fourth draft, 6 March, report prepared for the Edison Electric Institute by Laurits R. Christensen Associates, Madison, WI.
-

-
- McDonald T. (1991). Electricity industry restructuring and pricing: the fundamental mistakes. Paper presented at *Power Industry Reform Conference*, August, Wellington.
- McLachlan A. (1992). Competition in generation: the Geotherm story. Paper presented at *AIC Conference on Electricity Reform*, October, Wellington.
- McLay J. (1993). Electricity deregulation and the open market: the New Zealand experience; is it transportable? Paper presented at *AESIEAP Chief Executives' Conference*, September, Maramarua.
- Mercury Energy Lines Business (1999). *Regulation Tailored for the Future*. Submission on the Government's Discussion Paper "Operation of the Specific Thresholds for Price Control of Electricity Line Businesses", Mercury Energy, Auckland.
- Miyazaki H. (1990). *Rents and Instability in Contestable Markets*. Discussion Paper 233, The Institute of Social and Economic Research, Osaka University, Osaka.
- Murray K., Hansen C., Cheng W.L., Irwin T. (2001). *A Critique of the Commerce Commission's Draft Findings on the Efficiency of Price Control of WIAL*. Report submitted by Law and Economics Consulting Group (LECG) on behalf of Wellington International Airport Ltd. to the New Zealand Commerce Commission, 14 August, Wellington.
- Noble J. (1991). Electricity reform: pricing. The Clayton's competition pricing game. Paper presented at *Electricity Reform Conference*, December, Wellington.
- Noble J. (1992). Energy trading and brokerage: prospects and problems. Paper presented at *AIC Conference on Electricity Reform*, October 1992, Wellington.
- NZBR (1998). *Submission on the Electricity Industry Reform Bill*. New Zealand Business Roundtable, Wellington.
- NZBR (1999). *Submission on the Discussion Paper on the Operation of the Specific Thresholds for Price Control for Electricity Line Businesses*, New Zealand Business Roundtable, Wellington.
- NZIER (1997). *Metering and Profiling. Competition for Small Electricity Consumers*. Report to the Ministry of Commerce, by the New Zealand Institute of Economic Research (NZIER), September, Wellington.
- NZPA (2001a). NGC crisis will change face of electricity sector – analysts. *New Zealand Press Association*, 28 June, Wellington.
- NZPA (2001b). Line companies could be forced to lower prices – electricity lines companies wary after Commission assessment. *New Zealand Press Association*, 10 July, Wellington.
- O'Sullivan F. (1995) ECNZ monopoly set to become duopoly? *National Business Review*, 16 June, Wellington.
- OECD (1997). *Application of Competition Policy to the Electricity Sector*. Report OCDE/GD(97)132, Organisation for Economic Co-operation and Development (OECD), Paris.
- Outhred H.R., Kaye R.J. (1994). Incorporating network effects in a competitive electricity industry—an Australian perspective. Paper presented at the *Institute of Management Sciences Conference*, Alaska.
- Parsons Brinckerhoff Associates (2002). *Recalibration of Large Electricity Line Owners – Closing Report*, Report Prepared for Commerce Commission, August, Wellington.
- Patterson R.H. (1995). How Chicago School theology hijacked New Zealand competition law and policy. Paper presented at *Joint Conference of the New Zealand Association of Economists, and the Law and Economics Association of New Zealand*, Lincoln University, 29 August, Christchurch.
- PricewaterhouseCoopers (2002). *Electricity Line Business 2001 Information Disclosure Compendium*, March, Auckland.
- Regulatory Assistance Project (2000). *Best Practices Guide: Implementing Power Sector Reform*. Prepared by the Regulatory Assistance Project, Maine, Vermont, for the Energy and Environment Training Program,
-

-
- Office of Energy, Environment and Technology, United States Agency for International Development, Washington, DC.
- Russell D. (1991a). The residential consumers' view of electricity distribution and pricing. Paper presented at *Price Interactions in the Energy Market, New Zealand National Committee, World Energy Council Seminar*, March, Wellington.
- Russell D. (1991b). Where are the gains to the consumer? Paper presented at the *Electricity Reform Conference*, December, Wellington.
- Saha G. (1991). Competition in the power industry. Paper presented at the *Power Industry Reform Conference*, August 1991, Wellington.
- Saha G. (1993). Asset valuation and ownership transfer: key issues. Paper presented at *AESIEAP Chief Executives' Conference*, Maramarua.
- Sharp B.M., Baker W.R., Bodger P.S. (1985). *The Economic Cost of Energy Non-Supply*. New Zealand Energy Research and Development Committee, Report No. 117, Auckland.
- Simon Terry Associates (2000). *How Will the Initial Ratebase be Determined?* Submission to the Ministerial Enquiry into the Electricity Industry, Simon Terry Associates Ltd., March, Wellington.
- Simon Terry Associates (2001). *Submission to the Commerce Commission on Price Control of Airfield Activities at Auckland, Wellington and Christchurch International Airports – Draft Report*. Report prepared by Simon Terry Associates Ltd., August, Wellington.
- Small J.P. (1998a). *Credibility on the Line: a Commentary on the Electricity Reform Package*. Discussion Paper, Centre for Research in Network Economics and Communications (CRNEC), May, University of Auckland, Auckland.
- Small J. (1998b). *Real Options and the Pricing of Network Access*. Discussion Paper, Centre for Research in Network Economics and Communications, University of Auckland, Auckland.
- Small J.P. (1999a). *"Specific" Thresholds for Price Control of Electricity Line Business*. Submission on the Ministry of Commerce's Discussion Paper on the Operation of the Specific Thresholds for Price Control for Electricity Line Businesses, Centre for Research in Network Economics and Communications (CRNEC), February, University of Auckland, Auckland.
- Small J.P. (1999b). *The Design of Economic Regulation for Utilities: a Primer*. Paper written for the CRNEC Policy Conference, Centre for Research in Network Economics and Communications (CRNEC), September, University of Auckland, Auckland.
- Small J.P. (1999c). *Regulation and Competition Law of Networks in New Zealand*. Paper written for the CRNEC Policy Conference, Centre for Research in Network Economics and Communications (CRNEC), September, University of Auckland, Auckland.
- Small J.P., Ergas H. (1999). *The Rental Cost of Sunk and Regulated Capital*, Discussion Paper, Centre for Research in Network Economics and Communications, University of Auckland, Auckland.
- Small V. (1995). Utilities in orgy of corporate capitalism. *National Business Review*, 17 February, Wellington.
- SOLEC (1992). *Guide to Derivation of Line Charges*. Separation of Line and Energy Charges (SOLEC) Working Party, Wellington.
- Spong R. (1998). *Interconnection Under Light Handed Regulation*. Unpublished BCom(Hons) dissertation, Department of Economics, University of Auckland, Auckland.
- Steiner F. (2000). *Regulation, Industry Structure and Performance in the Electricity Supply Industry*. Economics Working Papers No. 238, ECO/WKP(2000)11, Organisation for Economic Co-operation and Development (OECD), Paris.
-

-
- Strong N., Gale S. (1999). *Electricity Line Business Performance*. Report for the Ministry of Commerce, New Zealand Institute of Economic Research, June, Wellington.
- Teplitz-Sembitzky, W. (1990). *Regulation, Deregulation, or Reregulation – What is Needed in the LDC’s Power Sector?* Energy Series Paper No. 30, Industry and Energy Department, The World Bank, Washington, DC.
- Teplitz-Sembitzky, W. (1992). *Electricity Pricing: Conventional Views and New Concepts*. Energy Series Paper No. 52, Industry and Energy Department, The World Bank, Washington, DC.
- Terry S. (1991). *Making a Market for Energy Efficiency*, New Zealand Planning Council, Wellington.
- Trebing H.M. (1967). Competition: an asset to utility regulation? Paper presented at *Public Utilities Seminar of American Marketing Association*, Chicago.
- US Supreme Court (1944). *Federal Power Commission v. Hope Natural Gas Co.*, Supreme Court of the United States of America, 320 U.S. 591 (1944).
- Valletti T.M, Estache A. (1999). *The Theory of Access Pricing: an Overview for Infrastructure Regulators*, Policy Research Working Paper 2097, Governance, Regulation, and Finance Division, The World Bank Institute, The World Bank, Washington, DC.
- WEMS (1992). *Wholesale Electricity Market Reform: Analysis of Options*. WEMS, Wellington.
- Willig R.D. (1979a). Consumer equity and local measured service, in J.A. Baude (ed.), *Perspectives on Local Measured Service*, Telecommunications Industry Workshop, Kansas City, MO.
- Wilson J.W. (1994). Valuation and regulation of New Zealand electricity companies: progress and issues. *10th CEPSI Conference Proceedings*, Christchurch, Vol. 2, 156-168.
- Wilson J.W. (2000a). *Review of Optimisation Rules for ODV Valuations of Electricity Networks – Comments on Discussion Paper*. Letter from Worley International to the Ministry of Commerce, 15 February, Auckland.
- Wilson J.W. (2000b). *Submission to the Ministerial Inquiry into the Electricity Industry*, 13 March, Auckland.
- Wilson P. (2001). Government fights its way through another energy policy debate. *New Zealand Press Association*, 28 June, Wellington.
- Zijl T. van, Irwin T. (2001). *Historic Cost and Replacement Cost: Efficiency Implications of their Use in Price Setting*. Report submitted by Law and Economics Consulting Group (LECG) on behalf of Wellington International Airport Ltd. to the New Zealand Commerce Commission, 14 August, Wellington.

Published Works

- Aigner D.J. (1984). The welfare econometrics of peak-load pricing for electricity. *Journal of Econometrics* 26, 1-15.
- Aigner D.J., Leamer E.E. (1984). Estimation of time-of-use pricing response in the absence of experimental data. *Journal of Econometrics* 26, 205-227.
- Andersson R., Bohman M. (1985). Short- and long-run marginal cost pricing. On their alleged equivalence. *Energy Economics* 7(Oct), 279-288.
- Areeda P., Turner D.F. (1975). Predatory pricing and related practices under Section 2 of the Sherman Act. *Harvard Law Review* 88, 637-733.
- Armstrong M., Doyle C., Vickers J. (1996). The access pricing problem: a synthesis. *Journal of Industrial Economics* 44(2), 131-150.
- Arrow K.J. (1968). Optimal capital policy with irreversible investment, in J.N. Wolfe (ed.), *Value Capital and Growth, Essays in Honour of Sir John Hicks*, Edinburgh University Press, Edinburgh.

-
- Averch H., Johnson L. (1962). Behavior of the firm under regulatory constraint. *American Economic Review* 52, 178-183.
- Bailey E.E. (1981). Contestability and the design of regulatory and antitrust policy. *American Economic Review* 71(2), 178-183.
- Bain J.S. (1954). Economies of scale, concentration and the conditions of entry in twenty manufacturing industries. *American Economic Review* 64, 15-39.
- Banks C. (1994). Electricity pricing in Sweden in theory and practice. *Energy Sources* 16, 519-530.
- Baumol W.J. (1968). Reasonable rules for rate regulation, plausible policies for an imperfect world, in A. Philips and O. Williamson (eds), *Prices: Issues in Theory, Practice and Public Policy*, University of Pennsylvania Press, PA.
- Baumol W.J. (1971). Optimal depreciation policy: pricing the products of durable assets. *Bell Journal of Economics and Management Science* 2, 638-656.
- Baumol W.J. (1977). On the proper cost tests for natural monopoly in a multiproduct industry. *American Economic Review* 67(5), 809-822.
- Baumol W.J. (1979). Minimum and maximum pricing principles for residual regulation. *Eastern Economic Journal* 5(1-2), 235-248.
- Baumol W.J. (1986). *Superfairness: Applications and Theory*. MIT Press, Cambridge, MA.
- Baumol W.J., Bailey E.E., Willig R.D. (1977). Weak invisible hand theorems on the sustainability of multiproduct natural monopoly. *American Economic Review* 67(3), 350-365.
- Baumol W.J., Bradford D.F. (1970). Optimal departures from marginal cost pricing. *American Economic Review* 60, 265-283.
- Baumol W.J., Koehn M.F., Willig R.D. (1987). How arbitrary is 'arbitrary'?—or, toward the deserved demise of full cost allocation. *Public Utilities Fortnightly* 120(5), 3 September, 16.
- Baumol W.J., Panzar J., Willig R. (1988). *Contestable Markets and the Theory of Industry Structure*, 2nd edn. Harcourt Brace Jovanovich, New York.
- Baumol W.J., Sidak J.G. (1994a). *Toward Competition in Local Telephony*. MIT Press, and American Enterprise Institute Press, Washington, DC.
- Baumol W.J., Sidak G.J. (1994b). The pricing of inputs sold to competitors. *Yale Journal of Regulation* 11, 171-202.
- Baumol W.J., Sidak J.G. (1995a). *Transmission Pricing and Stranded Costs in the Electric Power Industry*. American Enterprise Institute Press, Washington, DC.
- Baumol W.J., Sidak J.G. (1995b). Stranded costs. *Harvard Journal of Law and Public Policy* 18, 835-851.
- Baumol W.J., Willig R. (1981). Fixed cost, sunk cost, entry barriers and the sustainability of monopoly. *Quarterly Journal of Economics* 96, 405-432.
- Beesley M.E., Littlechild S.C. (1989). The regulation of privatized monopolies in the United Kingdom. *Rand Journal of Economics* 20(3), pp. 454-472.
- Berg S.V., Tschirhart J. (1988). *Natural Monopoly Regulation: Principles and Practice*. Cambridge University Press, Cambridge, UK.
- Berg S.V., Tschirhart J. (1995). Contributions of neoclassical economics to public utility analysis. *Land Economics* 71(3), 310-30.
- Berrie T.W. (1992). *Electricity Economics and Planning*, 2nd edn., Peter Peregrinus, London.

-
- Bertram G. (1991a). Comalco and Manapouri. *Victoria Economic Commentaries* 8(2), 81-88.
- Bhattacharya S.C. (1995). Power sector privatization in developing countries: will it solve all problems? *Energy Sources* 17, 373-389.
- Black B. (1994). A proposal for implementing retail competition in the electricity industry. *Electricity Journal*, October, 58-72.
- Black B., Pierce R. (1993). The choice between markets and central planning in regulating the US electricity industry. *Columbia Law Review* 93(6), 1339-1441.
- Blaug M. (1980). Kuhn versus Lakatos or paradigms versus research programmes in the history of economics, in S. Latsis (ed.), *Method and Appraisal in Economics*, Cambridge University Press, Cambridge, UK.
- Blaug M. (1990). Marginal cost pricing: no empty box, in M. Blaug, *Economic Theories, True or False? Essays in the History and Methodology of Economics*, Edward Elgar Publishing, Vermont.
- Blaug M. (1992). *The Methodology of Economics: or How Economists Explain*, 2nd. edn. Cambridge Surveys of Economic Literature, Cambridge University Press, Cambridge, UK.
- Boiteux M. (1949). Peak-load pricing, translated and republished in J.R. Nelson (ed.) (1964), *Marginal Cost Pricing in Practice*, Prentice Hall, NJ.
- Boiteux M. (1956). Marginal cost pricing, translated and republished in J.R. Nelson (ed.) (1964), *Marginal Cost Pricing in Practice*, Prentice Hall, NJ.
- Boiteux M., Stasi P. (1952). The determination of costs of expansion of an interconnected system of production and distribution of electricity, translated and republished in J.R. Nelson (ed.) (1964), *Marginal Cost Pricing in Practice*, Prentice Hall, NJ.
- Bollard A.E., Pickford M. (1995). New Zealand's 'light-handed' approach to utility regulation. *Agenda* 2(4), 411-422.
- Bös D. (1986). *Public Enterprise Economics: Theory and Application*, North-Holland, Amsterdam.
- Bradley M.D., Colvin J., Panzar J.C. (1999). On setting prices and testing cross-subsidy with accounting data. *Journal of Regulatory Economics* 16, 83-100.
- Braeutigam R. (1980). An analysis of fully distributed cost pricing in regulated industries. *Bell Journal of Economics* 11, 182-196.
- Braeutigam R. (1983). A dynamic analysis of second-best pricing, in J. Finsinger (ed.), *Public Sector Economics*, Macmillan, London.
- Braeutigam R. (1989). Optimal policies for natural monopolies, in R. Schmalensee and R.D. Willig (eds), *Handbook of Industrial Organization*, North-Holland, Amsterdam.
- Braeutigam R., Panzar J. (1989). Diversification incentives under 'price-based' and 'cost-based' regulation. *Rand Journal of Economics* 20(3), 373-391.
- Brennan M., Schwartz E. (1982). Consistent regulatory policy under uncertainty. *Bell Journal of Economics* 13, 506-521.
- Broadman H., Kalt J. (1989). How natural is monopoly? The case for bypass in natural gas distribution markets. *Yale Journal of Regulation* 6, 181-208.
- Bronfenbrenner M. (1971). The 'structure of revolutions' in economic thought. *History of Political Economy* 3, 137-152.
- Brown G., Johnson M.B (1969). Public utility pricing and output under risk. *American Economic Review* 59, 119-128.
-

-
- Brown S.J., Sibley D.S. (1986). *The Theory of Public Utility Pricing*, Cambridge University Press, New York, NY.
- Burness H.S., Patrick R.H. (1992). Optimal depreciation, payments to capital and natural monopoly regulation. *Journal of Regulatory Economics* 4, 35-50.
- Cairns R.D., Mahabir D. (1988). Contestability: a revisionist view. *Economica* 55, 269-276.
- Calem P.S. (1988). Entry and entry deterrence in penetrable markets. *Economica* 55, 171-183.
- Canoy M. (1994). Natural monopoly and differential pricing. *Journal of Economics* 59(3), 287-309.
- Caramanis M.C., Bohn R.E., Schweppe F.C. (1982). Optimal spot pricing: practice and theory. *IEEE Transactions on Power Apparatus and Systems* 101(9), 3234-3245.
- Carlton D.W. (1977). Peak load pricing with a stochastic demand. *American Economic Review* 67(5), 1006-1010.
- Cave M., Doyle C. (1994). Access pricing in network utilities in theory and practice. *Utilities Policy* 4(3), 181-189.
- Caves D.W., Christensen L.R., Schoech P.E., Hendricks W. (1984). A comparison of different methodologies in a case study of residential time-of-use electricity pricing. Cost-benefit analysis. *Journal of Econometrics* 26, 17-34.
- Chamberlin J.H. (1992). Forecasting response to innovative rates, in C.W. Gellings (ed.), *Demand Forecasting for Electric Utilities*, Fairmont Press, Lilburn, GA.
- Chambers M.J., Gregg M.G. and Ling C.C. (1990). Time of use pricing and two-way communications technology, SEQEB moves towards home automation. *IEE 6th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Manchester, 21-25.
- Cicchetti C.J., Gillen W.J., Smolensky P. (1977). *The Marginal Cost and Pricing of Electricity: an Applied Approach*. Ballinger Publishing, Cambridge, MA.
- Claggett E.T., Hollas D.R., Stansell S.R. (1995). The effects of ownership form on profit maximisation and cost minimisation behavior within municipal and cooperative electrical distribution utilities. *Quarterly Review of Economics and Finance* 35(Special), 533-550.
- Coase R.H. (1937). The nature of the firm. *Economica* 4, 386-405.
- Cocklin C. (1993). Anatomy of a future energy crisis - restructuring and the energy sector in New Zealand. *Energy Policy* 21(Aug), 881-891.
- Cole B. (1993). To regulate or not to regulate – that is the question. *New Zealand Engineering* (Jun), 22-24.
- Copeland (1989). Commentary, in K. Nowotny (ed.), *Public Utility Regulation*, Boston.
- Crew M.A., Fernando C.S., Kleindorfer P.R. (1995). The theory of peak load pricing: a survey. *Journal of Regulatory Economics* 8, 215-248.
- Crew M.A., Kleindorfer P.R. (1975). On off-peak pricing: an alternative technological solution. *Kyklos* 28, 80-93.
- Crew M.A., Kleindorfer P.R. (1976). Peak load pricing with a diverse technology. *Bell Journal of Economics* 7, 207-231.
- Crew M.A., Kleindorfer P.R. (1979). *Public Utility Economics*. St. Martins Press, New York, NY.
- Crew M.A., Kleindorfer P.R. (1992). Economic depreciation and the regulated firm under competition and technological change. *Journal of Regulatory Economics* 4, 51-61.
- Culy J., Gale S. (1987). Regulatory change in the energy sector, in A. Bollard and R. Buckle (eds), *Economic Liberalisation in New Zealand*, Allen and Unwin, Wellington.

-
- Culy J.G., Read E.G., Wright B. (1997). Structure and regulation of the New Zealand electricity sector, in R. Gilbert and E. Kahn (eds), *International Comparison of Electricity Regulation*, Cambridge University Press, Cambridge, UK.
- Curien N. (1991). The theory and measure of cross-subsidies. An application to the telecommunications industry. *International Journal of Industrial Organization* 9, 73-108.
- Daryanian B., Bohn R.E. (1993). Sizing of electric thermal storage under real time pricing. *IEEE Transactions on Power Systems* 8(1), 35-43.
- Daryanian B., Bohn R.E., Tabors R.D. (1991a) An experiment in real time pricing for control of electric thermal storage systems. *IEEE Transactions on Power Systems* 6(4), 1356-1365.
- Daryanian B., Bohn R.E., Tabors R.D. (1991b). Control of electric thermal storage under real time pricing. *IEE International Conference on Advances in Power System Control, Operation and Management*, Hong Kong, 397-403.
- David A.K., Li Y.Z. (1991a). Electricity pricing with competitive supply conditions. *Electrical Power and Energy Systems* 13(2), 111-122.
- David A.K., Li Y.Z. (1991b). A comparison of system response for different types of real time pricing. *IEE International Conference on Advances in Power System Control, Operation and Management*, Hong Kong, 385-390.
- David A.K., Li Y.Z. (1991c). Consumer rationality assumptions in the real time pricing of electricity. *IEE International Conference on Advances in Power System Control, Operation and Management*, Hong Kong, 391-396.
- David A.K., Li Y.Z. (1993a). A rational approach for incorporating capital costs in the real-time pricing of electricity. *Electrical Power and Energy Systems* 15(3), 179-184.
- David A.K., Li Y.Z. (1993b). Effect of inter-temporal factors on the real time pricing of electricity. *IEEE Transactions on Power Systems* 8(1), 44-52.
- David A.K., Nutt D.J., Chang C.S., Lee Y.C. (1986). The variation of electricity prices in response to supply-demand conditions and devices for consumer interaction. *Electrical Power and Energy Systems* 8(2), 101-114.
- Davies E.G. (1987). The derivation and implementation of a circuit breaker tariff. *Proceedings of the IEE 5th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Edinburgh, 26-30.
- de Serpa A. (1988). *Microeconomic Theory: Issues and Applications*, 2nd edn. Allyn and Bacon, Inc., Boston, MA.
- Della Valle A.P. (1988). Short-run versus long-run marginal cost pricing. *Energy Economics* 10(Oct), 283-288.
- Demsetz H. (1968). Why regulate utilities? *Journal of Law and Economics* 11(Apr), 55-65.
- Dixit A.K. (1980). The role of investment in entry deterrence. *Economics Journal* 90, 95-106.
- Dixit A.K., Pindyck R.S. (1994). *Investment Under Uncertainty*, Princeton University Press, Princeton, NJ.
- Drayton G.R., Read E.G. (1996). Using LP to form a market for spinning reserve. *Proceedings of the Operational Research Society of New Zealand* 1996, 119-124.
- Economides N. (1996). The economics of networks. *International Journal of Industrial Organization* 14(6), 673-700.
- Edwards J., Kay J., Mayer C. (1987). *The Economic Analysis of Accounting Profitability*, Oxford University Press.

-
- Ekelund R.B, Hébert R.F. (1990). *A History of Economic Theory and Method*, 3rd edn., McGraw-Hill, New York, NY.
- Ely R.T. (1908). *Outlines of Economics*, revised edn., Macmillan, New York, NY.
- Farmer E.D., Cory B.J. (1987). Dynamic pricing in power system operation. *Proceedings of the IEE 5th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Edinburgh, 50-55.
- Farmer E.D., Cory B.J, Perera B.L (1995). Optimal pricing of transmission and distribution services in electricity supply. *IEE Proceedings-Generation, Transmission and Distribution* 142(1), 1-8.
- Faulhaber G.R. (1975). Cross-subsidization: pricing in public enterprises. *American Economic Review* 65(5), 966-977.
- Faulhaber G.R. (1979). Cross-subsidization in public enterprise pricing, in J. Wenders (ed.) *Pricing in Regulated Industries—Theory and Application II*, Mountain States Telephone and Telegraph.
- Faulhaber G.R., Baumol W.J. (1988). Economists as innovators: practical products of theoretical research. *Journal of Economic Literature* 26(Jun), 577-600.
- Faulhaber G.R., Levinson S.B. (1981). Subsidy-free prices and anonymous equity. *American Economic Review* 71(5), 1083-1091.
- Feinstein C., Morris P., Chapel S. (1997). Defining distributed resource planning. *The Energy Journal, Special Issue on Distributed Resources*, 41-62.
- Friedman M. (1953). The methodology of positive economics, in M. Friedman, *Essays in Positive Economics*, University of Chicago Press, Chicago.
- Giles D., Wyatt N. (1992). Economies of scale in the electricity distribution system, in P. Phillips (ed.), *Models, Methods, and Applications of Econometrics*, Blackwell, Cambridge, MA.
- Grout P. (1995). The cost of capital in regulated industries, in M. Bishop, J. Kay, C. Mayer (eds), *The Regulatory Challenge*, Oxford University Press, Oxford.
- Gunn C. (1997). Energy efficiency vs economic efficiency? New Zealand electricity sector reform in the context of the national energy policy objective. *Energy Policy* 25(4), 445-458.
- Gunn C. (1998). New Zealand's electricity sector: a decade of reform. *Power Economics* 2(1), 37-39.
- Gunn C., Sharp B. (1999). Electricity distribution as an unsustainable natural monopoly: a potential outcome of New Zealand's regulatory regime. *Energy Economics* 21, 385-401.
- Hartman R.S., Doane M.J., Woo C.K. (1991). Consumer rationality and the status quo. *Quarterly Journal of Economics* 106(1), 141-162.
- Hay D.A., Morris D.J. (1993). *Industrial Economics and Organization: Theory and Evidence*, 3rd ed. Oxford University Press, Oxford.
- Heald D. (1997). Public policy towards cross subsidy. *Annals of Public and Cooperative Economics* 68(4), 591-623.
- Hillier F., Lieberman G. (1986). Probabilistic dynamic programming, in *Introduction to Operations Research*, Holden-Day, Oakland, CA.
- Hobbs B., Schuler R. (1986). Deregulating the distribution of electricity: price and welfare consequences of spatial oligopoly with uniform delivered prices. *Journal of Regulatory Science* 26(2), 235-265.
- Hogan W.W., Ring B.J., Read E.G. (1996). Using mathematical programming for electricity spot pricing. *International Transactions in Operations Research* 3(3-4), 243-253.

-
- Hotelling H. (1925). A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20(Sep), 340-353.
- Jamison, M.A. (1996). General conditions for subsidy-free prices. *Journal of Economics and Business* 48, 371-385.
- Johnson M.B., Brown G. (1970). Public utility pricing and output under risk: reply. *American Economic Review* 60, 489-490
- Joskow P.L. (1976). Contributions to the theory of marginal cost pricing. *Bell Journal of Economics* 7, 197-206.
- Joskow P.L. (1978). Public Utility Regulatory Policy Act of 1978: electric utility rate reform. *Natural Resources Journal* 19(Oct), 787-810.
- Joskow P.L., Noll R.G. (1981). Regulation theory and practice: an overview, in G. Fromm (ed.), *Studies in Public Regulation*, MIT Press, Cambridge.
- Joskow P.L., Schmalensee R. (1983). *Markets for Power: an Analysis of Electric Utility Deregulation*, MIT Press, Cambridge, MA.
- Kahn A.E. (1970). *The Economics of Regulation: Principles and Institutions*, MIT Press, Cambridge, MA.
- Kay J.A. (1971). Recent contributions to the theory of marginal cost pricing: some comments". *The Economic Journal*, June, 366-371.
- Kaye R.J., Outhred H.R. (1989). A theory of electricity tariff design for optimal operation and investment. *IEEE Transactions on Power Systems* 4(2), 606-613.
- Kaye R.J., Outhred H.R., Bannister C.H. (1990). Forward contracts for the operation of an electricity industry under spot pricing. *IEEE Transactions on Power Systems* 5(1), 46-52.
- Kelsey J. (1993). *Rolling Back the State*, Bridget Williams Books, Wellington.
- King S.P., Maddock R. (1996). Competition and almost essential facilities: making the right policy choices. *Economic Papers* 15, 28-37.
- King S.P., Maddock R. (1998). Light-handed regulation of access in Australia: negotiation with arbitration. *Information Economics and Policy* 11, 1-22.
- Kirzner I.M. (1973). *Competition and Entrepreneurship*, University of Chicago Press, Chicago.
- Kirzner I.M. (1985). The perils of regulation-a market process approach, in I.M. Kirzner (ed.), *Discovery and the Capitalist Process*, University of Chicago Press, Chicago.
- Kirzner I.M. (1994). On 'The economics of time and ignorance', in P.J. Boettke and D.L. Prychitko (eds), *The Market Process: Essays in Contemporary Austrian Economics*, Edward Elgar, Aldershot, UK.
- Kleindorfer P.R., Fernando C.S. (1993). Peak load pricing and reliability under uncertainty. *Journal of Regulatory Economics* 5(1), 5-23.
- Klevorick A.K. (1971). The 'optimal' fair rate of return. *Bell Journal of Economics and Management Science* 2(1), 122-153.
- Kreps D.M., Scheinkman J.A. (1983). Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14, 326-337.
- Laffont J.J., Tirole J. (1993). *A Theory of Incentives in Procurement and Regulation*, MIT Press, Cambridge, MA.
- Laffont J.J., Tirole J. (1994). Access pricing and competition. *European Economic Review* 38, 1673-1710.
- Lesser J.A., Feinstein C.D. (1999). Electric utility restructuring, regulation of distribution utilities, and the fallacy of 'avoided cost' rules. *Journal of Regulatory Economics* 15, 93-110.
-

-
- Littlechild S.C. (1970a). A game theoretic approach to public utility pricing. *Western Economic Journal* 8(Jun), 162-166.
- Littlechild S.C. (1970b). Marginal-cost pricing with joint costs. *Economic Journal* 80(Jun), 323-335.
- Littlechild S.C. (1981). Misleading calculations of the social costs of monopoly power. *Economic Journal* 91, 348-363.
- Lockhart D., Mallon J. (1995). Electricity contracts in a de-regulated industry. *Current*, August, 17-19.
- Loube R. (1995). Price cap regulation: problems and solutions. *Land Economics* 71(3), 286-298.
- MacAvoy P., Spulber D., Stangle B. (1989). Is competitive entry free? Bypass and partial deregulation in natural gas markets. *Yale Journal of Regulation* 6, 209-247.
- McDonald J.R., Lo, K.L. (1990). Dynamic price structures and consumer load reaction. *IEE 6th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Manchester, 6-10.
- McDonald J.R., Whiting P.A., Lo K.L. (1994a). Spot-pricing: evaluation, simulation and modelling of dynamic tariff structures. *Electric Power and Energy Systems* 16(1), 23-34.
- McDonald J.R., Whiting P.A., Lo K.L. (1994b). Optimized reaction of large electrical consumers to spot-price tariffs. *Electric Power and Energy Systems*, 16(1), 35-48.
- McDonald R., Siegel D. (1986). The value of waiting to invest. *Quarterly Journal of Economics* 101, 707-728.
- Martzoukos S.H., Teplitz-Sembitzky W. (1992). Optimal timing of transmission line investments in the face of uncertain demand. *Energy Economics* 14, 3-10.
- Meyer J.R., Tye W.B. (1985). The regulatory transition. *American Economic Review Papers and Proceedings* 75, 46-56.
- Michaels R.J. (1989). Reorganizing electricity supply in New Zealand: lessons for the United States. *Contemporary Policy Issues* 7(Oct), 73-90.
- Miller E.S. (1995). Is the public utility concept obsolete? *Land Economics* 71(3), 273-285.
- Mirman L.J., Samet D., Tauman Y. (1983). An axiomatic approach to the allocation of a fixed cost through prices. *Bell Journal of Economics* 14, 139-151.
- Monts K. (1991). An empirical procedure for the temporal aggregation of electric utility marginal energy costs. *IEEE Transactions on Power Systems* 6(2), 658-661.
- Morgan M.G., Talukdar S.N. (1979). Electric load management: some technical, economic, regulatory and social issues. *Proceedings of the IEEE* 67(2), 241-313.
- Munasinghe M. (1990). *Electric Power Economics*, Butterworths, London.
- Munasinghe M., Sanghvi A. (1988). Reliability of electricity supply, outage costs and value of service: an overview. *Energy Journal* 9, 1-18.
- Munasinghe M., Warford J. (1982). *Electricity Pricing – Theory and Case Studies*, The World Bank, John Hopkins University Press, Baltimore.
- Murphy L. Kaye R.J., Wu F.F. (1994). Distributed spot pricing in radial distribution systems. *IEEE Transactions on Power Systems* 9(1), 311-317.
- Neal R.J., Friend J.F., Hall R.G. (1987) A short notice pricing experiment. *IEE 5th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Edinburgh, 45-49.
- Nelson J.P., Roberts M.J. (1989). Ramsey numbers and the role of competing interest groups in electricity regulation. *Quarterly Review of Economics and Business* 29(3), 21-42.
- Nelson J.R. (ed.) (1964). *Marginal Cost Pricing in Practice*. Prentice-Hall, Englewood Cliffs, NJ.
-

-
- Neuberg L.G. (1977). Two issues in the municipal ownership of electric power distribution systems. *Bell Journal of Economics* 8, 303-323.
- Neufeld J.L. (1987). Price discrimination and the adoption of the electricity demand charge. *Journal of Economic History* 47(3), 693-709.
- O'Driscoll G.P. Jr., Rizzo M.J. (1985). *The Economics of Time and Ignorance*, Basil Blackwell, Oxford, UK.
- Oi W.Y. (1967). The neoclassical foundations of progress functions. *Economic Journal* 77, 579-594.
- Oi, W.Y. (1971). A Disneyland dilemma: two-part tariffs for a Mickey Mouse monopoly. *Quarterly Journal of Economics* 85(1), 77-96.
- Outhred H.R., Bannister C.H., Kaye R.J., Lee Y.B, Sutanto D., Manimaran R. (1988). Electricity pricing. Optimal operation and investment by industrial consumers. *Energy Policy* 16, 384-393.
- Palmer K. (1991). Using an upper bound on stand-alone cost in tests of cross-subsidy. *Economics Letters* 35, 457-460.
- Panzar J. (1976). A neoclassical approach to peak load pricing. *Bell Journal of Economics* 7, 521-30.
- Panzar J. (1980). Sustainability, efficiency, and vertical integration, in B.M. Mitchell and P.R. Kleindorfer, *Regulated Industries and Public Enterprise: European and United States Perspectives*, Lexington Books, Lexington, MA.
- Panzar J., Willig R. (1977). Free entry and the sustainability of natural monopoly. *Bell Journal of Economics* 8, 1-22.
- Panzar J., Willig R. (1981). Sustainability analysis. Economies of scope. *American Economic Review* 71(2), 268-272.
- Park R.E., Acton J.P. (1984). Large business response to time-of-day electricity rates. *Journal of Econometrics* 26, 229-252.
- Park R.E., Weitzel D. (1984). Measuring the consumer welfare effects of time-differentiated electricity prices. *Journal of Econometrics* 26, 35-64.
- Parsons S.G. (1998). Cross-subsidization in telecommunications. *Journal of Regulatory Economics* 13, 157-182.
- Paulo S. (1992). The weighted average cost of capital: a caveat. *Engineering Economist* 37(2), 178-182.
- Peet J. (1992). The free market in energy: can it work out the cost? *New Zealand Engineering*, June, 24-26.
- Perloff J., Salop S. (1985). Equilibrium with product differentiation. *Review of Economic Studies* 52, 107-120.
- Perry M. (1984). Sustainable positive profit multiple-price strategies in contestable markets. *Journal of Economic Theory* 32, 246-265.
- Peyton Young H. (ed.) (1985a). *Cost Allocation: Methods, Principles, Applications*. Elsevier Science Publishers B.V., North Holland.
- Peyton Young H. (1985b). Producer incentives in cost allocation. *Econometrica* 53, 757-765.
- Phillips A. (1975). *Promoting Competition in Regulated Markets*, The Brookings Institution, Washington, DC.
- Pickford M. (1996). Pricing access to essential facilities. *Agenda* 3(2), 165-176.
- Primeaux W. (1975). A re-examination of the monopoly market structure for electric utilities, in A. Phillips (ed.), *Promoting Competition in Regulated Markets*, The Brookings Institution, Washington, DC.
- Read E.G. (1997). Transmission pricing in New Zealand. *Utilities Policy* 6(3), 227-235.
- Rees R., Vickers J. (1995). RPI-X price-cap regulation, in M. Bishop, J. Kay, and C. Mayer (eds), *The Regulatory Challenge*, Oxford University Press, Oxford.
-

-
- Renz B.A. (1993). Power engineering trends and challenges. *IEEE Power Engineering Review* 13(5), 20-24.
- Ring B.J. (1995). A dispatch based pricing model for the New Zealand electricity market, in R. Siddiqi and M. Einhorn, *Transmission Pricing and Access*, Kluwer, Boston.
- Roberts J., Elliot D., Houghton T. (1991). *Privatising Electricity - The Politics of Power*, Belhaven Press, London.
- Roberts M.J. (1986). Economies of density and size in the production and delivery of electric power. *Land Economics* 62(4), 378-387.
- Rogerson W.P. (1992). Optimal depreciation schedules for regulated utilities. *Journal of Regulatory Economics* 4, 5-33.
- Routh G. (1975). *The Origin of Economic Ideas*. Macmillan, London.
- Salinger M. (1998). Regulated prices to equal forward-looking cost: cost-based prices or price-based costs? *Journal of Regulatory Economics* 14(2), 149-164.
- Salvanes K.G., Tjotta S. (1994). Productivity differences in multiple output industries: an application to electricity distribution. *Journal of Productivity Analysis* 5(1), 23-43.
- Sassower R. (1988). Ideology masked as science: shielding economics from criticism. *Journal of Economic Issues* 22(1), 167-179.
- Scherer C.R. (1977). *Estimating Electric Power System Marginal Costs*, North-Holland, Amsterdam.
- Schmalensee R. (1981). Monopolistic two-part pricing arrangements. *Bell Journal of Economics* 12, 445-466.
- Schmalensee R. (1989). An expository note on depreciation and profitability under a rate-of-return constraint. *Journal of Regulatory Economics* 1, 293-298.
- Schramm G. (1991). Marginal cost pricing revisited. *Energy Economics* 13(Oct), 245-249.
- Schumpeter J.A. (1950). *Capitalism, Socialism and Democracy*, 3rd edn., Harper and Row, New York.
- Schwartz P. (1971). *The New Political Economy of J.S. Mill*, 1st English edn., Weidenfeld and Nicolson, London.
- Schwepe F.C. (1978). Power systems 2000: hierarchical control strategies. *IEEE Spectrum* 15(7), 42-47.
- Schwepe F.C., Caramanis M.C., Tabors R.D., Bohn R.E. (1988). *Spot Pricing of Electricity*. Kluwer Academic Publishers, Boston.
- Schwepe F.C., Tabors R.D., Kirtley J.L., Outhred H.R., Pickel F.H., Cox A.J. (1980). Homeostatic utility control. *IEEE Transactions on Power Apparatus and Systems* 99(3), 1151-1163.
- Scott T.J., Read E.G. (1996). Modelling hydro reservoir operation in a deregulated electricity sector. *International Transactions in Operations Research* 3(3-4), 209-221.
- Sharkey W.W. (1982a). *The Theory of Natural Monopoly*. Cambridge University Press, New York.
- Sharkey W.W. (1982b). Suggestions for a game theoretic approach to public utility pricing and cost allocation. *Bell Journal of Economics and Management Science* 13, 57-68.
- Sharkey W.W., Telser L.G. (1978). Supportable cost functions for the multiproduct firm. *Journal of Economic Theory* 18, 23-37.
- Sheen J.N, Chen C.S., Yang J.K (1990). Time-of-use pricing for load management programs in Taiwan Power Company. *IEEE Transactions on Power Systems* 9(1), 388-395.
- Shepherd, W.G. (1984). 'Contestability' vs. competition. *American Economic Review* 74(4), 572-587.
- Shepherd, W.G. (1995). Contestability vs. competition—once more. *Land Economics* 71(3), 299-309.
- Shleifer A. (1985). A theory of yardstick competition. *Rand Journal of Economics* 16, 319-327.
-

-
- Sidak J.G., Spulber D.F. (1997). *Deregulatory Takings and the Regulatory Contract. The Competitive Transformation of Network Industries in the United States*. Cambridge University Press, New York.
- Sorenson J., Tschirhart J. and Whinston A. (1976). A game theoretic approach to peak load pricing. *Bell Journal of Economics* 7, 497-520.
- Spence A.M. (1977). Entry, capacity, investment and oligopoly pricing. *Bell Journal of Economics* 8, 534-544.
- Spence A.M. (1984). Contestable markets and the theory of industry structure: a review article. *Journal of Economic Literature* 21, 981-990.
- Spicer B., Bowman R., Emanuel D., Hunt A. (1991). *The Power to Manage: Restructuring the New Zealand Electricity Department as a State-Owned Enterprise - The Electricorp Experience*, Oxford University Press, Auckland.
- Spulber D.F. (1984). Scale economies and the existence of sustainable monopoly prices. *Journal of Economic Theory* 34, 149-163.
- Spulber D.F. (1989). *Regulation and Markets*. MIT Press, Cambridge, MA
- Spulber D.F. (1993). Monopoly pricing. *Journal of Economic Theory* 59, 222-234.
- Steiner P.O. (1957). Peak loads and efficient pricing. *Quarterly Journal of Economics* (Nov), 585-610.
- Stern P.C. (1986). Blind spots in policy analysis: what economics doesn't say about energy use. *Journal of Policy Analysis and Management* 5(2), 200-227.
- Stigler G.J. (1939). Production and distribution in the short run. *Journal of Political Economy* 47, 305-327.
- Stigler G.J. (1968). *The Organization of Industry*, Richard D. Irwin, Homewood, IL.
- Stigler G.J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science* 2, 3-19.
- Stiglitz J.E. (1987). Technological change, sunk costs, and competition. *Brookings Papers on Economic Activity*, 883-937.
- Stiglitz J.E. (1993). *Economics*, W.W. Norton, New York, NY.
- Stokey N. (1979). Intertemporal price discrimination. *Quarterly Journal of Economics* 93, 355-371.
- Tabors R. (1994). Transmission system management and pricing: new paradigms and international comparisons. *IEEE Transactions on Power Systems* 9(1), 206-215.
- Taggart R. (1985). Effects of regulation on utility financing: theory and evidence. *Journal of Industrial Economics* 22, 257-276.
- Teisberg E.O. (1993). Capital investment strategies under uncertain regulation. *Rand Journal of Economics* 24(4), 591-604.
- Tirole J. (1988). *The Theory of Industrial Organisation*. MIT Press, Cambridge, MA.
- Trebing H.M. (1984). Public control of enterprise: neoclassical assault and neoinstitutional reform. *Journal of Economic Issues* 18(Jun), 353-368.
- Trebing H.M. (1987). Regulation of industry: an institutionalist approach. *Journal of Economic Issues* 21(4), 1707-1737.
- Trebing H.M. (2000). A review essay on 'Deregulatory Takings and the Regulatory Contract'. *Telecommunications Policy Online* 24(2).
- Tromop R.W., White G.A., Gunn C.I (1996). Demand side management for the electricity industry. *Transactions of the Institution of Professional Engineers New Zealand* 23(1/Gen), 12-20.
-

-
- Turvey R. (1968). *Optimal Pricing and Investment in Electricity Supply: an Essay in Applied Welfare Economics*, Allen and Unwin, London.
- Turvey R. (1969). Marginal cost. *Economic Journal* 79(Jun), 282-299.
- Turvey R. (1970). Public utility pricing and output under risk: a comment. *American Economic Review* 60, 485-486.
- Turvey R. (1971). *Economic Analysis and Public Enterprise*, Allen and Unwin, London.
- Turvey R., Anderson D. (1975). *Electricity Economics*, John Hopkins University Press, Baltimore, MA.
- Turvey R., Anderson D. (1977). *Electricity Economics: Essays and Case Studies*. The World Bank, John Hopkins University Press, Baltimore, MA.
- Uusitalo J., Yrjölä P. (1990). Research on real-time pricing of electricity. *IEE 6th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, University of Manchester, 48-51.
- Vickers J., Yarrow G. (1988). *Privatization: an Economic Analysis*, MIT Press, Cambridge, MA.
- Vickrey W.S. (1955). Some implications of marginal cost pricing for public utilities. *American Economic Review* 4, 114.
- Vickrey W.S. (1971). Responsive pricing of public utility services. *Bell Journal of Economics and Management Science* 2(1), 337-346.
- Vogelsang I. (1994). Profit sharing regulation of electrical transmission and distribution companies, in M.A. Einhorn (ed), *From Regulation to Competition: New Frontiers in Electricity Markets*, Kluwer Academic Publishers, Boston, MA.
- Weinberg C.J., Iannucci J.J., Reading M.M. (1993). The distributed utility: technology, customer, and public policy changes shaping the electric utility of tomorrow. *Energy Systems and Policy* 15, 307-322.
- Weintraub S. (1970). On off-peak pricing: an alternative solution. *Kyklos* 23(3), 501-517.
- Weisman D.L. (1991). A note on first-best marginal cost measures in public enterprise. *Energy Economics* 13(Oct), 250-253.
- Weiss L.W. (1975). Antitrust in the electric power industry, in A. Phillips (ed), *Promoting Competition in Regulated Markets*, The Brookings Institution, Washington, DC.
- Wenders J.T. (1976). Peak load pricing in the electric utility industry. *Bell Journal of Economics* 7, 232-241.
- Wenders J.T., Taylor L.D. (1976). Experiments in seasonal time-of-day pricing of electricity to residential users. *Bell Journal of Economics* 7, 531-552.
- Weyman-Jones T.G. (1990). RPI-X price cap regulation. *Utilities Policy* 1(Oct), 65-77.
- Weyman-Jones T.G. (1995). Problems of yardstick regulation in electricity distribution, in M. Bishop, J. Kay, C. Mayer (eds), *The Regulatory Challenge*, Oxford University Press, Oxford, UK.
- Williamson O.E. (1966). Peak load pricing and optimal capacity under indivisibility constraints. *American Economic Review* 56, 810-827.
- Williamson O.E. (1986a). *The Economic Institutions of Capitalism*, Free Press, New York, NY.
- Williamson O.E. (1986b). The economics of governance: framework and implications, in R. Langlois (ed.), *Economics as a Process – Essays in New Institutional Economics*, Cambridge University Press, Cambridge, UK.
- Willig R.D. (1978). Pareto-superior nonlinear outlay schedules. *Bell Journal of Economics* 9, 56-69.
- Willig R.D. (1979b). The theory of network access pricing, in H.M. Trebing (ed), *Issues in Public Utility Regulation*, Michigan State University Public Utility Papers, East Lansing, MI.
-

-
- Willis H.L., Tram H., Engel M.V., Finley L. (1995). Optimization applications to power distribution. *IEEE Computer Applications in Power*, October, 12-17.
- Wirl F. (1991). Needle peaking caused by time of day tariffs. *Electrical Power and Energy Systems* 13(3), 175-181.
- Wiseman J. (1957). The theory of public utility pricing: an empty box, reprinted in J. Buchanan and G.F. Thirlby (eds) (1973), *LSE Essays on Cost*, Weidenfeld and Nicolson, London.
- Witteloostuijn A. van (1990). Contestability and investment in average cost reduction. *European Journal of Political Economy* 6, 23-40.
- Woo C-K., Lloyd-Zannetti D., Orans R., Horii B., Heffner G. (1995). Marginal capacity costs of electricity distribution and demand for distributed generation. *The Energy Journal* 16(12), 111-130.
- York D. (1994). Competitive electricity markets in practice: experience from Norway. *The Electricity Journal*, June.
- Zajac E.E. (1985). Perceived economic justice: the example of public utility regulation, in H. Peyton Young (ed), *Cost Allocation: Methods, Principles, Applications*, Elsevier Science, North-Holland.