# 4 A framework for benchmarking

|  |
| --- |
| Key points |
| * Regulatory benchmarking encompasses any method for comparing firms to each other, to themselves over time, or to an ideal firm, in order to measure, and (potentially) encourage, efficiency in the regulated business. * The literature on benchmarking is confused. There are: * multiple methods for benchmarking, with little consensus about which is best * divergent views about the appropriate inputs and outputs of electricity network businesses. * Nevertheless, in Australia and overseas, network regulators have made extensive use of benchmarking, mainly as a tool for assisting their regulatory judgments. * In Australia, the more sophisticated use of benchmarking in regulating the National Electricity Market has been frustrated by inadequate data (and potentially by limitations in the National Electricity Rules) and limited sample sizes. * International comparisons can only partially address these limited sample sizes, and create their own challenges in reaching valid comparisons of business performance. * Incentive regulation aims to improve managerial performance and business efficiency. Accordingly, benchmarking needs to take into account factors outside the control of the businesses. * Such an exercise is useful in its own right, as it may help identify policies or regulations that stifle efficiency. * The path to better benchmarking depends on specifying clear criteria for appropriate benchmarking and systematically applying them. This needs to recognise that the degree of rigour required is dependent on the extent to which benchmarking is used to determine the regulatory outcomes for businesses. * Benchmarking is not about identifying a single number denoting the efficiency of a business, but rather the potential range of likely numbers. * Any benchmarking exercise must take into account the consequences of being wrong. |
|  |
|  |

This chapter is one of five relating to benchmarking of business efficiency. It:

* introduces the concept. Despite its worldwide use, there are many complexities in defining what benchmarking is and its purposes (section 4.1)
* discusses briefly the different methods that are routinely used to undertake benchmarking, and their major technical pitfalls and advantages (section 4.2)[[1]](#footnote-1)
* considers the appropriate measures of network performance and the cost drivers the businesses face (section 4.3)
* reviews the historical use of benchmarking by Australian regulators, including the Australian Energy Regulator (AER) (section 4.4). This section also describes the relevant clauses in the Rules regarding benchmarking. However, whether in fact the AER can actually use benchmarking effectively depends on interpretation of other facets of the Rules — an issue addressed in the next chapter
* discusses the appropriate criteria for the coherent evaluation of benchmarking approaches and the methods used to test these criteria (sections 4.5‑4.8). The question of appropriate benchmarking *processes* is distinct, and is addressed in chapter 8.

## 4.1 Benchmarking managerial efficiency and performance

Regulatory benchmarking encompasses any method for comparing a firm to other businesses, to itself over time (or between its various divisions) or to an ideal firm. Utility regulators around the world use static (and dynamic) ‘benchmarking’ to encourage regulated businesses to achieve the long-run efficiency outcomes of decentralised, workably competitive, markets. Benchmarking is a well-established method for analysing network businesses’ performance, and has been used by:

* Australian regulators, including state based electricity regulators (Meyrick and Associates 2005, Pacific Economics 2008)
* international regulators such as OFGEM (United Kingdom), CER (Ireland), NZCC (New Zealand), and OEB (Ontario Canada)
* various academics in the Australian, European, American and other contexts (see AMP Capital, sub. DR55, Chanel 2008, Farsi and Filipini 2005, London Economics 1999, Mountain and Littlechild 2010).

Figure 4.1 depicts the broad types of comparisons underpinning benchmarking. In the left hand chart, a group of utilities have a distribution of efficiency from poor to best (the ‘frontier’). The challenge of benchmarking is to estimate the distribution and make a judgment about some level of acceptable efficiency — the ‘benchmark’. It could be at the frontier — but as explained in chapter 5, there are grounds for an incentive gap to provide a reward for dynamic efficiency and to address regulatory error — an issue touched on later and in chapter 8.

Sometimes, benchmarking will relate to a categorical, rather than a continuous measure of performance (or proxy for it). The example in the right hand chart is simply of the yes/no kind, but there could be more than two categories. The surveys used to measure management performance shown in figure 4.1 are based on Likert scales from one to five (Bloom and van Reenen 2010, p. 206).

Figure 4.1 Benchmarkinga

|  |
| --- |
| density  sub-optimal  A  yes  frequency  no  benchmark  Most  efficient  benchmark  **Continuous measure of performance**  (for example, labour productivity)  **Categorical measure of performance**  (for example, does the business use  internal benchmarking?)  actual  B  E  Under-  performance  The incentive gap |

a The distribution relates to the measured productivity outcomes for a group of businesses from some benchmarking exercise, with A denoting the measurement for a given business. A is a point estimate — it does not show the degree to which the measured efficiency of A or B is reliable.

*Data source*: Bloom and van Reenen (2010).

### What does efficiency mean and how is it relevant to incentive regulation?

Since benchmarking aspires to increase efficiency, it is critical to unpack this concept. Efficiency has three main dimensions, and relates to the extent to which a business has:

* achieved the maximum possible output from their given inputs (productive or ‘technical’ efficiency). A simple example would be the labour productivity of vegetation clearance around transmission towers (in a common terrain)
* allocated resources to their highest value purposes (‘allocative’ efficiency). The extent to which a business achieves this depends on whether it prices its outputs efficiently (for example, with no cross-subsidies) and uses input mixes that truly reflect their relative prices. The current prices paid for power and capacity from photovoltaic cells provides a good illustration. The physical energy outputs of photovoltaic panels is maximised with time invariant feed-in tariffs (an example of productive efficiency). But the *value* of the output (largely derived as network cost savings and as reduced needs for costly peaking generators) is highest at peak use times (chapter 13). In the case of monopoly network businesses, allocative efficiency particularly relates to the potential for businesses to charge prices well above the total costs of output with adverse impacts on businesses using inputs, consumer surplus and the optimum amount of output (chapter 3). In the regulated case, allocative inefficiency is most likely the outcome of inefficient mixes of capital and operating expenditures
* maximised the potential for increasing efficiency over time (dynamic efficiency). For example, this could occur through product and process innovation; investment in management and labour skills that allow flexible responses to changing economic circumstances, and exposure to the risks of insolvency and downsizing.

While all three efficiency measures are important, most benchmarking analysis focuses on the first aspect of efficiency, because it is tractable, and can be used in so-called incentive regulation to encourage the other forms of efficiency. Chapter 5 examines incentive regulation in more detail,[[2]](#footnote-2) but the thrust of the idea is that the regulator:

* calculates some measure of the ‘efficient’ cost of producing the output of the business using benchmarking, controlling factors outside the influence of the business, so as to target managerial inefficiency
* allows the business to set prices to recover those efficient costs (including a return on capital).

Under this form of incentive regulation, there is no direct link between the business’s actual costs and the prices it charges, encouraging businesses to minimise their costs in order to maximise their profits. If the regulator sets a benchmark close to the attainable technical efficiency level, the business has little prospect of setting excessive prices (thus encouraging allocative efficiency), technically inefficient businesses exit or learn quickly (achieving technical efficiency), while businesses that aggressively reduce costs through better management or innovation make higher than ‘normal’ profits (dynamic efficiency). This ideal is not a feasible option, but at least it illustrates the principle.

Many of the network businesses covered in this inquiry have queried the suggestion that they are inefficient in any substantive way. Were that true, the main goal of benchmarking would be to eliminate excessive rents (allocative inefficiency) over time, rather than to penalise technical inefficiency. This would suggest no gain from benchmarking differences in levels of productivity between various businesses. Rather, it would imply that a regulator reimburse the business’s existing costs in a base year and then adjust allowed revenues over succeeding periods, using an estimate of the rate of total factor productivity (TFP) change, based on historical data, evolving network industry productivity trends, or another unregulated industry. (This is often referred to as the CPI-x or TFP methodology, where x is the relevant TFP growth rate — AEMC 2011b).[[3]](#footnote-3) The Australian Energy Market Commission (AEMC) has carefully reviewed the various disadvantages and (many) advantages of this approach. Its largest advantages are its apparent relative simplicity, the strong incentives it creates for dynamic efficiency and its amenability to being used in negotiated settlements (Cunningham, sub. 28; Kaufmann 2006; Pacific Economics Group, sub. 35 and sub. DR48).

Nevertheless, the CPI-x approach has several drawbacks. First, its simplicity may be more apparent than real. As Bruce Mountain (representing the ESAA) noted:

I’ve just seen it degenerate into the most arcane arguments of definitions and inputs and outputs (trans. p. 109).

While a strong advocate for the CPI-x approach, Paul Fearon (at the time, the chief executive of the Victorian Essential Services Commission), acknowledged that:

Introducing TFP methodologies does present some operational and transitional challenges. In the United States, pure TFP based price caps … are unusual and have been modified by regulators. For example, they have introduced off-ramps, earnings and revenue sharing mechanisms and ‘z’ and stretch factors, to deal with differing circumstances in relation to risk, firm performance and other uncontrollable exogenous factors (e.g. environmental factors). (Fearon 2007, p. 10)

Second, regulators have more often applied CPI-x to opex than capex. As Joskow (2008, pp. 553‑54) observes, there are many difficulties in operationalising a CPI-x approach that incorporates the complexities of investment and the heterogeneity of the businesses (for example, in relation to the vintage of their assets):

The limited attention paid to capital-related costs in the academic literature on price cap regulation provides a potentially misleading picture of the challenges associated with implementing a price-cap mechanism effectively. … Thus, the implementation of price cap mechanisms is more complicated and their efficiency properties more difficult to evaluate than is often implied and places a significant information collection, auditing and analysis burden on regulators.

It is notable that transmission capex is particularly lumpy at the firm level. Any incentive mechanism based on industry-wide productivity growth rates would need to ensure sufficient revenue flows to underpin such investments. In particular, unlike electricity distribution, the inherent reliability of a transmission network may not be observed until a major system failure, suggesting the requirement for some regulatory oversight of required investment (chapter 16).

Finally, and probably most importantly, the assumption of efficiency in the base year is a strong one, in which case, some kind of return to benchmarking of the level of efficiency re-emerges. As noted by the Major Energy Users (MEU):

TFP is a form of benchmarking which does result in less prescription but is dependent on the initial allowance being efficient first. If the initial allowance is not efficient, TFP provides for the inefficiency to be perpetuated. Therefore benchmarking is seen as the essential first step to reach the efficient frontier. (sub. 11, p. 30)

If the assumption of efficiency is not correct, the regulator should re-adjust base year costs to the efficient level before applying the conventional CPI-x approach. Without such an adjustment, businesses could either still survive while remaining inefficient, or make permanent profits above the normal return by closing the efficiency gap.[[4]](#footnote-4) Neither would be desirable. For example, were the total industry costs around $15 billion in a base year, but efficient aggregate costs were 10 per cent lower, then under reasonable assumptions, using the TFP methodology without base year adjustment would result in one of three outcomes:

1. At worst, if businesses’ average level of productivity did not catch-up to the frontier, the economy would forgo an efficiency dividend of around $22 billion in net present value terms.[[5]](#footnote-5)
2. Alternatively, if the businesses quickly reached the efficient frontier, they would earn rents of $22 billion — a transfer from customers to the network businesses. This would involve lower efficiency costs (although higher prices would still have some inefficiency impacts — chapter 3).
3. A combination of the two could occur.

Regardless, the incorrect assumption of efficiency in the base year would be equivalent to foregoing the opportunity for customers to receive free electricity network services for one and half years.

The hypothesis that network businesses are currently efficient would be an astonishing result on several grounds.

* As discussed in chapter 3, incentives for cost efficiency are blunted when businesses are not exposed to fierce actual or potential competition and have legacy ‘cultures’ reflecting their past status as effectively non-corporatised government departments.
* The usual factors encouraging dynamic efficiency are, at best, weak in an industry, where, even *after* regulation:
* businesses cannot be allowed to become insolvent (with the disruption costs and difficulties of ensuring continuity of service quality were that to occur)
* many businesses, through their state-owned status, cannot ever be taken over in hostile takeovers or merge across state boundaries
* revenue allowances have tended to reward over-capitalisation
* network businesses are free from normal market competition given their monopoly status, and through the regulatory arrangements, have greater security about returns on their sunk investments than would occur in unregulated markets. Lower levels of competition are generally associated with reduced efficiency (Bloom and Van Reenen 2010)
* the industry is characterised by large lumpy investments, which makes businesses susceptible to long-lived efficiency impacts from poor decisions. The existing international empirical evidence for highly capital-intensive industries — such as electricity, airlines and telecommunications — suggests that wide divergences in efficiency are unexceptional, notwithstanding the sophisticated nature of the businesses concerned.

Even in competitive markets, there is a continuum of efficiency and performance (Bloom and van Reenan 2010; Syverson 2011). This is demonstrated by figure 4.2 for a large sample of Australian businesses operating in typical market conditions. There is a long tail of poorly performing businesses. A similar dispersion of performance is apparent in similar studies using the same methods over a large number of countries.[[6]](#footnote-6)

Figure 4.2 Management performance of Australian businesses generally

2009

|  |
| --- |
| 1  2  3  4  5  0.2  0.4  0.6  0.8  1.0  Distribution of overall management performance of businesses (score 1-5)  Density (or relative frequency) |

a The data relate to a mixture of business ownership types, sizes and sectors (though mainly in sectors, such as manufacturing, where competitive forces are high). The data were collected as part of a collaborative performance study overseen by the London School of Economics, Stanford University. McKinsey & Company. A higher score represents better performance.

*Data source*: Macquarie Graduate School of Management et al. (2009, p. 30).

And regardless of the sample size and data deficiencies that might limit the usefulness of benchmarking analysis of Australian network businesses, most overseas studies of electricity network businesses find many are not even close to the efficiency frontier (figure 4.3 presents an example). For example, putting aside the sometimes disputed Australian results of Mountain (2011) and Mountain and Littlechild (2010), a large sample study by Chanel (2008, p. 7) found that the least efficient distribution businesses in the European Union had costs more than 80 per cent higher than the most efficient. It is hard to argue that Australian network businesses are exempt from what appears to be a ubiquitous pattern among their international peers (and businesses generally).

The key question for benchmarking is not whether there is inefficiency, but whether there is enough to matter for regulatory purposes. One of the attractions of privatisation is that it may strip away enough inefficiency that simpler benchmarking methods — such as conventional CPI-x — might become more realistic options (chapter 7 and 8).

Figure 4.3 Technical efficiency in electricity distribution in the United Kingdoma

|  |
| --- |
|  |

a For both axes in the chart, a business with a score of one is at the efficiency frontier for the relevant measure of expenditure, while scores below this indicate the degree of inefficiency. For example, firm A is 37 per cent below best practice totex and 44 per cent below best practice opex. In principle, and considering totex, firm A could increase its output by around 59 per cent (that is 1/0.63 x 100), without any changes in its inputs.

*Data source*: Yu et al. (2006).

## 4.2 Benchmarking techniques

Aside from engineering models, the methods used to benchmark electricity networks are similar to productivity analysis of any industry, with the major difference being the relevant variables (figure 4.4).

There are four basic approaches to benchmarking:

* statistical methods provide estimates of parameters of a production or cost function, but with information about the imprecision of those parameters

Figure 4.4 Different benchmarking approachesa

|  |
| --- |
| Figure 4.4 Different benchmarking approaches. This figure shows different benchmarking methods and outputs, such as simple indicators, ordinary least squares, data envelopment methods and stochastic frontier measures. |

a OLS is a regression approach that minimises the sum of squares of the errors of a line passing through the data. The dependent variable could be in log form or some other transformation. (While not shown above, non-linear least squares — a variant of OLS — does not impose linearity.) Corrected OLS shifts the line up just until there is one observation with a measured efficiency index of one. Stochastic frontier analysis takes account of the fact that errors around the regression line comprise statistical noise and systematic (one-sided) inefficiency. Structural time series models take account of the fact that parameters may shift over time (for example due to structural shifts or slowly declining or increasing productivity growth). Non-parametric methods are simple ratios (akin to those used in chapter 6). Total factor productivity (TFP) indexes are based on weighted inputs and outputs over time. Data envelopment analysis is a linear programming approach that connects the outer envelope of productively efficient businesses, while stochastic DEA takes account of the influence of statistical noise. The approaches can apply to a snapshot of businesses at a given time, time series for specific businesses, or panel data methods that use cross-sectional and time series data.

* non-parametric methods do not make any assumptions about the distribution of the population. Such approaches include simple ratios, indexes and linear programming methods
* hybrid methods combine non-parametric and parametric methods (Daraio 2012)
* engineering and reference models are bottom‑up models, based on expert knowledge about the operation of networks and the efficient costs of building and operating them. These models (such as the Swedish NPAM model) — create an artificial firm based on engineering and cost information to use as a benchmark, and then feeds in the characteristics of a given network business — such as its customer numbers, location, and capacity for each transformer station (Jamasb and Pollitt 2007a; Strbac and Allan 2001). In Australia, Elder and Beardow (2003) developed a model of an idealised electricity distribution network for regulatory pricing (among other reasons), but it is not clear it has been used for that purpose.

Figures 4.5 and 4.6 depict graphically how the various methods measure efficiency.

The literature on the various methods is vast. This inquiry does not examine these methods in any detail since there are multiple textbooks, software tools and review papers in this area. The AER has been responsible for two of the most comprehensive and useful papers.[[7]](#footnote-7)

Notwithstanding the volume of research and empirical work, there is no consensus about the best benchmarking measures (or even the appropriate inputs and outputs — section 4.3). In 2001, two leading economists in the area effectively said they did not know what approach to advocate:

… there is no ﬁrm consensus on how the basic functions of the utilities are to be modelled. (Jamasb and Pollitt 2001, p. 114)

That view did not change much in the ensuing decade. For example, in their consideration of the issue, Growitsch et al. (2010, p. 3) reiterated:

Although, from a technical point of view, distribution networks can be regarded as relatively simple activities, there is no consensus in the academic literature or among the regulatory practitioners as how to model this activity.

In the United Kingdom, Frontier Economics (2010b, p. 6) recommended ordinary least squares (OLS) or corrected OLS (COLS), but not stochastic frontier analysis and data envelopment analysis (DEA). Other analysts suggest that DEA is more robust and makes fewer underlying assumptions. Yet again, others regard the TFP index approach as the most simple because it uses overall industry productivity growth as the basis for incentive regulation, without reliance on the specific performance of the business.[[8]](#footnote-8)

The various preferences of regulators are revealed by significant variations in benchmarking practices around the world (table 4.1). Even in countries where benchmarking has been routinely applied, regulators of different utilities have tended to adopt different approaches, without any sign of convergence. As one analyst in this area noted in the United Kingdom context:

Moreover, it is not obvious that over time there has been a movement towards some form of consensus across the regulatory ofﬁces on the role of benchmarking. (Dassler et al. 2006, p. 172)

Figure 4.5 Benchmarking methods

|  |
| --- |
| Figure 4.5 Benchmarking methods. This figure shows the different measures of performance, based on a set of data on the costs and output of different businesses. Some productivity measures depend on just drawing a trend line through the points. Others use the most efficient business as the benchmark. |

*Source*: CREG & SUMICSID (2011, p. 26).

Figure 4.6 Illustrating the measures of inefficiencya

|  |
| --- |
| Figure 4.6 Illustrating the measures of inefficiency. This figure is an example of how efficiency is measured by comparing the varying costs of achieving a given output. |

a An input-oriented measure of inefficiency is the reduction in inputs for a given output level that a firm could achieve if it became efficient. So, firm A produces output Y1 at cost C1. Using the linear regression approach, it could produce Y1 at cost C2, so that its current inefficiency is C2/C1. Using data envelopment analysis, it could produce Y1 at cost C3, so that its inefficiency is C3/C1. Similar measures can be obtained by using the same approaches for the other estimation methods shown in the previous chart.

Table 4.1 What do other countries do?

Network benchmarking around the world

|  |  |  |
| --- | --- | --- |
| Location | Method | Source |
| Chile | Engineering | Farsi et al. 2007 |
| Denmark | Ratios | Nordic Energy Regulators 2011 |
| Finland | DEA/SFA/SDEA | Nordic Energy Regulators 2011 |
| Netherlands | DEA/Ratios | Farsi et al. 2007 |
| Norway | DEA | Nordic Energy Regulators 2011 |
| Sweden | DEA/Engineering | Jamasb and Pollitt 2007b |
| Spain | Engineering | Schweinsberg et al. 2011 |
| Austria | COLS/DEA | Schweinsberg et al. 2011 |
| Ireland | COLS | ACCC/AER 2012b |
| New Zealanda | TFP index | ACCC/AER 2012b |
| Ontario | TFP, ratios | ACCC/AER 2012b |
| Japan | Economtric | ACCC/AER 2012b |
| California | TFP, econometric | ACCC/AER 2012b |

a The New Zealand Commerce Commission (2010, pp. 640‑45) has recently proposed a novel application of benchmarking. Independent verifiers, effectively external auditors (paid for by the businesses), must provide independent evidence in support of the business’s proposal — and are free to use a range of quantitative techniques, including benchmarking.

In this context, the AER’s verdict about the state of the technology is not surprising:

There is no clear consensus in the literature in relation to which benchmarking approach should be used by economic regulators. As identified previously, each method has relative strengths and weaknesses. (ACCC/AER 2012a, p. 136)

## 4.3 What should be benchmarked?

In many industries, it is clear what inputs and outputs are. This is not so straightforward in electricity networks (and other networks, such as rail[[9]](#footnote-9) and telecommunications). In part, this is because electricity networks are very complex. For instance, there may be discontinuities in cost functions and counterintuitive relationships:

… the effects of Kirchhoff’s laws lead to bewildering irregularities in the relationship between outputs and capacities … As a result, transmission cost functions are likely to have strange properties that make them interesting for an audience outside electricity. Where else can you expect to have negative marginal costs? (Rosellón et al. 2009, p. 1)

These esoteric complexities aside, a comprehensive analysis of benchmarking prior to 2001 found no unanimity about the appropriate variables to include (Jamasb and Pollitt 2001, p. 114). Given the ACCC/AER’s (2012a and b) recent literature review, this situation has persisted. As Turvey (2008a, p. 2) put it:

Comparisons between networks of the costs of these activities can only illuminate differences in the efficiency with which operations and maintenance are carried out if the magnitudes of the tasks of operation and maintenance can be compared. This is a platitude, yet failure to articulate it has led some authors to scrabble around among available data to select a set of “explanatory” variables without displaying any understanding of what an enterprise does and how it does it. Confusion about these matters is rife, as witnessed, for example, by the fact that while some econometric efficiency estimates for electricity distribution treat MWh distributed, km of overhead lines or number of customers as an input, others treat one or more of these variables as an output!

As an illustration, while some consider that energy transfer across networks is an output, in fact, network businesses do not determine how much power is transferred across their lines (in contrast with generators). The APA Group (sub. 2, pp. 1‑2) observed:

The benchmarking of electricity business’ productivity (and TFP benchmarking in particular) almost always uses energy delivered as the output measure … This is highly problematic, as energy delivery (and hence apparent network productivity) is responsive to price changes for reasons such as the carbon tax and fuel prices.

Fundamentally, network businesses provide capacity and quality to customers, analogous to the services provided by roads (Frontier Economics 2010b, pp. 41ff). Accordingly, outputs (or cost drivers) would include:[[10]](#footnote-10)

* customer numbers (new connections raise costs) by type (commercial and household)
* the capacity to carry power to dispersed customers (through transformers, network kilometres, by the level of kV) when it is required. This would need to consider network types, such as the mix of central business district, urban, short rural or long rural assets (ETSA Utilities et al., sub. 6, p. 32)
* ensuring adequate quality (reliability requirements) and low transmission losses
* the capacity to cater for peak demand (the load factor or peak to average demand measures).[[11]](#footnote-11) If this variable is not available, then energy supplied may be a practical control variable, since otherwise an over-engineered network with excessive capacity may appear to be efficient
* input prices (as suggested by Ergon Energy, sub. 8, p. 10 and the ENA, sub. 17, p. 30).

If there is an insufficient sample for estimation, composite cost drivers may be relevant (widely used by OFGEM, but declared to be ‘arbitrary’ by NERA 2007, p. 3, pp. 29ff).

It is also important to distinguish short-run efficiency (where the focus is on operating expenditures) from long-run efficiency (where capex *and* opex are the relevant cost). While generally in favour of the greater use of benchmarking in regulatory determinations, the MEU acknowledged this limitation:

… benchmarking of capex is less readily applied due to the “lumpiness” inherent in some of the capital investments required in the energy transport sector. This, of course, should not be a reason not to benchmark. … This requires the excision of the large augmentation projects (the “lumpiness”) from the capex program and then benchmarking becomes essentially straightforward and useful. The few “lumpy” elements can be assessed in their own right. (sub. 11, pp. 17‑18)

The ratio of replacement investment to the age-weighted value of existing capital stock provides one indicator of whether rising capex reflects the (efficient) need for the orderly replacement of older assets (Ausgrid, sub. 19, p. 6). Others — such as NERA (appendix B of ENA, sub. 17) — argue that the relevant metric is the share of assets close to expiry.

## 4.4 The use of benchmarking for Australian electricity networks

### The use of benchmarking under previous regulatory regimes

Prior to the creation of the AER, state and territory regulators often used benchmarking of electricity networks as a tool, but not in determining revenue allowances (box 4.1). Several regulators used benchmarks to test the reasonableness of proposals (ICRC, IPART, QCA). Others used benchmarks to flag areas that might need further analysis (ETSA) or to assess a base year expenditure (ERA). In other instances, benchmarking has apparently not been influential (OTTER, NTUC). Chapter 6 examines these studies’ empirical estimates of network inefficiency.

Notably, the Northern Territory relies particularly heavily on comprehensive benchmarking to determine the rate of growth in their CPI-x framework.[[12]](#footnote-12)

### Benchmarking in the current regulatory framework

The Rules specify that from late 2014, the AER *must* prepare annual benchmarking reports of the relative efficiency of distribution and transmission network businesses in the NEM (box 4.2). Amongst other sources of information, the AER must ‘have regard’ to these reports in assessing the reasonableness of business’s building block and revenue proposals. Prior to the recent Rule change, there was no requirement for formal and regular benchmarking. Nevertheless, the AER used benchmarking in its regulatory determinations (box 4.3), albeit with some limitations.

* While the AER has commissioned various consultants to undertake benchmarking analyses in the various distribution determinations, the consultants have used different benchmarking techniques, with varying access to industry-knowledge and data.
* Not all networks have been analysed using the same modelling processes. For instance, the Repex model was only developed in time for use in the Victorian distribution determination, and has yet to be applied to several other jurisdictions (AER 2011a, p. 11).
* The results of any benchmarking exercise have typically only informed the AER’s state-specific revenue determinations, even though there may be clear implications for other jurisdictions.
* The AER (and others) have indicated that it has made limited use of benchmarking and has focused on ratio analysis (AER, sub. 13, p. 14).[[13]](#footnote-13) The MEU (sub. 11, p. 12) claimed the regulatory use of benchmarking had ‘been supported in principle but has become somewhat inconsequential in practice.’ The AER has claimed that the Rules have frustrated its capacity to set a revenue allowance using benchmarking (sub. 13, p. 14), thus limiting its usefulness. The Rules have now been amended.

## 4.5 Criteria for judging benchmarking

There are many benchmarking methods (as discussed above), multiple uses for these, varying processes for testing them and different regulatory practices for gathering data, communicating results, and addressing compliance burdens. A key question is how to separate the wheat from the chaff among the various competing approaches, recognising that this will typically involve balancing various criteria (figure 4.7).[[14]](#footnote-14)

|  |
| --- |
| Box 4.1 State and territory regulators’ use of regulatory benchmarking |
| State and territory regulators have used partial productivity indicators to benchmark electricity network performance:   * IPART compared various expenditure ratios over time for New South Wales distribution networks, using data from the years 1999‑2000, 2003‑04, and 2008‑09. The results were used to test the reasonableness of expenditure allowances. * ICRC compared ActewAGL with five Victorian distribution networks across various opex ratios, using data from 2002‑03. The results were used to test conclusions about expenditure allowances. * ESC (Essential Services Commission Victoria) compared a deconstruction of operating expenditure (opex) growth to measure partial factor productivity. Trend analysis was also carried out for capex, comparing networks from Victoria, New South Wales and New Zealand. The opex comparisons were carried out to determine the ‘rate of change’ and ‘growth factor’. * QCA compared opex rations between Energex, AGL, United Energy and EnergyAustralia. Capital expenditure (capex) ratios were compared between Energex, Ergon and Victorian distribution networks. The results contributed to the assessment of reasonableness for expenditure proposals. * ESCOSA compared expenditure ratios between ETSA and 10 other distribution networks. The results were used to identify areas requiring further analysis. * Office of the Tasmanian Economic Regulator (OTTER), conducted trend analysis for opex and capex using data from 2002‑03 to 2011‑12. The results were not considered in the determination. * Economic Regulation Authority (ERA WA) compared actual opex between Western Power and all other Australian distribution networks for 2007‑08. The results were used to assess the base year expenditure for 2007‑08. * NTUC used a ‘multilateral unit opex’ method to benchmark opex for water and energy utilities.   Regulators have also undertaken benchmarking using comprehensive indices such as TFP:  The Northern Territory operates a CPI-x framework for the escalation of network prices, using estimates of productivity to determine the X factor. The industry TFP growth trend is used in the calculation of the x factor (see footnote 12 above). |
| *Source*: ACCC/AER (2012a), tables 2.2 and 3.3. |
|  |
|  |

|  |
| --- |
| Box 4.2 The Rules and benchmarking |
| The National Electricity Rules (v. 54) makes 53 references to ‘benchmarking’ or ‘benchmark’, predominantly relating to the financial variables that determine the weighted average cost of capital. However, for the purposes of this inquiry, the most important references are to the opex and capex of distribution and transmission networks. Considering distribution networks, the Rules stipulate various decision-making stages:  **Publication of an annual benchmarking report:**  The AER must produce an annual benchmarking report of the relative efficiency of each distribution network business (s. 6.27) — a Rule change that occurred after the Productivity Commission’s draft report. The AER must publish the first by September 2014. In preparing the report, there must be consultation with NSPs and the jurisdictions, and the capacity, but not the requirement to consult with people with an interest in the subject and the public (s. 8.7.4(b)).  **The AER must accept reasonable opex:**  Clause 6.5.6(c): The AER must accept the forecast of required operating expenditure of a distribution network business that is included in a building block proposal if the AER is satisfied that the total of the forecast operating expenditure for the regulatory control period reasonably reflects: (1) the efficient costs of achieving the operating expenditure objectives; (2) the costs that a prudent operator would require to achieve the operating expenditure objectives; and (3) a realistic expectation of the demand forecast and cost inputs required to achieve the operating expenditure objectives.  **The AER must reject unreasonable proposals:**  Clause 6.5.6(d): If the AER is not satisfied as referred to in paragraph (c), it must not accept the forecast of required operating expenditure of a distribution network business that is included in a building block proposal.  **The AER must have regard to benchmarking when evaluating reasonable and unreasonable proposals:**  Clause 6.5.6.(e): In deciding whether or not the AER is satisfied as referred to in paragraph (c), the AER must have regard to the following (the operating expenditure factors): … (4) the annual benchmarking report published under rule 6.27 and the benchmark operating expenditure that would be incurred by an efficient distribution network business over the regulatory control period [among 10 categories].  **The AER can make its own reasonable estimate (clause. 6.12.1(4)(ii))**  While it must give reasons for its rejection of the business proposal (cl 6.12.2), the AER has the discretion to accept or reject any element of a proposal (cl. 6.12.3(a))  There are parallel requirements for capex for distribution network businesses (clauses 6.5.7(c), 6.5.7(d), 6.5.7(e), and 6.12.1(3)(ii). |
|  |
|  |

|  |
| --- |
| Box 4.2 continued |
| **Transmission businesses**  There are also similar, but not identical, clauses for transmission network businesses:   * an annual benchmarking report (clause 6A.31) * for opex (clauses 6A.6.6 (c), 6A.6.6 (d), 6A.6.6 (e)(4)) * for capex (clauses 6A.6.7 (c), 6A.6.7 (d), 6A.6.7 (e)(4)) * a capacity for the AER to put in its own estimate (cl. 6A.13.2), with reasons (cl. 6A.14.2)   The two important differences are that for transmission network business, the clauses relate to the business’s revenue proposal, and the issue of discretion is implicit, not explicit. |
| *Source*: National Electricity Rules, version 54. |
|  |
|  |

|  |
| --- |
| Box 4.3 Examples of benchmarking in AER determinations |
| On behalf of the AER, Nuttall Consulting (2011) benchmarked Aurora’s expenditure levels against several other networks. This included various ratios of capex and opex. Comparisons were made against other networks, states, or regions (for example, rural Victoria). Nuttall Consulting (2010a) also benchmarked Victorian distribution expenditure levels and ratios against those of other networks and States.  Wilson Cook (2008) analysed measures of expenditure for distributors in the Australian Capital Territory and New South Wales, and compared them with a subset of other networks. A bottom-up analysis was also undertaken, and recommendations were based on both sets of results.  Parsons Brinckerhoff (2009a, 2009b and 2010) compared opex measures for Queensland and South Australian distributors against other networks.  ETSA Utilities et al. (sub. 6, p. 24) noted that network businesses themselves have submitted benchmarking results, for example in relation to opex. These businesses (pp. 60ff) also gave a comprehensive account of the benchmarking undertaken by the AER in respect of Victorian distribution network service providers (and how these results demonstrated the efficiency of the Victorian businesses). |
| *Sources*: Nuttall (2010a, 2011); Parsons Brinckerhoff (2009a, 2009b, 2010); Wilson Cook (2008). |
|  |
|  |

Figure 4.7 Evaluation criteria for benchmarks and benchmark practices

|  |
| --- |
| Figure 4.7 Evaluation criteria for benchmarks and benchmark practices. This figure shows several evaluation criteria for benchmark measures, agency benchmarking practices and statistical benchmarking practices. |

It is useful for any discussion of such criteria to break them into three subgroups:

* the characteristics of a good benchmark from a scientific perspective, of which validity is the initial consideration (sections 4.6 and 4.7)
* the statistical processes used by the analyst to test whether any result is likely to be useful for its regulatory purpose (section 4.8)
* the processes used by the regulator in undertaking benchmarking. (This issue is examined in chapter 8, which relates to how the AER should use benchmarking.)

## 4.6 Validity — does the measure test what it claims to?

A valid benchmark should relate to the relevant concept — efficiency (or conversely inefficiency) in one or more meaningful dimensions.

A low failure rate (in say minutes lost per customer) is not a measure of efficiency if customers do not value the reliability benefits above the costs of achieving them (chapter 14).

Similarly, while higher revenues per connection *could* be a reasonable measure of inefficiency:

* that is not necessarily true. For example, higher revenues of one business compared with others may reflect a regulator’s reasonable decisions about the weighted average cost of capital (WACC) at the time of the regulatory determinations
* it is not clear to what type of inefficiency it relates. A high revenue per connection could reflect an inefficiently high WACC (leading to excessive prices, with deadweight costs for customers, but not necessarily excessive investment), or low productivity, or both. Accordingly, it is not a valid measure of its separate components. Data permitting, a better approach would be to break revenue per connection down into its price and productivity components, since these have different implications for policy. For instance, if the business were technically efficient, but the WACC was too high, the regulator would concentrate on the latter in its determinations.

An overarching concern is that any valid benchmarking measure should reflect the way that the businesses are run. As Turvey (2008a, pp. 2‑3) put it in relation to opex benchmarking:

Applying different econometric methods to find which method and which of such variables give the ‘best’ results is very different from the down to earth approach of understanding the industry sufficiently well to identify and describe the immediate determinants of the operations and maintenance work actually required.

### Valid benchmarks of efficiency will often need to take account of time

Efficiency has a time dimension. It can be efficient to ‘over-invest’ in certain assets ahead of their full utilisation because investment must precede production or because it can be less costly to build in spare capacity at a given time, than to re‑invest at a later time to add further capacity.[[15]](#footnote-15) One of the reasons for the recent slowdown in measured productivity of the Australian economy is that mining companies made large investments ahead of the extraction of output. Few suggest that the Australian mining industry is inefficient for this reason.

Benchmarks that fail to recognise the implications of timing can be misleading. For example, suppose that a business makes its investment too early compared with a peer that invests optimally, but that both start and finish with the same capital stocks. Assume for illustrative purposes that the outputs of the businesses are identical in each year. In the example shown in figure 4.8, the comparative investment levels required to achieve a given output — a measure of their relative investment efficiency — suggests that the ‘inefficient’ firm is grossly inefficient in the initial years and then more efficient in later years. Yet in present value terms over the relevant period, the early investing firm’s relative inefficiency is only around four per cent. Using snapshot efficiency measures for the first few years would not provide a valid measure of the real efficiency gap between the businesses. Static measures are still useful, but need to be carefully interpreted.

### Controlling for factors outside the control of business and their relevance to valid conclusions about inefficiency

There is a large literature on estimating the comparative costs of businesses, with much of that literature concentrating on using the ‘right’ techniques. However, it is equally important to be clear about how to interpret benchmarking results for policy purposes because the misuse of good technical analysis can result in adverse outcomes for consumers and businesses. In particular, comparing the costs between businesses in different jurisdictions without accounting for factors outside the control of the business could provide misleading indicators of managerial efficiency. If used in incentive regulation, this could lead to underinvestment or unwarranted transfers from consumers to the businesses.

The measured level of comparative business performance across Australian and international jurisdictions reflect four broad factors.

Figure 4.8 Efficiency often needs to be dynamically assessed

|  |  |
| --- | --- |
|  |  |

a This figure compares the efficiency of a firm that ‘prematurely’ invests (firm A) and a firm that invests at the optimal time (firm B). The illustration assumes that both firms have the same initial and final capital stocks and investment levels, and the same output levels throughout the period. However, firm A builds up capacity ahead of its productive use. In contrast, firm B uses its capacity efficiently throughout, but must increase its investment rate towards the end of the period to maintain its capital stock and productive capacity. A period-by-period measure of the inefficiency of firm A is the difference between its investment level and that of firm B as a share of output. This is clearly worst at the point where its investment growth rates are high early in the period. For example, looking at year 4, firm A has a level of investment to output ratio more than 40 per cent higher than firm B — an apparently very low level of efficiency. However, by year 15, a snapshot view suggests that firm B is less efficient. An overall view of the relative inefficiency of firm A is the difference between the net present value of its investment stream over the period and that of firm B, divided by the net present value of output. In this illustration, firm A uses just over four per cent more resources to produce the same output as firm B. Snapshot measures can therefore mischaracterise the longer-run level of efficiency differences between businesses.

*Data source*: Commission estimates.

#### Managerial inefficiency

Managerial inefficiency may be manifested in many ways. A network business may not organise itself efficiently, resulting in higher than necessary costs and/or poorly structured prices. It is important to emphasise that managerial inefficiency is primarily not about the performance of managers per se, but primarily about how the governance and organisation of a business can affect any aspect of its performance. For example, a business may be overstaffed, have insufficient expectations of workplace productivity, be overly combative or passive in its dealing with its unions, and over-invest. Managerial inefficiency does not always mean ineptitude or lack of industriousness — highly competent people can pursue inefficient outcomes with hardworking gusto. Moreover, a business could make the most efficient use of its resources at a given time, yet still be encumbered with an inefficient (sunk) capital stock that reflected investment decisions made by previous managers.

Inefficiency also comes in different sizes. The critical concern is not the presence of inefficiency per se, as no industry is perfectly efficient, but whether the divergence in efficiency matters from a policy perspective.

#### Environmental differences outside the control of the businesses and government

Network businesses are justifiably concerned about ‘like with like’ comparisons.[[16]](#footnote-16) Many of the criticisms levelled by network businesses against the findings of Mountain and Littlechild (2010) reflect their concerns that the analysis did not control sufficiently for extraneous differences in the operating environments of the businesses.[[17]](#footnote-17)

Network businesses operate in different environments that are outside their and government control, including the physical aspects of that environment (topography, forestation, climate, soil type, temperature and wind), customer density and type, the form and location of generators, and the prices of inputs, such as wage costs and the costs of wires and substations.[[18]](#footnote-18) For example, the efficient costs per customer of providing network services to highly dispersed customers in a rural region are greater than the costs of providing services in a city suburb (figure 4.9). In relation to its own network, Ergon Energy indicated that its:

… network area covers more than one million square kilometres, which is over six times the size of Victoria. This network characteristic would impact on our performance against reliability standard targets due to accessibility issues and distance to travel to faults. (sub. 8, p. 3)

Some participants pointed out the need to engage with the business’s actual operations at a detailed level:

… it may be that poles are twice as expensive in one jurisdiction compared with the other (e.g. due to different proximity to a hardwood industry ‑ noting that poles are expensive to transport). (ENA, sub. 17, p. 28)

Figure 4.9 Customer density differs between networks

Customers per km of distribution line

|  |
| --- |
|  |

*Data sources*: AER (2011b); distribution businesses’ websites and annual reports.

In certain instances, it is difficult to categorise definitively whether a factor is fully outside the control of the business:

* peak demand can be partly controlled by the businesses given they can undertake demand management programs (AER 2010a). Failing to account for this in benchmarking could discourage further improvement of demand management programs (chapters 9 to 12)
* businesses provide advice on reliability standards
* in many cases governments, rather than network businesses, make the decision about whether to put cables underground — which lies clearly outside the control of the businesses. However, in some other instances, network businesses will make the decision to install cabling underground (‘undergrounding’) because it increases reliability and reduces vegetation clearing costs, or because there is a mutual benefit to them and to co-contributing customers from undergrounding. Consequently, were businesses sometimes to overinvest in undergrounding then this would not show up as inefficiency were a benchmarking exercise to control for all undergrounding.

#### Policy differences

Network businesses in different jurisdictions are sometimes subject to different regulatory and policy frameworks, which are outside the business’s control, but within governments’ control. Such policies and regulations can affect a network business’s practices, costs and productivity. For example, governments may set reliability standards (ENA, sub. 17, p. 7), may own the business concerned (a policy choice of government, not the business), and where state-owned, stipulate their governance arrangements or give them non-commercial directions (chapter 7). Governments may also mandate various environmental policies — such as feed-in tariffs for household photovoltaic generation — that may physically affect the efficiency of the network and add to overall customer prices, reducing power demand, and utilisation of the network (chapter 13).

This is relevant to policy in two ways:

* benchmarking for incentive regulation needs to control for any important policy/regulatory arrangements that affect business performance
* the information from controlling for such arrangements in any benchmark modelling can be used to improve aspects of the policy environment (such as divergent and unjustified differences in reliability standards; or limits on the use of demand management through price controls). When reporting on its benchmarking modelling results, the AER should highlight publicly any major effects on productive efficiency from policy and regulatory settings.[[19]](#footnote-19) A major reason for the Commission’s consideration of ownership, demand management, and reliability is that these may be a source of greater inefficiencies than the managerial practices per se. For example, IPART (2010, p. 49) commented that the quality of network planning and the decisions on the licence conditions that drive capital expenditure will be critical for future productivity performance in electricity distribution. It is also possible that the parameters indicating the productivity and cost impacts of policy settings are more statistically reliable than the efficiency estimates for any given business.[[20]](#footnote-20)

The policy benefits from exploiting benchmarking models in this way have a long lineage. Indeed, where regulations and policies particularly affect business performance, this is sometimes the major reason for benchmarking.[[21]](#footnote-21) Some of the impetus for regulatory reform of Australian utilities a decade ago arose from evidence about its beneficial impacts in overseas countries.

#### Interactions between policy and managerial inefficiencies may also be important

While it is important to distinguish managerial inefficiency from economic efficiency generally, it can be equally important to consider their interaction. Trade policy provides a well-understood illustration of the issue. Import tariffs have adverse impacts on efficiency by distorting people’s consumption choices and diverting resources to industries that have a lower comparative advantage. That type of inefficiency need not involve any managerial inefficiency in that the business managers may still seek to minimise costs and set prices as efficiently as they can in the constrained world they face. However, tariffs may also have a dynamic effect on efficiency by lowering the incentives for managerial performance. This might occur because business managers feel insulated from competition or believe that governments will adapt tariffs (or provide subsidies) to maintain their business’s profits regardless of their competitiveness. The managers may be intrinsically highly competent, but they respond to the incentives they face.

In the case of network businesses, the factors outside managers’ control that may affect their managerial efficiency include, among other things, private or state ownership (and any associated issues with labour relations and governance), and the regulatory arrangements used to constrain monopoly profits. An enterprise has less pressure to rigorously test all investment and other spending options for its efficiency if it is partly shielded from the consequences of its decisions. This represents a form of ‘moral hazard’ of the kind that makes people less careful where they are insured against the consequences of certain risks.

For instance, to the extent that an enterprise can link much of its spending to the imperative to meet reliability standards (thereby receiving regulated returns for its spending), then it has weaker incentives to control that spending. This is more likely where a reliability standard is not anchored to the customer value of reliability, and where the knowledge about how to achieve the standard lies mainly with the business, rather than the regulator. If the business is particularly sensitive to public perceptions of instances of unreliability then it may spend too much to reduce these risks, even if the standard did not require that. Moreover, as high reliability standards require greater capital expenditure, they provide a capacity for greater costs from inefficient management practices.

Once an external factor outside the control of the business also affects managerial efficiency, then fully controlling for that external factor in benchmarking analysis means that managers do not face the full incentives for efficient behaviour — subverting the goal of benchmarking.

### Data problems

Any data inputs into benchmarking models are subject to error due to measurement problems, small differences in the definitions used by different businesses and the periods to which the data relate, and simplification of the relationship between costs, inputs and outputs. Businesses may not report reliable information. The AEMC (2009a, p. 8) pointed out:

One regulator stated that the data provided by a service provider was so unreliable that it was difficult to conduct an independent audit of the data.

Before the AER assumed responsibility as the distribution network regulator, state and territory regulators collected data on network businesses, but did not employ a common framework (AER, sub. 13, p. 18). Lawrence (2009) notes that coverage of key variables such as opex has varied over time (p. v). Ergon Energy (sub. 8, p. 4) observed that the AER has only recently begun to implement consistent data collection. Such data limitations lay behind the AEMC’s view that TFP methods could not yet be used to set x factors in CPI-x methods.

There are also significant data gaps. Some of the recent capital expenditure has reflected the need to replace assets close to their end of life (chapter 2; Topp and Kulys 2012), but data on asset vintages is incomplete.[[22]](#footnote-22) More generally, leading benchmarking practitioners identified gaps in physical network data, which the AER is only now systematically addressing:

For instance, regulatory reporting guidelines deal almost exclusively with financial matters required for building block regulation with little or no mention on any physical system data. Thus, while they examine the moneys received and spent, and the financial characteristics of assets used, created and depreciated, there is no quantification of *what* assets are built, maintained or operated to deliver the network service. Jurisdictional regulators noted that their efforts to date had almost exclusively been directed at obtaining the financial data required for building blocks regulation and they had had little time to assemble data on physical characteristics and outputs. (Lawrence and Kain 2009, p. 19)

Others point to the expedient use of easily available data for benchmarking, rather than the variables most suited to efficiency measurement (APA Group, sub. 2, pp. 1‑3).

Nevertheless, *any* quantitative method is subject to error and data problems. The key question is whether it matters for the regulator’s intended purpose (an issue to which this chapter returns). The MEU identified data limitations but — with some reasonableness — pointed to the fact that:

There is a need for a consistent approach for the gathering, manipulation and display of data to make the best use of benchmarking. However, even imperfect data can provide useful insights and should not be excluded, even though its use might be minimised. … ‘rough and ready’ can provide a strong indication as to whether the proposed opex, capex and WACC outcomes are grossly inefficient and whether deeper analysis is required to ensure a more efficient outcome is possible. (sub. 11, p. 25)

#### Australia has few observations

There are only 13 distribution businesses, five regional transmission businesses and three separate direct current interconnectors in Australia.[[23]](#footnote-23) This reduces the feasibility for more elaborate models that take into account the multiple environmental factors affecting inter-firm performance.[[24]](#footnote-24) As the Consumer Action Law Centre argued:

However, with a relatively low number of network businesses to start off with, this may cause a particular challenge for the NEM that has been more easily overcome in overseas markets. (sub. 5, p. 4)

The case of interconnectors is even more problematic:

In the case of ‘single’ electricity transmission assets such as Murraylink and Directlink, benchmarking is not an appropriate way to establish the price or revenue path. Their ‘output’ is the continued availability of full interconnection capability … These assets are unique and have no logical comparators on which benchmarking could be based. (APA Group, sub. 2, p. 2)

Given sample size problems, benchmarking is most likely to be effective in its application to distribution businesses (Grid Australia 2009), although Ergon Energy argued that even here, the numbers were too small to provide meaningful results (sub. 8, p. 3). The latter suggested that benchmarking reliability might be improved by including both transmission and distribution businesses in the analysis (sub. 8, p. 10).

#### Would international benchmarking help?

While international benchmarking might increase the number of observations, it raises other challenges for valid comparisons. For example, ETSA Utilities et al. (sub. 6, p. 27) and the ENA (sub. 17, p. 25, p. 49) noted that results are affected by differences in exchange rates, reliability standards, accounting policies, tax laws and corporate structures.[[25]](#footnote-25) The question of whether Australia has high, medium or low electricity prices — a potential test of efficiency — has proved difficult to resolve (Mountain 2012a in contrast to NERA 2012b). However, it appears that the results are sensitive to choices about the appropriate exchange rate and the retail tariff (regulated versus market rates). It merely serves to highlight some of the complexities of comparisons of this kind.

On the other hand, the Australian Academy of Technological Sciences and Engineering cited the value of the International Transmission Operation and Maintenance benchmarking (ITOMs), which is conducted every two years, covering about 30 businesses (ATSE, sub. 9, pp. 1‑2). Australian transmission businesses are included in ITOMs and use it for internal purposes. ATSE also recommended the inclusion of Western Australia and the Northern Territory as comparators. While the MEU (sub. 11, p. 25) noted the difficulties of international comparisons, it considered that it could ‘still provide some useful information’.

The Council of European Energy Regulators (2012) has undertaken extensive benchmarking of the quality of electricity supply — arguing that successive benchmarking analyses in this area may have prompted improvements in quality. However, even in this narrow area, it encountered significant differences in definitions, which complicated its benchmarking.

Overall, international benchmarking should not be dismissed, especially as the businesses use it themselves. However, it is likely to be most useful in specific aspects of the performance of the businesses where measures of physical inputs and outputs are available (such as reliability or certain labour productivity measures), for assessing business processes (such as the use of tendering and outsourcing) and for raising issues of concern that the regulator might pursue in further detail.

## 4.7 Other scientific criteria for judging benchmarking

While the validity of any benchmarking measure is a prerequisite for its use, there are also other criteria important for assessing the quality and interpretability of various benchmarks.

### Accuracy and reliability

Accuracy is the degree to which a benchmark provides an unbiased estimate of efficiency, while the reliability (used here in the normal sense of reproducibility) is about the variance of the measure. For example, a bathroom scale that is consistently out by exactly 10 per cent is inaccurate, but reliable. As discussed above, a failure to adequately control for differences in operating environments can lead to heavily biased measures. More subtly, certain statistical models of productivity may use underlying assumptions at odds with the known properties of the variable of interest.[[26]](#footnote-26)

### Robustness

This is a subset of accuracy and reliability, but worth emphasising in its own right. A particularly useful robust measure is one that provides information about the efficiency of an enterprise regardless of its operating environment. Although affected by other kinds of errors, some measures of management performance (such as those underpinning figure 4.2) might fall into that category. However, such measures are likely to be the exception rather than the rule.

If the results of a model are sensitive to small perturbations in the underlying data, the addition of control variables with little expected impact, the removal or addition of a single network business, or to modest changes in assumptions and estimation techniques, then benchmarking results are at best indicative, and at worst, useless.

* Farsi and Filippini (2005) found that in a study of 52 distribution companies, the efficiency scores and rankings were significantly different across various methods, and especially between parametric and non-parametric models.
* A study of 33 Polish distribution businesses found a large dispersion in efficiency results for any given business (figure 4.10). These results would not be usable for regulatory purposes unless model specification tests were able to eliminate most of the models. This underlines the importance of testing.
* IPART’s (2010) analysis of state-owned corporations, which included electricity networks, showed that the choice of input variables changed the rankings of firms and their TFP scores.
* The use of partial productivity indicators is also sensitive to specification. For instance, in their report to the WA Economic Regulation Authority, Wilson Cook (2009, p. 86) made several partial productivity comparisons between Western Power and network operators in other states. Three comparisons were made of distribution opex, separately accounting for customer numbers, line length (km), and electricity consumption (kWh). While these comparisons produced consistent *rankings* at a state level, they did not provide consistent measures of opex efficiency.
* Complex non-linear models appear to be particularly prone to problems. For example, in their review of European benchmarking of electricity networks, Schweinsberg et al. (2011, p. 55) found unstable convergence on parameters in stochastic frontier analysis (a common benchmarking approach), such that it was not possible to distinguish inefficiency from statistical noise.

### Limited susceptibility to manipulation or gaming

As in all systems where rewards and punishments depend on incomplete measures of performance, the measured party has incentives to ‘look’ like a highly performing entity (whether it be a hospital, school, a student, CEO, employee, or in this case, network businesses). As the Consumer Action Law Centre observed:

It is thus important to be mindful of this when designing benchmarking mechanisms, but even more so when interpreting and applying the results. If given the opportunity, we can assume that the network businesses would seek to strategically influence the benchmark indicators and as such, the result could be just another layer of ‘gaming’. (sub. 5, p. 5)

Figure 4.10 Estimates of inefficiency vary depending on the method

Results for various measures of efficiencya

|  |
| --- |
|  |

a The results are based on applying 14 benchmarking methods (variants of DEA, COLS and SFA) to 33 Polish electricity distribution businesses. Each vertical line shows the various efficiency measures corresponding to the 14 methods. (Often there will not appear to be 14 observations per business, because the measures sometimes coincide.)

*Data source*: Cullmann et al. (2006).

Accordingly, the regulator should consider the capacity of any particular benchmarking measure to create unforeseen business behaviours.

* A benchmark measure might be a proxy for some hard-to-observe characteristic of efficient businesses or be sufficiently vaguely represented that firms can meet the benchmark, but not the underlying goal (CREG and SUMICSID 2011, p. 36). For example, a business may be able to meet some benchmarks by changing their cost-allocation methods (Jamasb et al. 2003), even if the underlying resource allocation is not optimal.
* The regulator also needs to consider the interdependence between some efficiency measures.[[27]](#footnote-27) Measuring one aspect of efficiency (such as opex per kilometre of line), but ignoring another due to measurement difficulties (capital productivity), may result in inefficient substitution in distribution systems. This is one of the advantages of the CPI-x methodology and aggregate benchmarking.
* Distorted information may directly affect the incentives of complying firms (NERA, appendix B of ENA, sub. 17, p. 34).

### Parsimony

A good model should be no more complex than required. This is important in assisting interpretability, avoiding data mining, achieving robust results, reducing data collection costs and allowing greater comparability of results across countries (since it is easier to ensure common definitions for a few variables).

An important facet of parsimony is to forgo complexity where it adds little explanatory power. For example, a Productivity Commission Staff Working Paper (Sayers and Shields 2001, p. 178) investigated the sources of price differences between Australian electricity distribution businesses, and found that many factors had small impacts. For example, at that time, the study found vegetation growth and its management had small impacts. (Whether that result remains true is uncertain, but the point is that it will sometimes be possible to identify factors that are largely immaterial for costs, and that can be eliminated as control variables.)

Similarly, in an exhaustive study of Belgium distribution network businesses involving multiple measures of outputs and inputs, the only relevant outputs were total circuit length of lines, customer numbers, and the total number of connections (CREG & SUMICSID 2011). One of the valuable aspects of international benchmarking analyses is they may be able to identify the variables that matter for networks more generally.[[28]](#footnote-28) This may then allow greater confidence in using a small range of variables in the Australian context where the sample size is low.

### Fitness for purpose

As discussed further in chapter 8, benchmarking has multiple purposes. Some require great accuracy, reliability and robustness. This is particularly important where benchmarking is used to determine a business’s revenue allowance. Such benchmark estimates should be highly reliable across time, business types and jurisdictions. The concerns are less where benchmarking is indicative — used to identify areas for possible future investigation, or to reach some prima facie judgment. This is how the Commission regards the preliminary results presented in chapter 6.

## 4.8 Testing the credibility of results

Benchmarking models do not actually estimate inefficiency, although this is how they are generally interpreted. The results of any benchmarking model show the extent to which the model *fails* to explain performance:

… it is incorrect and misleading to ascribe the residual to ‘inefficiency’, or to describe the benchmark as a measure of ‘efficient costs’. Instead, one must acknowledge that the residual measures no more than the element of observed costs that the model has *failed to explain*. (Shuttleworth 2005, p. 315)

… an observation of difference does not itself constitute *diagnostic* evidence, the gold dust of effective assessment. (Yarrow 2012, p. 4)

That is, the inefficiency of any business is the difference between the business’s observed performance and that predicted by a set of cost drivers. This can reflect missing cost drivers, data errors, incorrect estimation methods, and invalid assumptions about the functional form and error distributions.

The way of appraising the credibility of the results is through systematic investigation and reporting, including:

* explanation and graphical presentation of inputs and outputs and their main statistical features (averages, variances)
* divulgence of model selection processes, how data may have been manipulated, and why potentially relevant variables have been omitted
* comparisons with alternative models, and why the ultimately selected model(s) is superior to others
* tests of model adequacy, such as tests of linearity, normality or otherwise of the residuals, parameter stability with different sub-samples, the influence of outliers, and, subject to the caveats spelt out later, statistical significance
* corroboration tests, which include assessment of the consistency in business’s efficiency measures rankings based on different methods and over different periods. For example, if efficiency measures are reasonably stable over short periods of time, this increases the confidence in the results (since most businesses cannot improve their efficiency in very short periods of time)
* explanation of what the results practically mean, and the possible flaws in the modelling.

The Commission examined a host of benchmarking studies. Very few undertook comprehensive testing of the models. This is acceptable where the results are ‘indicative’, but not where a regulator might use them to determine a business’s revenues.

### The capacity for statistical inference

This is probably the most neglected issue in interpreting various benchmarking measures, although essential to its meaningful use. This is why it is worth carefully dissecting.

Many people think of a benchmark as ‘a’ number. However, given the points made above, no model or measure is perfect. Accordingly, any benchmarking methodology should be able to identify a justifiable confidence interval around a predicted benchmark result. As an illustration, a benchmarking model might predict an efficiency score of 75 per cent for a given business compared with an industry average of 85 per cent (say the ‘benchmark’). By themselves, these are not useful numbers. There are two broad ways of considering such estimates.

It is frequent for economists to argue that the difference between two estimates is, or is not, ‘significant at the 0.05 significance level’.[[29]](#footnote-29) The implicit notion is that 0.05 or some other lower probability that the difference could have arisen purely by chance somehow legitimises the value of the estimate. Under this approach, were the 85 per cent estimate to have a wide confidence interval then the regulator would be reluctant to use the 75 per cent estimate as a sign of genuine inefficiency for the business concerned (and probably be right in the context). On the other hand, if the confidence interval around the benchmark was very tight, they might regard 75 per cent as a reasonable measure of inefficiency for benchmarking.

However, this approach is often not useful in deciding what policy action to take (McCloskey 1985a and McCloskey and Ziliak 1996, 2008). The appropriate framework to use is the so-called ‘loss function’, which considers the costs of errors around any point estimates. This is not esoterica — failing to do this can have major adverse effects on the economic efficiency and the distributional impacts of regulatory policy. Part of the reason that the Rules were designed in their current form was the view that making an error that led to lower investment would be more costly than the alternative. An international review of benchmarking methods and their practical application observed:

The principal disadvantage of benchmarking is the potential that a model of poor quality can expose utilities to undue risk. While regulation must protect consumers from monopoly abuses, it must also not compromise the ﬁnancial viability of regulated entities. … regulatory opportunism that violates the [need for the utility] to raise funds, operate successfully and reward its investors for the risk they assume, must be prevented. (Lowry and Getachew 2009, p. 1325)

Biggar has questioned whether the costs of errors lie in this direction (as discussed in appendix B), but accepts that the costs of errors is an important one, a perspective that is unrecognised in many benchmarking studies.

These issues have several key implications for any benchmarking practices where the regulatory stakes are high, such as determining revenue allowances.

* The regulator should test assumptions about the nature of the distribution around a benchmark estimate. It is common to assume errors are normally distributed, though in practice, this is impossible for some benchmarking measures. For example, any simple regression — say of opex against customer numbers — cannot have normally distributed errors, because opex cannot be negative. That may not matter much in many applications, but it will in some.[[30]](#footnote-30)
* The regulator should attempt to estimate, or at least make explicit its assumptions about, the loss function it believes is reasonable. That is not a trivial exercise, but at least transparent assumptions would be a useful step.
* Consultation with independent engineering experts with good knowledge of network business operations can help provide a basic credibility test of model results or of assumptions by parties not familiar with the actual operations of businesses. Similarly, as Pollitt (2005, p. 283) notes, as much as possible, it is desirable to test the consistency of a business’s benchmark with financial market perceptions of its relative performance. The 1986 Challenger Shuttle disaster provides a graphic illustration of the potential divergence between expert engineering advice and hunches or statistical misunderstandings. NASA management claimed that the chance of a catastrophic failure of a shuttle was 1 in 100,000 based on misinterpretations of safety factors and unjustified optimism. The engineers thought it was between 1 in 50 and 1 in 100 (Feynman 1986), a verdict that was found to be more compelling. It is quite conceivable that over-confident benchmarking modellers might make errors analogous to this — at least in terms of the consequences for a business (or consumers). This suggests the importance of engineering and financial analysis as a supplement in interpreting statistical benchmark results (and one of the reasons that the AER needs further resources to access such expertise).

### Explanation of inefficiencies

Surprisingly, this is a rarely mentioned aspect for evaluating alternative benchmarking models,[[31]](#footnote-31) and yet one of the most crucial for policy (and management purposes). Even if the measured inefficiencies of various businesses are regarded as accurate, it leaves open the question of why some businesses are managed less well than others. Some of the factors that may be relevant include the use of obsolete technology, little innovation, weak corporate governance or the form of ownership. If a benchmarking model can credibly unearth the behaviours that lead to managerial inefficiency, it corroborates the measures of inefficiency, and is useful for the businesses themselves. As Berg notes: ‘performance rankings … are catalysts for promoting critical thinking about the sources of inefficiency’ (2010, p. 54).

## 4.9 No perfect measure is possible

Benchmarking is a demanding quantitative (and qualitative) task. As in many other cases of firm-based modelling, the results are often fragile to data errors, statistical assumptions and variable choices. Prima facie, this appears to doom benchmarking as a useful tool, at least for the time being. However, this is overly pessimistic.

* Criteria (such as those above) can be used to separate poor from better benchmarking.
* It is possible to address inaccuracy and unreliability in using benchmarking (applying the loss function principles spelt out previously).
* Regardless of whether satisfactory aggregate benchmarking models can be estimated, benchmarking will often still be useful for specific performance measures (such as the efficiency of vegetation management or the use of tendering).
* Improvements in data collection are likely to improve benchmarking models.
* It may have a role in policy-relevant areas other than revenue determinations (section 4.6 and chapter 8).
* It can provide a rough test of the reasonableness of building block and revenue proposals, which could be the basis for further more detailed assessment. In that vein, in his testimony to the Commission, Bruce Mountain, representing, the Energy Users Association of Australia, pointed out that it was easy to dismiss benchmarking, but that this ignored its role of giving a *sense* of whether business proposals were right or wrong:

Benchmarking can always be criticised and shot down for some reason or the other. It's intrinsic, it's the very nature of it. … I think those who stand to lose from benchmarking comparisons will be able to mount a convincing argument that it's never quite good enough, it's always not quite satisfactory. I think that that completely misses the point. Users and consumers in all number of industries benchmark all the time. They do it crudely and it's part of the business process. People select and they choose and they make decisions. … the question arises: how is [the regulator] going to do the more aggregate assessment? That always will translate into some sort of benchmarking exercise. You can't get a sense of a big number as being right or wrong unless you actually compare it to others. I would also add, just going back to this 1999 distribution reset in Britain where Stephen Littlechild was castigated for having done such a simple benchmarking exercise. Ofgem has since then spent a great deal of money on benchmarking. … In the most recent price reset they compared the results they got using these more advanced technologies against what Stephen said back in 1999 and they found the answer was in fact jolly similar. (trans. pp. 91‑2)

1. At times, this chapter covers some technical matters in order to be useful to practitioners, but the treatment is as simple as possible. The chapter provides references to more comprehensive technical material. [↑](#footnote-ref-1)
2. Joskow (2007) provides a summary of the issues. [↑](#footnote-ref-2)
3. The terminology can be misleading. For example, OFGEM’s RPI-x mechanism (where RPI denotes the retail price index) is not much different in practice to the approach currently adopted by the AER. This is because the approach still requires an assessment of the efficiency of any given business, and then some decision about the allowable revenue over the future regulatory period based on a judgment about the pace of convergence to an efficient benchmark. Fearon (2007, p. 7) defines two types of CPI-x approaches — one based on the building blocks approach as used currently by the AER and one that calibrates x on an industry-wide TFP trend. In this report, we refer to the latter as the CPI-x approach. [↑](#footnote-ref-3)
4. IRIC (2003) explains why it is therefore necessary to have a base year adjustment or apply higher initial TFP growth rates for businesses estimated as less efficient than the static benchmark in the base year. This means that some static benchmarking would be required. [↑](#footnote-ref-4)
5. This illustration assumes a TFP rate of 1.5 per cent per annum, a discount rate of 8 per cent and an economic growth rate of 3 per cent per annum. [↑](#footnote-ref-5)
6. Bloom et al. (2007), Bloom and Van Reenen (2007), Bloom and Van Reenen (2010). [↑](#footnote-ref-6)
7. The ACCC/AER (2012a) provides a comprehensive update on benchmarking approaches. The AER also commissioned a thorough analysis of European approaches (Schweinsberg et al. 2011) as input into ACCC/AER (2012b). In this inquiry, Cunningham (sub. 28) provides an accessible treatment of the different methods and their use in incentives regulation. Filippini et al. (2005) also covers the various techniques. There are numerous software packages for undertaking benchmarking, such as the various programs from the Centre for Efficiency and Productivity Analysis at the University of Queensland (DEAP, DPIN, Frontier, and TFPIP); LIMDEP, STATA, and OpenSolver for Excel. [↑](#footnote-ref-7)
8. Kaufmann (2009) has argued that TFP benchmarking is less vulnerable to data inconsistencies than the building block process currently employed. This is true if TFP benchmarking is based on the growth rate of State-wide TFP, as had been suggested by the Victorian Essential Services Commission. Lawrence’s judgment is that it depends on the context and that TFP results can still be quite sensitive to data errors (Lawrence 2009, p. ii). [↑](#footnote-ref-8)
9. As noted by IRIC (2003, p. 18). [↑](#footnote-ref-9)
10. Several participants made suggestions about the appropriate variables. These resembled the diversity of those suggested in an international context (Ergon Energy sub. 8, p. 10; ATSE sub. 9, p. 1). [↑](#footnote-ref-10)
11. EnerNOC (sub. 7, p. 4) considered that the best measure was the proportion of the peak load that appears for 40 hours or less in a year, but acknowledged that 40 hours was a fairly arbitrary choice (with values from 10 to 80 hours all being reasonable). [↑](#footnote-ref-11)
12. The X factor is of the form X1 + X2 – X3 where: X1 is the difference between the TFP growth for the electricity distribution industry in Australia and that for the economy as a whole; X2 is the difference between the best observed opex partial productivity level in comparable electricity distribution businesses in Australia and that of Power and Water Power Networks (PWPN); and X3 is the difference between the input price growth for PWPN and that for the economy as whole (GHD Meyrick 2008, p. ii). [↑](#footnote-ref-12)
13. What ‘limited’ means is a matter of judgment. ETSA Utilities et al. (sub. 6, p. 3) indicated that the AER has used benchmarking extensively. The difference of view centres on the sophistication and function of the benchmarking, not the number of times it has been used. As the ENA (sub. 17, p. 4) pointed out ‘the key question is how the AER’s use of benchmarking can be enhanced in order to improve the accuracy of [network expenditure forecasts] (sub. 17, p. 4). [↑](#footnote-ref-13)
14. Participants in the inquiry had overlapping criteria for judging benchmarking. For example, ETSA Utilities et al. (sub. 6, p. 25) identified nine criteria: communication, consultation, consistency, predictability, flexibility, independence, effectiveness and efficiency, accountability and transparency. Ausgrid (sub. 13, p. 5) and the ENA (sub. 17, pp. 4‑5) recommended robustness, transparency, promotion of efficiency, consistency with the wider regulatory framework, reasonableness of data requirements, adaptability and resource costs. These are similar criteria identified by Frontier Economics (2010b) for OFGEM in the United Kingdom. Kaufmann and Beardow (2002) noted similar criteria, but added the importance of capturing business conditions — an issue to which this chapter returns later. [↑](#footnote-ref-14)
15. Growth rates in past demand also have implications for the asset type and scale economies (ENA, sub. 17, p. 30). A business experiencing low and steady growth will tend to (efficiently) have a large number of smaller transformers and other equipment that will need to be operated and maintained. In contrast, a business that has experienced large waves of growth will (efficiently) have a smaller number of large-scale assets that need to be operated and maintained (because large-scale assets can be built with less average underutilisation when demand is growing faster). [↑](#footnote-ref-15)
16. Some accepted that it was possible to meaningfully benchmark the weighted average cost of capital (ETSA Utilities et al., sub. 6, p. 29). The Commission considers some of the problems in this area in chapter 5, as they relate mostly to the capacity for the AER to effectively use other kinds of benchmarking in incentive regulation. [↑](#footnote-ref-16)
17. On the other hand, the MEU (sub. 11, p. 24) argued that ‘networks try to minimise the use of benchmarking on the grounds that their network is different.’ This is why statistical testing and engineering appraisal of benchmarking models to test for important omitted variables is critical. [↑](#footnote-ref-17)
18. This was reflected in submissions from Ergon Energy (sub. 8, p. 7), ActewAGL (sub. 14, p. 2), the AEMC (sub. 16, p. 2), the ENA (sub. 17, pp. 30ff), Ausgrid (sub. 19, p. 8), the EUAA (sub. 24, p. 7), Essential Energy (sub. 30, p. 4), Grid Australia (sub. 44, p. 3) and GDF Suez Energy Australia (sub. DR68, p. 4). As an illustration, the difficulties in controlling for environmental factors meant that the Brattle Group (2012a, p. 48) found there was no clear relationship between the costs of the distribution networks they reviewed and the reliability performance they achieved. [↑](#footnote-ref-18)
19. Such reporting may also be helpful in that it emphasises that managers of network businesses may sometimes be quite efficient given the policy environment they face. [↑](#footnote-ref-19)
20. As an illustration, the parameter estimates of an OLS regression on productivity or costs will be unbiased regardless of the variance of the error terms (the error terms being the measures of business inefficiency). This is not true for any given business’s efficiency measure. [↑](#footnote-ref-20)
21. This approach has been used across many industries and contexts, such as in electricity, rail freight, telecommunications, road freight, waterfront, coastal shipping, aviation and gas supply industries (BIE 1995), privatisation (Mota 2004), reliability and pricing (McLennan Magasanik Associates 2007), and international comparisons of stevedoring (PC 2003). [↑](#footnote-ref-21)
22. An observation also made by the AER in its justification for accepting substantial expenditure increases in various regulatory determinations. [↑](#footnote-ref-22)
23. In contrast, Germany has nearly 900 distribution network operators (Schweinsberg et al. 2011, p. 29). [↑](#footnote-ref-23)
24. In a comprehensive survey of the use of benchmarking methods by individual countries, Haney and Pollitt (2011) argued that small numbers of network companies act as a constraint on the use of advanced benchmarking methods. The ENA observed that the greater the level of aggregation in the analysis, the greater the number of variables that should be included in the analysis, such as the type of street transformer, the distribution of asset vintages and climatic conditions (ENA, sub. 17, pp. 24‑5). Of course, some factors that affect performance must be omitted from any benchmarking analysis given sample size limits. For example, an inefficient local government may be slow at processing planning proposals for a given network business, and no practical benchmarking approach can fully take account of such micro-factors. The extent to which a model should include variables must therefore depend on their materiality and statistical testing. [↑](#footnote-ref-24)
25. There was general opposition to international benchmarking by network businesses in both the Commission’s informal consultations and in submissions made by them to this inquiry (for example, Ergon Energy, sub. 8, p. 9). However, as shown later, they do participate in international benchmarking for commercial reasons. [↑](#footnote-ref-25)
26. For example, using ordinary least squares regression for categorical data (for example, yes/no answers, or ratings from one to five) will lead to biased estimates and a poor capacity for statistical inference. [↑](#footnote-ref-26)
27. The ENA observed that the ‘level of aggregation needs to be high enough such that material interdependencies between different expenditure types are captured in the analysis. For example, network augmentation expenditure on the sub-transmission system may be a substitute for augmentation expenditure on the distribution system and vice versa’ (sub. 17, p. 24). [↑](#footnote-ref-27)
28. So long as there is sufficient variation in the cost drivers concerned (such as varying levels of forestation, wind, and temperature). [↑](#footnote-ref-28)
29. The interpretation of such a statement is that the probability that the estimate is truly different from zero (or some other favoured null hypothesis) is 95 per cent. [↑](#footnote-ref-29)
30. A preferable approach would be to consider the likely distribution (for example, log normal) or to data-intensive methods in inference, such as bootstrapping (Varian 2005). [↑](#footnote-ref-30)
31. Berg (2010, p. 54) is a notable exception. [↑](#footnote-ref-31)