Disability Employment Services

International Literature Review for
Benchmarked assessment models and
comparison to Star Ratings

2014

Department of Social Services
Table of Contents

Acronyms .......................................................................................................................... 3  
Executive Summary ........................................................................................................ 4  
1. Introduction ..................................................................................................................... 7  
3. Comparative Performance Model .................................................................................. 11  
   3.1 External Factor Adjustment .......................................................................................... 11  
   3.2 Ex-post Performance Data ......................................................................................... 12  
   3.3 Creaming and Parking ............................................................................................... 12  
   3.4 Business Reallocation ............................................................................................... 13  
   3.5 Issues with the Star Ratings Model ........................................................................... 13  
4. Benchmarking Models ................................................................................................... 15  
   4.1 Absolute Benchmarking Model .................................................................................... 15  
      4.1.1 Target Setting ....................................................................................................... 16  
      4.1.2 External Factor Adjustment ............................................................................... 22  
      4.1.3 Forecasting .......................................................................................................... 24  
      4.1.4 Creaming and Parking ....................................................................................... 26  
      4.1.5 Certainty of Performance .................................................................................... 28  
   4.2 Negotiated Benchmarking Model .............................................................................. 29  
      4.2.1 Target Setting ....................................................................................................... 32  
      4.2.2 External Factors ................................................................................................... 33  
      4.2.3 Forecasting .......................................................................................................... 34  
      4.2.4 Resourcing Impost .............................................................................................. 35  
   4.3 Adjusted Benchmarking Model ................................................................................... 37  
      4.3.1 External Factors ................................................................................................... 38  
      4.3.2 Forecasting .......................................................................................................... 39  
5. Discussion ....................................................................................................................... 42  
   5.1 Certainty of performance .......................................................................................... 42  
   5.2 Transparency ............................................................................................................ 43  
   5.3 Adjustment for external factors ................................................................................ 45  
   5.4 Quality ..................................................................................................................... 45  
   5.5 Equity ....................................................................................................................... 46  
   5.6 Continuous Improvement ......................................................................................... 47  
   5.7 Competition and Performance ................................................................................. 49  
   5.8 Innovation ................................................................................................................. 50  
6. Conclusion ....................................................................................................................... 52  
References ........................................................................................................................ 53
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMO</td>
<td>Annual Measurable Objective</td>
</tr>
<tr>
<td>AMS</td>
<td>Arbeitsmarktservice</td>
</tr>
<tr>
<td>AYP</td>
<td>Adequate Yearly Progress</td>
</tr>
<tr>
<td>CA&amp;G</td>
<td>Comptroller and Auditor General</td>
</tr>
<tr>
<td>CESI</td>
<td>Centre for Economic and Social Inclusion</td>
</tr>
<tr>
<td>CPI</td>
<td>Consumer Price Index</td>
</tr>
<tr>
<td>CSE</td>
<td>Child Support Enforcement</td>
</tr>
<tr>
<td>DEEWR</td>
<td>Department of Education, Employment and Workplace Relations</td>
</tr>
<tr>
<td>DES</td>
<td>Disability Employment Services</td>
</tr>
<tr>
<td>DoE</td>
<td>Department of Employment</td>
</tr>
<tr>
<td>DSS</td>
<td>Department of Social Services</td>
</tr>
<tr>
<td>DOL</td>
<td>Department of Labor</td>
</tr>
<tr>
<td>DWP</td>
<td>Department for Work and Pensions</td>
</tr>
<tr>
<td>EETYF</td>
<td>Employment, Education, Training and Youth Affairs</td>
</tr>
<tr>
<td>ESS</td>
<td>Employment Service System</td>
</tr>
<tr>
<td>FND</td>
<td>Flexible New Deal</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross Domestic Product</td>
</tr>
<tr>
<td>GPRA</td>
<td>Government Performance and Results Act</td>
</tr>
<tr>
<td>IPS</td>
<td>Individual Placement and Support</td>
</tr>
<tr>
<td>ITT</td>
<td>Invitation to Tender</td>
</tr>
<tr>
<td>JSA</td>
<td>Job Services Australia</td>
</tr>
<tr>
<td>JTPA</td>
<td>Job Training Partnership Act</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>LWA</td>
<td>Local Workforce Area</td>
</tr>
<tr>
<td>NCLB</td>
<td>No Child Left Behind Act</td>
</tr>
<tr>
<td>NCLP</td>
<td>No Cure Less Pay</td>
</tr>
<tr>
<td>NCNP</td>
<td>No Cure No Pay</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Cooperation and Development</td>
</tr>
<tr>
<td>PCT</td>
<td>Primary Care Trusts</td>
</tr>
<tr>
<td>PES</td>
<td>Public Employment Services</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>Research and Development</td>
</tr>
<tr>
<td>RBA</td>
<td>Reserve Bank of Australia</td>
</tr>
<tr>
<td>RPE</td>
<td>Relative Performance Evaluation (South Korea)</td>
</tr>
<tr>
<td>TANF</td>
<td>Temporary Assistance for Needy Families</td>
</tr>
<tr>
<td>U.K.</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>WIA</td>
<td>Workforce Investment Act</td>
</tr>
<tr>
<td>WIB</td>
<td>Workforce Investment Board</td>
</tr>
<tr>
<td>WISARD</td>
<td>Workforce Investment Act Standard Reporting Database</td>
</tr>
<tr>
<td>WP</td>
<td>Work Programme</td>
</tr>
</tbody>
</table>
Executive Summary

The Star Ratings model employed by the Australian Government is the formal performance management apparatuses to assess the performance of employment services. It does this through a relative assessment of provider performance after taking account of differences in participant characteristics and labour market conditions via statistical regression.

This paper undertakes an international literature review of the use of benchmarking systems in the performance management of areas of public policy. The benchmarking models were evaluated against each of the advantages and disadvantages that are said to be associated with the current Star Ratings model.

There are three broad types of benchmarking systems that can be employed to assess provider performance; absolute, negotiated and adjusted benchmarking (Industry Reference Group, 2009).

- **An absolute benchmarking** model consists of expressing one or more simple, nationally consistent pre-determined performance measures and required performance levels.
- **A negotiated benchmarking** system is where performance targets are negotiated between the government and each individual provider, taking into account changes in economic conditions and changes in the characteristics of the participants served by the program.
- **An adjusted benchmarking** system also takes account of local labour market conditions and individual job seeker characteristics across providers. However, adjustments are calculated on a consistent basis across the national network of providers using regression analysis.

It is important to acknowledge what are recognised as the strengths of the current Star Ratings model (OECD 2012, 2013) and make an assessment of whether a benchmarking system will allow for the continuation of those strengths. Broadly speaking the identified benefits of the Star Ratings model include:

- **External Factor Adjustment:** The use of logistics regression in the Star Ratings model assists in disentangling the variation in outcomes attributable to participant characteristics and labour market conditions (not under the control of providers) from factors under providers’ control, such as effectiveness and quality of servicing. The evidence indicates that external factors have a significant effect on provider performance and that it will be difficult to account for differences in participant characteristics and labour market conditions using a benchmarking model. This is most problematic for an absolute benchmarking model that does not attempt to adjust for these external factors. The evidence though also indicates that negotiated performance standards that can adjust for external factors fail to accurately do so. Adjusted benchmarking models are the most technically sophisticated benchmarking models, using logistic regression to take into account external factors. However, there accuracy is dependent on forecasting future changes in future economic conditions and changes in the participant characteristics of the provider’s caseload.
- **Ex-post data:** The Star Ratings model compares the performances of each individual provider relative to the average performance of providers at the end of the performance period. This does not require the arbitrary selection of baseline performance data that is required to be used in setting a benchmark. As the setting of a benchmark occurs at the beginning of the performance period it therefore requires determining the type of baseline performance data to be used. The different types of baseline data include the past performance of each provider, the average performance of all providers, the average performance of the top providers and the performance of similar labour market programs. The selection of the type of baseline data used can have significant effect on the resulting performance standards estimated. Further, the use of ex post performance data does not require forecasting future economic conditions and changes in participant characteristics for the outlet or contract (i.e.
the Star Ratings model. They reflect both operational and technical concerns. This paper has also sought to assess whether the previously raised disadvantages of the Star Ratings model would be resolved under a benchmarking framework.

- **Certainty of performance:** As it is a relative model, the Star Ratings cannot provide an expected performance that the provider needs to achieve into the future. Under a benchmark model, the department would relieve providers of the risks involved in forecasting future expected performance standards, by providing certainty to providers about the level of performance they are required to achieve. Though benchmarks may provide performance certainty, it reduces the ability of the department to accurately determine which providers are performing unsatisfactorily. The evidence indicates that business reallocations are an important method for improving performance.

- **Transparency:** The technically rigorous approach of statistical adjustments to performance measures may be said to be too complicated, leaving performance analysis out of reach of providers and the public. This limits the transparency of results and hence how providers and the program itself are performing. A lack of transparency though isn’t confined to the Star Ratings model. Benchmarking models also have transparency issues because a benchmarking model still requires a methodology for determining the performance standard.

- **Controlling for other factors:** It may be unrealistic to believe that variations in provider effectiveness can be isolated by adjusting for participant characteristics and labour market conditions. This is because a regression model cannot take into account all external factors affecting provider performance. It is recognised that there will be prediction errors, but the counterfactual of using raw performance outcomes as the basis for assessment would
introduce far greater problems. Further, problems that the Star Ratings faces in adjusting for external factors are the same problems affecting benchmarking models.

- **Equity in Outcomes:** It may be argued that if it is important to place highly disadvantaged job seekers into employment, then the performance standard should be the same across all organisations, regardless of the challenges this poses to some organisation to achieve minimum levels of performance. An absolute benchmark can achieve this objective by setting the same performance standard for all providers. A review of the literature however suggests that this can lead to serious gaming issues by providers such as encouraging the creaming and parking of participants.

- **Continuous improvement:** As the Star Ratings model assesses performance against the mean as opposed to fixed benchmarks of expected performance, this allows for the possibility that all services drift towards low outcomes because providers can still maintain their Star Rating. Results show that over time outcome rates have also improved significantly, suggesting that provider performance has not been averaging down. Further, the literature suggests that the competitive effects introduced by a comparative performance model have a greater positive impact on performance than a benchmarking model.

- **Collaboration:** The promotion of competition in employment services under the Star Ratings model may reduce collaboration between providers leading to a reduction in performance of employment services over the long term. The evidence regarding competition in other public sector areas however is that higher competition leads to an improvement in sector performance.

The evidence about the challenges in the use of benchmarking performance management models internationally in areas of public policy points to the challenges that are associated with them. The effectiveness public service delivery impacts on people’s capacity to benefit from accessing and receiving assistance from these services, making it important that public sector performance is accurately assessed. The performance measurement approaches used vary, with varying levels of success. What is nevertheless evident across all the examples looked at in this international literature review is that the means to, as accurately as possible to, account of differences in operating environment (i.e. external factors) is the cornerstone that distinguish between those performance management frameworks that are successful and those that had or currently do experience substantial operating challenges.
1. Introduction

Australia’s employment services are designed to assist people into employment by working with employers, training providers and participants to find individuals sustainable employment and where necessary address participants’ vocational and non-vocational barriers to employment. The Australian Government currently invests around $10 billion over a four year cycle in employment services programs such as Job Services Australia (JSA), Disability Employment Services (DES) and the Remote Jobs and Communities Program (RJCP) (DEEWR 2013).

Services are provided to a diverse range of job seekers over the whole of Australia. The Australian Government contracts profit and non-for-profit organisations to deliver employment services across Australia. The size of this market is significant. At any point in time around 950 000 people are receiving assistance from JSA, DES and RJCP providers, and over the course of a year, 1.6 million Australians will access these services. JSA delivers services to around 800 000 job seekers at any point in time. There are currently 93 providers delivering services across 2070 geographic locations across Australia. 151 providers deliver DES across 2600 locations to around 150 000 job seekers with temporary or permanent disability, injury or health condition who require the assistance of a disability employment service, and who may require ongoing support in the workplace. These providers compete in a ‘quasi-market’ with business reallocated from poor performing providers to high performing providers.

The significant role of employment services in Australia’s labour market entails the need to ensure that contracted providers are meeting the goals of the government through maximising employment outcomes. The applicable departments work in partnership with JSA and DES providers to drive performance and continuous improvement in the delivery of quality services to all job seekers. Underpinning this is the need for a robust performance framework that accurately measures provider performance.

The purpose of this review is to examine the current Star Ratings model used by the department to assess provider performance and review potential alternatives to this model advocated within and outside of the sector. Specifically this review:

- assesses the need for performance management in Australia’s employment services market
- examines the advantages and disadvantages of the Star Ratings model
- examines benchmarking models as an alternative to the Star Ratings model
- analyses whether benchmarking models can meet the advantages of the Star Ratings model and
- analyses whether benchmarking models offers a solution to the concerns suggested by stakeholders about the Star Ratings model.

When assessing the Star Ratings model and its alternatives, it is assumed that the department continues to operate a quasi-market that providers compete in and that the performance indicators used to assess performance remain unchanged. It is recognised that some stakeholders have proposed changes to the overall structure of employment services. An examination of these proposals is outside the scope of this paper.
The next sections are outlined as follows:

- **Section 2** discusses why provider performance is measured
- **Section 3** outlines the current Star Ratings model and the reasons why it is used to assess provider performance
- **Section 4** presents the different types of benchmarking models and whether they can match the advantages of the Star Ratings model
- **Section 5** outlines what are considered to be disadvantages of the Star Ratings model and examines whether the different types of benchmarking models can overcome these characteristics and
- **Section 6** concludes the paper, summarising the main issues the Department faces in implementing a benchmarking model to assess provider performance.

Government contracting of employment services to for-profit and not-for-profit providers is not isolated to Australia. In many countries, government departments and public agencies deliver employment services through grants or contracts with other public, non-profit and for-profit organisations (Finn, 2011). In line with the approach of the Australian Government, the activities typically contracted out include the delivery of conventional labour market programs and more intensive forms of support targeted at disadvantaged groups, including for example specialist programs for people with disabilities (OECD, 2007). Contracting systems seek to drive improved employment outcomes, service delivery and cost-efficiency by applying market mechanisms that promote competition, flexibility and choice (EETYF, 1996). However, this necessitates the need to monitor the efficiency and effectiveness of service provision by providers.

Within a private market, the monitoring problem is minimised as the consumer pays the producer directly for the good or service they consume. In a contestable market, an informed consumer who is dissatisfied with a product can seek these good and services from other producers. Loss of revenue thus acts as a discipline upon producers to cater for consumer needs in a cost-effective manner. When the government is paying but not consuming these services, it is more difficult to ensure that providers promote the program goals of the Government (Prendergast, 1999, Dixit, 2002 and Marschke, 2003). One accountability mechanism for private providers employed by the government is the use of performance management systems. Such systems measure performance, has standards that indicate acceptable performance, and rewards and sanctions (which need not be monetary) organisations that exceed or fail to meet the standards (Barnow and Smith, 2004).

Measuring provider performance is completed through the use of performance indicators. Care is required in choosing performance indicators. Focus should be on outcome indicators (such as the average duration of unemployment spells) and less on process related indicators (such as the number of referrals to vacant jobs) (Grubb, 2004). As a result, the Department of Social Services (DSS) assesses the performances of providers against three Key Performance Indicators (KPIs):

- **KPI 1**: an efficiency indicator which measures the amount of time a provider takes to place a job seeker into employment.
- **KPI 2**: an effectiveness indicator which evaluates employment and education related outcomes of the provider.
- **KPI 3**: a quality indicator that assesses the quality of a provider’s service.

KPI 3 – Quality is measured separately from the DES Star Rating calculations. These measures (in Star Ratings) are employed to distinguish between better and poorer performance of a single provider contract or outlet in meeting the program objectives compared to the performance across all providers. Standards that identify minimally acceptable performance must be associated with these measures. These standards are used to determine how well the provider is performing. A performance management framework in the context of contracted employment services should have the following features:

1. provides accountability for public funding of employment programs
2. aligns with the goals of the program
3. improves the performance of the employment services system
4. creates a level playing field by taking into account differences in the labour market environments and participant characteristics
5. provides information on performance of providers to be used to help determine business reallocations
6. creates disincentives for providers seeking to manipulate (sometimes referred to as gaming) the framework.

Broadly speaking, there are two available methods in determining standards to assess provider performance; a comparative performance model and a benchmarking model. These models and their strengths and weaknesses in assessing provider performance in regards to Australia’s employment services will be analysed in the following sections.
3. Comparative Performance Model

The Department of Employment and DSS employ a comparative performance model in the assessment of JSA and DES provider performance called Star Ratings. The system of Star Ratings used in Australia is a form of adjusted performance measure based on regression analysis. Essentially provider performance is adjusted for labour market conditions and job seeker characteristics. Expected outcomes based on the regression analysis are compared to actual outcomes. Providers that have significantly better than expected performance are given a higher score.\(^1\) The Star Ratings model is used to assess provider performance because it offers the features that should exist in a performance management system as discussed above. These include:

- it is used for accountability purposes by providing the government with information on the programs relative performance of providers
- it provides a measure of effectiveness and efficiency in achieving outcomes for job seekers against the priorities set out by the Australian Government
- it improves the performance of the employment services system through fostering pseudo-competition between providers (Productivity Commission 2002)
- it creates a level playing field by taking into account the labour market conditions and participant characteristics applicable for each contract and outlet, leading to providers being compared on their true performance
- by accurately assessing provider performance it provides the department with information that helps determine business reallocations and
- it reduces the incidence of gaming responses such as parking and creaming as a providers performance rating is not negatively affected by operating in poor labour markets serving hard to place participants.

These advantages are discussed in detail in the next sections.

3.1 External Factor Adjustment

A key argument in favour of the use of Star Ratings is that it is an effective methodology for estimating a provider’s true performance (OECD 2012, 2013). Outcome measures reflect contributions from provider-related and non-provider related factors. For example, an employment outcome for a participant reflects not only the effort the provider, but the participants’ characteristics and how buoyant the labour market is. The use of regression in the Star Ratings model aims to account for non-provider related factors when comparing performance across providers.

Conceptually, outcomes represent a function of multiple interacting factors:

\[
\text{Outcomes} = f(\text{effectiveness of provider, participants characteristics, labour market conditions, random chance})
\]

The regression assists in disentangling the variation in outcomes attributable to participant characteristics and labour market conditions (not under the control of providers) from factors under providers’ control, such as effectiveness of servicing. This gives the outcome measures attributional validity, where the observed outcome differences causally relate to differences in provider performance rather than other contributing factors. It is critical to adjust for these factors before using outcomes to draw inferences about the relative efficiency and effectiveness of providers.

---

\(^1\) The methodology to calculate the stars is given in Appendix A
When these adjustments are made, the performance of a provider can be assessed relative to that of its peers without influence from the provider’s operating environment. By adjusting for labour market conditions and participant characteristics, the ranking provided through Star Ratings are much more valid than a ranking based on unadjusted performance figures (OECD, 2012).

3.2 Ex-post Performance Data

Another benefit of a comparative system is that it determines provider performance using ex-post performance data. This means there is no requirement for the setting of a benchmark that defines good or adequate performance as the ratings of providers is based on past performance. This has two benefits. First, the Star Ratings model compares the performances of each individual provider relative to the average performance of providers at the end of the performance period. In contrast, the setting of a benchmark occurs at the beginning of the performance period and therefore requires determining the type of baseline performance data to be used. For example, a benchmark could be based on the past performance of the provider or the average performance of all providers. This determination is arbitrary yet has a significant impact on the performance standard the provider is subject to. The rating of providers is as a result less robust under benchmarking relative to Star Ratings as it is highly dependent on the choice of data used.

The second benefit is that rating providers does not require forecasting changes to the operating environment which impact on provider performance. Setting performance standards at the start of the contract (or performance) period means they cannot take into account subsequent changes to the operating environment, except to the extent they are linked to specified independent economic indicators. Forecasting the level of performance required to achieve a certain Star Rating at a site level would require forecasting many variables including the nature of the participants who will inflow and changes in labour market conditions. Further, forecasts of economic indicators may differ significantly to actual future economic conditions. For instance, if the department set the outcome rates based on their expectations of the labour market, and the labour market was worse than expected, then many providers would not meet the performance standard through no fault of their own. These forecasts would introduce errors in measuring performance that are not present in the current system. This would lead to performance standards set at the beginning of the performance period becoming an unsuitable measure of performance. Conversely, the use of ex-post data allows provider performance to be adjusted based upon economic conditions that have already occurred. This improves the accuracy of the Star Ratings assigned to each provider.

3.3 Creaming and Parking

Parking refers to focusing on particular clients on the basis of performance on measured outcomes instead of value added according to the program’s stated objectives, resulting in the most disadvantaged job seekers not getting serviced (Anderson, Burkhauser, and Raymond, 1993; Cragg, 1997; Heckman, Heinrich, and Smith, 2002). In a performance based contract, providers are paid by results, based upon a range of measures. Performance indicators can include outcome, process or cost indicators. Outcome measures are used in Australia’s employment services system for examining provider performance, which have implications with regards to creaming and parking. Outcome based indicators encourage providers to maximise profits by getting the best outcomes as measured by the contract indicators. These incentive payments can encourage providers to focus on participants most likely to generate an outcome (creaming) and alternatively to provide little assistance to clients with low employment probabilities (parking). In this context, creaming refers to

2 As an example of the challenge in forecasting, note the significant revisions in the short period between the Commonwealth 2013–14 Budget on 14 May 2013 and the release of the Economic Statement on 2 August 2013.
picking participants most likely to generate an outcome after being placed with a provider, as a provider does not have the option of selecting participants.

The Star Ratings model encourages providers to focus on disadvantaged participants. A provider has an incentive to park hard to place participants, in a benchmark model, if that portion of their caseload handicaps their performance relative to other providers. In the calculation of Star Ratings, however, the actual outcome of each participant is compared to their expected performance based upon participant characteristics and labour market factors. This means that a participant in an area with poor labour market conditions has a lower expected probability of achieving an employment outcome. As a result, a provider with a caseload of hard to place participants will have a lower expected outcome rate relative to a provider with a caseload of easy to place participants. The former provider can therefore have a lower actual outcome rate and still achieve 5-Stars despite its case-mix of participants. Because the provider is not disadvantaged in the model, it acts as a disincentive for the provider to cherry pick job ready clients.

3.4 Business Reallocation

A key use of the Star Ratings is to act as a performance measurement tool for the purpose of assisting business reallocations. The purpose of business reallocation is to focus the market on continuous improvement and assist participants by moving them from low performing providers to high performing providers. In determining who the high performing providers are, it is imperative that provider performance is accurately assessed. Because the Star Ratings properly reflect the different circumstances that providers operate in, it more accurately estimates provider performance relative to raw measures. The advantage of business reallocation from a service delivery perspective is that, under the current Star Ratings system, providers are given sufficient incentive to continually improve performance in line with the policy priorities set out by the Government. A good rating is crucial for a provider to be included in the next round of tenders (or other purchasing process) and thereby stay in the market. If a provider is able to sufficiently raise performance from period to period, not only does it stand to benefit from the financial transactions associated with achieving outcomes, but it may also attract additional clients as a result of being allocated new business or participant choice.

This process results in providers competing to maximise performance, with the best performing providers receiving new business while poor performing providers exit the market. Further, because providers make profits from their business, there can be large benefits in promoting performance by using these profits to further drive performance. If future profits (through future business share) are dependent on performance in the current period (as measured by the Star Ratings), then future profits can act an incentive to promote performance over and above the payments received in the current period.

3.5 Issues with the Star Ratings Model

Conversely, stakeholders suggest that the Star Ratings model may suffer from a number of problems. These issues include:

- Providers do not have certainty over expected performance levels (Flentje, Cull and Giuliani 2010). As the Star Ratings model is a relative model, it cannot provide an expected performance that the provider needs to achieve.
- Lack of transparency over the calculations used to assess provider performance (Shulock, 1999). The Star Ratings adjusts for differences in external factors that affect the provider performance using regression analysis. This methodologically demanding process may be difficult to understand by providers and the public.
Regression models are subject to measurement error (Brooks 2000). Regression-based performance adjustment models have been criticized for having low explanatory power and flawed specifications, suggesting that sometimes adjustments may be biased or unreliable.

Adjustments for economic conditions does not provide for equity in outcomes (Barnow and Heinrich, 2010).

A high rating is not indicative of a high quality of service (Ben-Tovim et al., 2009). Differences in performance between providers may not reflect differences in quality between them but is a result of other factors affecting performance.

Does not promote continuous improvement in performance (Waghorn 2011)

For one provider to ‘succeed’ another must ‘fail’, creating competition leading to opportunistic behaviour and risk selection (Bredgaard and Larsen 2008)

Reduced scope for collaboration and cooperation among providers (Bowman and Horn 2010). A reduction in collaboration between providers may hinder the level of innovation in the sector and hence the long-term performance of the sector as a whole.

If these issues are a significant problem in Australia’s employment services, this may offset the advantages that the Star Ratings model affords. A detailed examination of these concerns and an in-depth analysis of whether benchmarking provides solutions is conducted in Section 5.
4. Benchmarking Models

An alternative performance measurement system used to assess provider performance is the benchmarking model. Cowper and Samuels (1997: 11) define benchmarking as ‘an efficiency tool based on the principle of measuring the performance of one organisation against a standard’. In employment services, this equates to setting targets for different performance indicators that a provider is expected to achieve. For example, the department may set a provider a 13-week outcome target of 20 per cent. The performance of the provider is compared to this target. This the key difference between a benchmark model and the Star Ratings model as benchmarks give providers a predetermined required performance level that they needs to achieve over the performance period. The performance of the provider is then compared to this target. There are three broad types of benchmarking systems that can be employed to assess provider performance; absolute, negotiated and adjusted benchmarking (Industry Reference Group, 2009).

- An absolute benchmarking model consists of expressing one or more simple, nationally consistent pre-determined performance measures and required performance levels.
- A negotiated benchmarking system is where performance targets are negotiated between the department and each individual provider, taking into account differences in economic conditions and changes in the characteristics of the participants served by the program.
- An adjusted benchmarking system also takes account of local labour market conditions and individual job seeker characteristics across providers. However, adjustments are calculated on a consistent basis across the national network of providers techniques, such as regression modelling.

An in-depth analysis is conducted on the three different types of benchmarking systems. This includes the issues involved in implementing a benchmarking model, whether they possess the features that a performance framework should possess as well as international case studies showing how each system has been implemented in areas of public policy.

4.1 Absolute Benchmarking Model

An absolute benchmarking model consists of expressing one or more simple, nationally consistent pre-determined performance measures and required performance levels. For example, DSS may institute a target of 40 per cent for payable 13 week full outcomes. This means that over the performance period, all providers are required to place a minimum 40 per cent of eligible participants into employment for a minimum of 13 weeks. Proponents of absolute benchmarking suggest a number of advantages in using this type of model.

- **Certainty of performance**: The key argument in favour of establishing an absolute benchmarking system is greater transparency for providers. It provides the greatest level of certainty for providers about the level of performance required to achieve certain ratings. Ultimately, the department would relieve providers of the risks involved in forecasting future acceptable performance standards. By knowing in advance what performance level is required to achieve the benchmark it allows providers to plan ahead to work towards achieving these targets.
- **Collaboration**: It allows for setting a target based on best practice, which would encourage collaboration amongst providers to share information that can improve each other’s performance in order to achieve the targets.
- **Transparency**: Under absolute benchmark systems, providers would have a better understanding of the process used to determine their performance assessment or rating.
- **Reduced Administration for providers**: An absolute benchmark is easy to apply as the performance targets are consistent across providers. This can reduce the Department’s administration costs in implementing the performance framework.

### International Case Study on Absolute Benchmarking: United Kingdom Work Programme

The United Kingdom (U.K.) Work Programme (WP) is an integrated welfare-to-work program implemented across England, Wales and Scotland by the UK Government in June 2011. It replaces a range of predecessor programs, including Pathways to Work and the Flexible New Deal (FND) (Newton et al., 2012). Similar to DES and JSA, it is a welfare-to-work program contracted through private and voluntary sector contractors (DWP, 2008).

The contract is paid through a payment model focussing on rewarding outcomes. This includes:

1. Attachment fees
2. Job Outcome payments
3. Sustainment Outcome payments and
4. Incentive payments.

In assessing the performance of the providers, the U.K. Department of Work and Pensions (DWP) employs an absolute benchmarking model. The Key Performance Measure employed to assess their performance is the job outcome rate. According to the Invitation to Tender (ITT), the job outcome rate is calculated by comparing job outcomes achieved in the previous twelve months to the number of job seekers referred to providers in the same period. The job outcome rate of each provider is compared to a minimum performance standard. The minimum performance standard sets the lowest performance that providers must achieve if they are to meet their contractual obligations. The DWP calculated the non-intervention performance level based on an analysis of historical job entry rates. The minimum performance standard is defined as the non-intervention performance level plus 10 per cent. For the first twelve months of the program (July 2011 to June 2012), the DWP expected providers to achieve a minimum performance level of 5.5 per cent for the main group of job seekers (Job seekers Allowance 25+). The performance level increases to 27.5 per cent in year two, 33 per cent in year three and 40 per cent in year 4 onwards.  

### 4.1.1 Target Setting

#### 4.1.1.1 Past Provider Performance

One of the major challenges is how to estimate an appropriate benchmark for providers. There exist a number of possible approaches for this. The first is to use the past performance of each provider as a baseline performance level and require a certain percentage increase in the benchmark over the next performance period to reflect the focus on continuous improvement. Using the past performance of providers is problematic however as it is not necessarily reflective of the organisation’s true performance. This has been demonstrated in the work of Courty, Heinrich and Marschke (2011) who developed a principal-agent model incorporating performance standards. Conceptually, this model involves the party who designs the measurement system as the principal and the party who is subject to the performance measurement system as the agent. The authors assume that the agent has some control over the performance outcome and that effort constitutes the agent’s choices and influences the performance outcome. The authors further assume that the

---

3 As outlined previously the DWP calculated the non-intervention performance level based on an analysis of historical job entry rates. This analysis has not been released by the DWP.
performance standard determines the level of acceptable performance below which sanctions are imposed. In the model, the agent receives a level of compensation independent of performance. In an efficiently functioning system, prevailing competitive forces determine the level of effort that a representative agent would expect from that base compensation. The performance standard is set at the level of performance that occurs when the agent provides the competitive level of effort.

Applying this framework, a provider’s benchmark should be measured by calculating the outcome rates of their caseload without the help from the provider, referred to as a counterfactual outcome rate. Courty, Heinrich and Marschke (2011) note that counterfactual experiments are practical in a few occupations typically involving manual work. Many work situations, however, involve non-manual work, group projects, non-standardised outputs and unstable production processes. In addition, the agent often works in multiple environments or the principal manages multiple agents with different work conditions. Australia’s employment services involve these types of work situations, making experimental and statistical studies to construct counterfactual performance benchmarks impractical (Bray, 2013).

Because of this, the approach typically used in determining the organisations base level of performance is examining what it has achieved in the past. However, determining performance standards using the past performance of each provider raises equity concerns, as poor performing providers will find it easier to improve performance relative to currently high performing providers.

This problem is highlighted in the performance management and progress reporting in relation to the No Child Left Behind Act of 2001 (NCLB) – the headline education reform package of the 2001–2009 United States Administration. The NCLB mandates that 100 per cent of American students will be proficient in English Language Arts (ELA) and mathematics by the year 2014 (Rosenberg, 2004). The system champions standards-based education; a system where high standards are set and measurable goals are defined to achieve better individual education outcomes. Rather than implementing a nation-wide performance framework, NCLB is administered by individual States which develop annual measurable objectives (AMOs) to determine whether or not a school, district or entire state is making adequate yearly progress (AYP) toward the 2014 projection (Cronin et al., 2007). States are fundamentally responsible for calculating AYP and determining the AMOs that feed into the calculation. The academic standards for what students should know and be able to do at each grade level are set out in each State’s Consolidated State Accountability Workbook. States then design or purchase a standardised testing system and establish a cut-off score for each test that indicates whether a student is deemed proficient in the subject matter. Yearly benchmarks are set out in a Timeline to Proficiency that documents a State’s intended progress toward the 100 per cent goal on a year to year basis.

A vital distinction here is that all States have different, self-calculated starting points of proficiency. Given the nature of the NCLB system, the proficiency standards and results of individual States are not directly comparable. Thus, there is no standardised benchmark of proficiency for school grades across the United States, but rather a series of different proficiency levels according to the States. Under NCLB, a student relocating from one State, in which they have been deemed proficient for the grade level, to another State with different proficiency standards, may be severely advantaged or disadvantaged when it comes to the next testing period. The outcome of allowing States to choose their own performance standards is that some States that set artificially low scores inflate their performance relative to other States that set more rigorous requirements.

To illustrate, over the 2004-2005 period the proportion of schools meeting AYP in the State of Wisconsin was 98 per cent. In contrast, the proportion of schools meeting AYP in the State of Massachusetts was 57 per cent. It is generally accepted that the State of Wisconsin has among the
lowest standards for proficiency and sets cut-off scores well below the average for all other States (Cronin et al, 2007). Thus, its schools and students do not necessarily have to perform as highly as those in other States in order to meet AYP. This is clearly reflected in its soaring percentage of schools meeting the AYP figure. Conversely, it is generally accepted that Massachusetts has the highest proficiency standards in the United States and sets cut-off scores well above average (Cronin et al, 2007). The State also provides more funding per student and has a lower student to teacher ratio than Wisconsin (PBS Newshour, 2010). The impact of the relatively more difficult standards is captured in the State’s percentage of schools meeting AYP figure. Even though Massachusetts allocates more funding and teacher interaction per student relative to Wisconsin, its higher proficiency standards mean that less students are able to meet the benchmark.

What the example of NCLB demonstrates is the challenge around setting the target and how that can influence behaviours. Beyond the autonomy for each State to set their own target (and measure their performance how they chose), the need and desire to achieve a specific performance level led to sub-optimal behaviours – relative to the intent of the public policy – on the part of the States.

### 4.1.1.2 Average Provider Performance

Another approach is to use the current average performance of all providers as a baseline performance level and require all providers to meet that target plus a certain percentage increase in performance. The principal concern here is that is does not adjust for factors outside the provider’s control in evaluating and managing performance (Deming, 1986). For example, two providers may have different performance but this may be due to differences in the labour market that they operate in. If a performance standard for providers does not take into account differences in individual participant characteristics and labour market factors, variation in provider performance relative to the performance standard will potentially be predominately attributable to uncontrollable factors.

Two concerns that arise from this are risk aversion and fairness. A benchmark that does not control for outside risks encourages creaming by increasing the probability that providers will become risk averse and refuse to help harder-to-place job seekers, as not doing so will lower their performance. Second, provider comparisons based on the benchmark can foster concerns about fairness as differences in participant characteristics and labour market factors can generate different ratings of providers who otherwise behaved identically in terms of effort and quality of assistance.

### 4.1.1.3 Best Provider Performance

An alternative method is to set a target based on the performance of the best performing providers. In this situation, for the purpose of demonstration, if it is believed that only the top 10 per cent of providers were high performing providers, then the performance standard may for instance be set at the 90th percentile of the distribution of past performance. The advantage of this approach is that it motivates poor performing providers to change their ways and aspire them to achieve higher standards of performance. It may also encourage the sharing of best practices amongst providers in order for providers to meet this demanding target.

For instance, one United States government program, Temporary Assistance for Needy Families (TANF), provided cash assistance to indigent (i.e. poor) American families with dependent children through the United States Department of Health and Human Services. In the TANF high performance bonus system, states that in the past had invested little to help clients achieve self-sufficiency had to work harder to meet performance requirements for client work participation, job entry, retention, and earnings gains (Barnow and Heinrich, 2010). The concern here again though is that no
adjustment is made for differences in client characteristics and labour market factors that affect provider performance and hence can overestimate the level of performance providers can reasonably achieve.

### 4.1.1.4 Other Labour Market Programs

A final methodological approach is to employ a target based on the performance achieved in other similar programs, either domestically or internationally. Studies examining outcome rates in other labour market programs provide a basis for setting targets in Australia’s employment services system. The concern generated here is that the performance of these other programs may not be an appropriate benchmark, due to differences in program design, performance indicators, economic conditions, demographic profile and characteristics of the job seekers in these programs.

To illustrate, Waghorn (2011) recommends that the Star Rating system be anchored to a fixed level of performance based upon high quality international studies. These studies document the performance that can be achieved by employment services when following best practices to support job seekers into employment. Waghorn, Stephenson Browne (2011) further argues that the most effective form of supported employment for people with severe mental illness as demonstrated in the literature is the Individual Placement and Support model (IPS). This model has a number of principles that should be practiced by providers to maximise employment outcomes based upon empirical evidence. This includes focusing on competitive employment, rapid job searching, intensive assistance, individualised assistance, maintaining follow-on support, integrating the employment program with mental health treatment and providing financial counselling. Bond (2004) reports that on average 56 per cent of people with mental illness receiving evidenced-based supported assistance managed to commence competitive employment relative to average outcomes of 24 per cent for other types or rehabilitation. In response Waghorn, Stephenson Browne (2011) states that ‘a key success criterion is the proportion attaining competitive employment. If this falls below the average 60 per cent reported in the research, more program development is indicated.’

Using this as an example though, there are four significant issues that arise from using such an approach to benchmark performance in Australia’s employment services. The first problem is that outcome rates of being placed into employment are not consistent across studies. Bond (2004) examined the effectiveness of evidence-based supported employment based on converging findings of the conversion of day treatment to supported employment and nine randomised controlled trials comparing supported employment to a variety of alternative approaches. The outcome rates differed substantially across studies with between 40 per cent and 60 per cent of participants receiving assistance achieving employment. Bond, Drake and Becker (2012) find that outcome rates also differed significantly across countries. Comparing the effectiveness of IPS in studies in the US to studies outside the US, the employment rate was significantly higher in the US studies than in non-US studies (62 per cent compared to 47 per cent).

The second issue is that the measurement framework employed in the literature differs to the performance indicators used by the department. Bond, Drake and Becker (2012) notes that some indicators are common across studies including measures; of job acquisition (e.g. percentage of clients obtaining competitive employment and time from study entry to first job start), duration (e.g. cumulative number of weeks worked in all jobs), intensity (e.g. percentage working at least 20 hours a week) and productivity (e.g. total hours worked/wages). However, the performance framework used in employment services is used to support program goals. There will therefore be some performance indicators that should be measured by the department that will not be identified in the literature. For these performance indicators it is impossible to know what benchmark to apply.
Third as Waghorn (2011) points out, the evidence-based practices identified (Bond, 2004; Bond, Drake and Becker, 2008) apply to volunteers not non-volunteers. These practices that work for volunteers might not apply to non-volunteers, meaning the outcome rate reported in the literature might not be meaningful. In the Australian context, universal access to, and participation requirements in, employment services means that there are significant differences between participants in Australia’s employment services and the sample of people examined in these international studies. Currently, the best example of a volunteer program in Australia’s employment services is the New Enterprise Incentive Scheme, a small business training and support program. Given that participants have to nominate for the support, and then have their small business plan approved before entry into the program, the selection processes at the start are an important contributing factor to its high employment outcomes results.4

Finally, these practices apply to individuals with a mental health condition. There are numerous other types of disadvantage that may require different types of evidence-based practices to best overcome those barriers. People with other types of disadvantages may therefore have different outcomes rates compared to what is reported in the above studies. This example shows that setting benchmarks based upon results reported in the academic literature is fraught with difficulties. This is not to say that these evidence-based practices do not have value. However, their applicability rests in informing program design, not in the performance framework that measures program success.

4.1.1.5 Continuous Improvement

In addition to determining the target, a further issue is how to ensure continuous improvement in provider performance. In performance management systems using a benchmarking model, the approach used is to add a premium to the base target each performance period determined by a subjective assessment about how much provider performance should improve (Courty, Heinrich and Marschke 2011). Setting the premium too low can result in provider performance being sub-standard over the long term. However, setting it too high can also perversely lead to reduced performance as a result of the dynamic interaction between the performance premium and the actions of providers.

With the use of benchmarking, providers will take into account the possibility of future changes in the benchmark and that current performance outcomes will be used in setting future standards. Assume for example that in the performance measurement system, a “stretch target” is set where providers have to significantly increase their performance over the period. Further assume that a provider systematically outperforms the standard and hence the standard for that provider is consequently increased. Their high performance may be because they are providing exceptional levels of effort. The provider, anticipating that their current performance influences future standards, has the incentive to purposefully under-perform in order to ensure less stringent future performance targets (Bouckaert and Halligan, 2010). This phenomenon is known as the ratchet effect (Holmstrom and Milgrom, 1987; Miller, 1992).

The way to eliminate the ratchet effect is to not change the standard, or to commit to strict rules for changing the standard, i.e. adjusting the benchmark to take into account changes in economic conditions. Not adjusting the standard each performance period however negates continuous improvement in provider performance. Target setting thus becomes a complex process that involves technical calculations of performance improvement that are both achievable and challenging.

---

4 In the Labour Market Assistance Outcomes report, the New Enterprise Incentive Scheme consistently reports around an 80 per cent employment rate.
Case Study: Estimating An Absolute Benchmark Under The Work Programme

The WP highlights the challenge in ensuring that the performance targets are set at an appropriate level that can be reasonably achieved by providers. This is important, as providers’ contracts can be cancelled due to underperformance. A benchmark set too high can threaten the viability of the program as an inordinate number of contracts get cancelled, as well as being unfair to providers who have had their contracts removed due to overly ambitious performance expectations.

In the ITT, the DWP made clear that it may cancel contracts of provider who do not meet the minimum performance standards. Given that contracts can be cancelled due to underperformance, Mulheirn (2011) examines the appropriateness of DWP’s performance standards and hence the viability of the scheme through forecasting the likely performance of the providers during the first three years. This is done through examining the actual performance achieved under FND, which has strong similarities in the job seekers it served and the way providers were renumeration.

FND operated over the period October 2009 to May 2011, with performance data available up to April 2011. Using actual performance data, the author forecasts the performance of the FND over the period of the WP based on forecasted referral numbers and assesses that forecasted performance against the minimum performance standards for the ‘JSA 25 and Over group’ in the WP.

The results show that the predicted performance of the WP over the first three years of the program lags significantly behind the minimum performance expected by the DWP. Over the three years, expected performance is 4.1 per cent (Yr 1), 20.5 per cent (Yr 2) and 27.8 per cent (Yr 3). This compares to the required performance standard over the same period of 5.5 per cent, 27.5 per cent and 33 per cent, respectively. This suggests that the performance benchmarks have been based on unrealistic performance benchmarks that providers are unlikely to achieve. This evidence indicates the significant problem in estimating a reasonable benchmark that seeks to improve provider performance and ensure that a provider’s targets are nevertheless achievable.

Performance results from the first year of the program show that no provider has currently met the minimum performance standard. Using the DWP’s methodology to calculate the minimum performance standards, the Comptroller and Auditor General (2012) estimates the minimum performance level to be 9.7 per cent for the period 1 June 2011 to 31 July 2012. Over this same period, a total of 31,000 job outcome payments were paid to WP providers, equivalent to 3.2 per cent of individuals referred to the programme in the three main groups achieving a job outcome. The outcome rate for all participant groups is 3.5 per cent. The figure below shows the performance of the 18 prime providers relative to the minimum performance level over the period 1 June 2011 to 31 July 2012. The best performing provider moved 5 per cent of people off benefit and into sustained employment, the lowest performing provider managing only 2.2 per cent (C&AG, 2012). This is significantly below the required 9.7 per cent outcome rate providers were required to achieve.
4.1.2 External Factor Adjustment

One challenge with an absolute benchmark is that it cannot single out program outcomes that are generated by providers’ efforts from outcomes that are generated by providers’ environments. This potentially penalises providers for subpar performance that is as a result of the provider’s environment. If government policies are to be assessed through evidence of their effectiveness, then it is imperative that Australia’s employment services system statistically models the relationship between the activities of providers and performance outcomes. This necessarily entails that factors that influence outcomes, but are not controlled by providers, are adjusted for in the performance framework. By adjusting for these external factors, performance estimates are more likely to reflect the provider’s true performance.

Brooks (2000) examines whether adjusted performance measures more accurately reflect true performance relative to unadjusted measures. The author finds that estimates of performance may be distorted unless the specification is very accurate. For policy makers however an adjusted model estimates can still be an improvement over an absolute benchmark even if estimates are biased. Bartik, Eberts and Kline (2004) develop an adjustment model for workforce performance standards for local workforce areas (LWAs) in the US which they suggest provides a better estimate of value.
added. The authors contend that even if the model is not perfectly specified, “the relevant issue is whether these estimates are still closer to the true relative value added than the estimates one would obtain by simply comparing LWA means on the common measures” (2004, 8). This is an observation also made by the OECD when looking at Australia and the Star Ratings model (OECD 2012, 2013).

The evidence indicates that adjusting performance across providers has even greater importance for people with a disability. It has been argued that if people have an impairment making them unable to work or persuade an employer to hire them, then they would always be disadvantaged regardless of the economic conditions in the place they lived. If this situation exists then differences in labour market conditions will make little difference to their probability of finding employment. Berthoud (2011) finds that in Britain, the employment rate of disabled people is highly sensitive to differences in labour demand across regions. Using data from the General Household Surveys over 1996 to 2005, the authors estimate a logistics regression of employment probabilities of the general population. The control variables of interest include the average unemployment rate across regions, a dummy variable that takes the value of one if the person has a disability and zero otherwise and an interaction term between the unemployment rate variable and the dummy variable. The authors model the interaction between the unemployment rate and whether a person is disabled. They find a significant and negative effect of this interaction term on employment probability, suggesting the employment probabilities of disabled people are strongly affected by regional variations. Further, the effect of regional variations in labour demand on the employment probabilities of disabled people is greater relative to non-disabled people. These results suggest that the importance of adjusting for differences in economic conditions is even higher for providers catering to job seekers with a disability.

Case study: Risk adjustment and the relative performance of hospitals

Research on hospital performance provides important insights for employment services. Similar to providers serving clients with different characteristics in different labour market environments, the patients served by hospitals differ significantly in terms of their characteristics and socio-economic background. As stated by Iezzoni (2009 p253-245)

‘Health plans, hospitals, general practitioner practices or other healthcare providers are not selected randomly. Many factors affect the way people link with their sources of care, including the nature of their health needs (e.g. acuity and severity of illness); financial resources; geography; previous health-care experiences; and their preferences, values and expectations of health services. Not surprisingly, there may be wide variations in the mix of persons covered by different health plans, hospitals, general practitioner practices or other healthcare providers. These differences can have consequences…..Most importantly from a quality measurement perspective, persons with complex illnesses, multiple coexisting conditions or other significant risk factors are more likely to do poorly than healthier individuals, even with the best possible care.’

As a result, risk adjustment has been used to adjust hospital performance for variations in patient characteristics. The evidence shows the importance of adjusting for factors outside the control of hospitals when assessing their performance. For instance, Pok et al. (2011) examine the results of operative treatment of patients with colorectal carcinoma within the context of hospital benchmarking. The results of resection of ontological surgery such as colorectal carcinoma can vary greatly from one hospital to another. However, this does not necessarily reflect differences in the quality of treatment. The patient’s demographic details, their general state of health (comorbidities) and the stage of the patient’s tumour at the time of diagnosis can vary greatly between groups of
patients. The authors examine the use of risk adjustment in comparing treatment outcomes between hospitals after colon carcinoma resection. To do this, a survey questionnaire was completed where the patient was treated that recorded information about the patient, the tumour, the operations and the postoperative course to discharge from the hospital. Three hundred and forty-six hospitals and 48,894 patients with colon carcinoma participated in the survey. The primary variable of interest was in-hospital mortality. Of the participating institutions, two hospitals were selected based on a significant difference in postoperative mortality and postoperative morbidity. A univariate comparison of hospitals A and B showed a statistically significant difference (at the 5 per cent level) for postoperative mortality, being 2.9 per cent for A and 6.4 per cent for B. This would suggest that the quality of surgical treatment of patients with colon cancer is significantly different across the two hospitals.

The authors then examined if the hospital and patient location affected the performance of postoperative mortality after risk adjustment. In a univariate comparison, patient populations differed significantly in terms of age, body mass index, American Society of Anaesthesiologists score, the frequency of pulmonary and renal concomitant disorders and the proportion of patients with insulin-requiring diabetes mellitus. No significant difference was found with respect to tumour-related variables. In assessing the effect of risk adjustment on hospital performance, logistic regression, propensity score analysis and the CR-POSSUM score was used. The results show that the significant difference in postoperative mortality in the univariate comparison did not hold under the risk adjusted analyses. This suggests that raw scores can give misleading conclusions of provider performance if factors outside a provider’s control are not taken into account.

4.1.3 Forecasting

The suitability of an absolute benchmark is affected by unexpected changes in economic conditions. To illustrate, assume that a target of 50 per cent is set for the job outcome rate over the next year. The calculation of this target will depend on expectations of economic growth, unemployment and other macroeconomic factors affecting the probability of job outcomes. Changes in economic conditions different to what is forecasted however will affect job outcome rates and hence the appropriateness of the benchmark. The worst case scenario is a situation in which labour market conditions deteriorate significantly in the performance period leading to poorer outcomes for affected providers in the area. Such sites would be disadvantaged by having their actual outcome rates compared to predicted outcome rates that were forecast in times of more favourable labour market conditions.

Prior research suggests forecasting is subject to significant uncertainty. For instance, each quarter the Reserve Bank of Australia (RBA) releases its forecasts for key economic variables in its Statement of Monetary Policy. Tulip and Wallace (2012) examine the uncertainty around the Reserve Bank of Australia’s (RBA) forecasts. The authors evaluate the RBA’s forecasts through constructing confidence intervals for forecasts of CPI inflation, underlying inflation, relative GDP growth, and the unemployment rate. The estimated confidence intervals span a wide range of outcomes, indicating significant uncertainty regarding the forecasts. Further, RBA forecasts have significant explanatory power for inflation over the first forecast year but do not significantly explain variations in GDP growth or medium-term changes in unemployment.

Forecasts of economic conditions offer particular challenges for disability employment services. This is because the effect of future economic conditions on employment rates of people with disabilities is itself affected by government policies. Burkhauser, Daly and Houtenville (2002) examine the employment rate and household income of the working age population (25-61 years) with and
without disabilities over the business cycles of the 1980s and 1990s in the United States. The authors find that the employment rate and household income of working age people with disabilities are more adversely affected by economic downturns relative to those people without disabilities. This contrasts to their findings regarding economic expansions, where employment and household income rose for both groups during the 1980s but only increased for working age people without disabilities during the 1990s. Stapleton and Burkhauser (2003) review the evidence examining the reason for the decline in the employment rate of the 1990s and conclude that it was a result of public policies implemented in the late 1980s to early 1990s. This is due to the evidence showing that the decline occurred across all demographic, education and health groups. The change in public policy was an expansion of the Social Security Disability Insurance and Supplemental Security Income programs. This included both eligibility expansions and expansion in the value of benefits relative to wages for low-wage workers. These results suggest that estimating a benchmark for disability employment services has the challenge of both forecasting future economic conditions and determining structural breaks in the relationship between employment rates for people with disabilities and future macroeconomic factors.

**Case Study: Forecasting Economic Conditions in The Work Programme**

Another issue that the current performance of the WP highlights is that unexpected changes in economic conditions can alter the suitability of the absolute benchmark currently in place to assess provider performance. As pointed out by CESI (2012), official growth forecasts are integral to the design of the WP because the funding of the WP is determined by outcomes. When the expected performance levels and the size of payment outcomes were being set in December 2010, the DWP and bidders (potential providers) needed to assess the future state of the economy in order to determine the likely level of job outcomes that could be reasonably achieved. As seen from growth forecasts by the Office for Budget Responsibility (OBR), growth expectations differed significantly to actual GDP growth. OBR forecasts for GDP were 2.1 per cent in 2010 and 2.6 per cent in 2011 and 2012, respectively. This compares to significantly lower actual GDP growth of 1.8 per cent in 2010, 0.8 per cent in 2011 and 0.3 per cent in 2012. The change in economic conditions has a significant effect on the appropriateness of the WP’s benchmarks. CESI (2012) document a significant positive correlation between GDP growth and job entry rates. They estimate a simple linear regression model of the form:

\[ P_t = \alpha + \beta X_t + \epsilon_t \]

where \( P_t \) is the Jobcentre Plus outcome rate and \( X_t \) is the annual change in GDP. CESI (2012) find that the coefficient estimate \( \beta \) is approximately 5.2, meaning that annual changes in job-entry rates are circa 5.2 times the change in GDP. For instance, if the job entry rate for long-term JSA claimants is 40 per cent and GDP falls by one per cent, there is a reduction in the job entry rate to 38 per cent. Although a simple model, the regression explains almost 75 per cent of the variation in job entry rates for long-term claimants. The positive relationship between GDP growth and job entry rates and the associated reduction in GDP growth relative to that forecasted lead the CESI (2012) to conclude that the minimum performance standards of the DWP were no longer applicable to appropriately assess provider performance. They estimate that the minimum performance standards for the three main groups need to fall by 15 per cent to accurately take into account the change in GDP growth. This highlights a significant problem with an absolute benchmark: it does not take accurately take into account external factors affecting provider performance.
4.1.4  Creaming and Parking

Another issue with absolute benchmarks is that it may encourage program operators to engage in parking. This refers to focusing on particular clients on the basis of performance on measured outcomes instead of value added according to the program’s stated objectives, resulting in the most disadvantaged participants receiving minimal service or not getting serviced (Anderson, Burkhauser, and Raymond, 1993; Cragg, 1997; Heckman, Heinrich, and Smith, 2002). Under absolute benchmarking, creaming is encouraged because performance outcomes are based on success. The difficulty in meeting the measured performance target is reduced when servicing clients with low employment barriers, allowing providers to maintain their relatively higher performance. Furthermore, providers are constrained by environmental conditions that interfere with outcomes (e.g. a high unemployment rate or an economically disadvantaged population), and thus providers in difficult employment areas would find it more necessary to target clients with good job prospects.

This is illustrated in the significant incidence of creaming during the beginning of the Job Training Partnership Act (JTPA), the primary workforce development program in the US from 1982 to June 2000. At the start of the program, many states used the secretary of labour’s performance standards without adjustments. By the 1990s, a majority realised that failing to adjust for economic conditions and participant characteristics lead to the enrolment of the easiest-to-place individuals (Barnow, 1992).

Courty, Heinrich and Marschke (2011) empirically investigated whether in the JTPA adjusted performance measures for different socio-economic subgroups of enrollee influence case workers’ choice of intake population. Under JTPA, job training services were administered by over 620 semi-autonomous sub-state training agencies each evaluated according to a set of performance measures. Job training agencies in this program are rewarded for improving the labour market performance of the clients they serve but the reward function also depends on the enrolment choice. A training agency enrolling less able applicants has to meet a lower level of performance, effectively setting a system of shadow prices that correct for the challenge that each demographic subgroup presents.

If for the purpose of demonstration it is assumed that an agency enrolled just two demographic subgroups, this results in a an adjusted performance model of the type

\[ M_0(X_j) = m_0 - (\beta_i X_i + \beta_j X_j) \]

where \( M_0 \) is the adjusted performance standard, \( X_i (X_j) \) is the per cent of the subgroup j enrolled at the agency and \( \beta_j \) is the adjustment weight for demographic characteristic j. If the adjustment factor is positive for subgroup j, then the agency is more likely to receive an award, ceteris paribus, if it enrols more of that subgroup. The authors find that increases in the incentive for enrolling members of a subgroup significantly increase the fraction of enrollees from this subgroup. Adjusting the performance benchmark reduces the incidence of creaming by ensuring that providers do not just focus on easier-to-place clients across multiple subgroups. Providers under Australia’s employment services do not have the ability to choose which job seekers are referred to them. However, it is possible for them to focus on the easiest-to-place job seekers they are managing. In this situation as suggested by Courty, Heinrich and Marschke (2011), an absolute benchmarking model does not act as a disincentive to stop this situation from occurring.

Koning and Heinrich (2010) examine the impact of 100 per cent-performance contingent reward schemes on provider job placement rates. The authors utilise a unique dataset on Dutch cohorts of unemployed and disabled workers that were assigned to private welfare-to-work providers in
performance-based contracts over 2002-2005. Over this period, procured contracts gradually moved from No Cure Less Pay (NCLP) to No Cure No Pay (NCNP) reward schemes, the latter of which constituted a fully (100 per cent) performance contingent contract, with payments made only for clients placed in jobs. The selection of easier-to-place clients by providers are expected to increase if payments or rewards are performance-based, as workers with bad a priori job prospects will increase the risk of no (or lower) payments (Heckman, Heinrich and Smith, 2002). Their results show that incentives are most likely to work when the financial risks of performance-contingent payments for providers are not too large to induce creaming behaviour. That is, for larger cohorts where the risk of failing to place clients in jobs is spread over a broader client base (i.e. the per-client risk is smaller), the authors find less of an effect of NCNP incentives on creaming. Conversely, for smaller contracted cohorts with greater risks of non-payment due to performance, the evidence points to greater levels of creaming. This indicates that the significant financial risk associated with failing to meet an absolute benchmark (i.e. non-renewal of the provider’s contract) may induce providers to largely ignore harder-to-place clients.

Case Study: Creaming and Parking in the Work Programme

The WP highlights the difficulty in an absolute benchmarking model restricting incidences of creaming and parking. Initial evaluations of the WP suggest that creaming and parking is occurring within the program. Comparing the performance of the Employment Support Allowance groups (which includes claimants with disabilities) with those on Job seeker’s Allowance (representing easier to help groups), show that the outcome rate of the Employment Support Allowance groups is further below the DWP’s performance expectations relative to the Job seeker’s Allowance Category. It is expected that the outcome rate of the group with greater barriers to employment will be low relative to the group with fewer barriers to employment. However, the DWP takes this difference into account by having different minimum performance levels across the two groups. Results show that the best performing Employment Support Allowance group achieved 20 per cent of the expected outcome rate set by the DWP, whereas the worst performing Job seeker’s allowance category achieved 24 per cent (C&AG, 2012).

DWP has put in place a differential payment regime where the level of payments made will vary between customer groups, based on the benefit they are receiving when they enter the programme, which is deemed to reflect the barriers they face to work. This is designed to restrict the incidence of creaming and parking. However, the DWP published preliminary findings from its first phase of qualitative research as part of its evaluation of the WP, which concluded that ‘the available evidence to date suggest that providers are engaging in creaming and parking, despite the differential payment regime’ and that ‘those assessed as hardest to help are in many cases left with infrequent routine contact with advisers, and often with little to or no likelihood of referral to specialist (and possibly costly) support, which might help address their specific barriers to work.’ In addition, the report noted that ‘some providers at least, took the view that (perhaps surprisingly, given the design and remit of the Work Programme) that it was inappropriate for the hardest to help to be referred to their services at all’ (Newton et al., 2012).

A number of measures exist to help reduce the incidence of creaming and parking in the WP including the use of differential payments. Other measures are the use of minimum service standards (though these are set by providers) and contract design to ensure providers cannot remain profitable unless they help claimants beyond those that could be categorised as those easy to help (Committee of Public Accounts, 2012). Interestingly, the latter measure is reminiscent of the JTPA examined by Courty, Heinrich and Marschke (2011), where there was an incentive to enrol more people from the same cohort, which was found to reduce the incidence of creaming. The absolute benchmarking model works against these measures however. Although minimum
performance standards vary across cohorts, there is the risk that creaming and parking will occur within the groups. This is because there is significant variation in characteristics within those groups that are not taken into account when assessing provider performance against those benchmarks. Thus, providers have an incentive to cherry-pick the easiest-to-place job seekers within each group to maximise its performance relative to the benchmark, leading to harder-to-place job seekers and the providers catering to these individuals being disadvantaged.

### 4.1.5 Certainty of Performance

A final point is that absolute benchmarking may be used as a relative model in reallocating business. If an absolute benchmark was used in business allocation then some providers would lose and some would gain business based on that benchmark. If such a benchmark did not properly reflect the different circumstances that providers operate in (as the current Star Ratings do) then such a benchmark may cause more, not less concerns, because it would less accurately reflect performance. As stated before in regards to the WP, a provider that performs below the minimum threshold is in danger of losing the contract from the beginning of June 2013. From June 2013 thereafter, the DWP’s contracts with the providers allows them to switch referrals from the poorer performing provider to the better performing provider in each of the 18 regions the WP is divided into. A determination therefore needs to be made as to who is the better performing provider. In the current Star Ratings model, top performing providers receive 5-Stars based on their measured outcomes after taking into account a number of factors that may affect their performance. Using an absolute benchmark, this would be determined using raw outcomes.

The use of raw outcomes in determining the relative performance of providers has two drawbacks. First, although an absolute benchmark tells providers the minimum requirement they need to achieve, it does not provide any guidance as to how well a provider needs to perform in order to increase their business share or receive new business. Second, the provider who receives these referrals (i.e. greater business allocation) will presumably be the one with the best outcome rate. Theoretically, a provider may be performing at a high level but be poorly rated based on raw outcomes as a result of external factors, such as their client mix consisting of a high proportion of hard to place job seekers and that they are operating in an area with difficult labour market conditions. At the same time, a provider may be performing at a low level yet be rated highly on a raw outcomes basis due to external factors positively affecting their performance. From this, a perverse situation may arise where a highly performing provider loses their contract, with referrals switched to the subpar performer, as factors affecting provider performance are not taken into account.

### Case Study: England’s National Health Service

England’s National Health Service (NHS) illustrate the problems in using benchmarking to compare provider performance. Significant problems were claimed to plague the NHS in the 1980’s, due to its hierarchical structure and lack of incentives (Enthoven, 1985). Enthoven (1985) recommended the introduction of incentives by requiring providers to compete in an internal market. This internal market based on the principal of provider competition was implemented between 1991 and 1997, where hospital performance information was made available to the public. The perception was that patients would act as consumers, shifting from low performing to high performing providers, forcing hospitals to improve performance to attract clients. The evidence indicated however that this did not occur with waiting times remaining a significant problem (Marshall, Shekelle, Davies, and Smith, 2003; Fung, Lim, Mattke, Damberg, and Shekelle, 2008). As a result, a new policy was introduced in 2000 with the goal of reducing waiting times for patients. This was completed though a system of Star Ratings, applying to NHS trusts over 2001 to 2005. Under this system, a star between zero and
three was given based on performance against ‘key targets’ and performance indicators in a balanced scorecard. The key targets reflect the minimum standards that all organisations are expected to achieve. At its introduction, the targets comprised various waiting time indicators, hospital cleanliness, cancelled operations, the financial state of the trust and the demonstration of a commitment to improve working conditions. Three sets of subsidiary indicators were also assessed, including those with a clinical focus (such as emergency re-admission rates); those with a patient focus (such as the resolution of written complaints) and those with a staff focus (such as junior doctors’ hours). An example of a key target was the target for category A calls (life threatening emergencies) applying to ambulance trusts. The target was that at least 75 per cent of calls be met within 8 minutes. The performance of NHS trusts against the targets lead them to be ranked in the following manner:

- trusts with the highest levels of performance in the measured areas are awarded a rating of three stars
- trusts with mostly high levels of performance, but not consistent across all measured areas, are awarded a rating of two stars
- trusts where there is some cause for concern regarding particular areas of measured performance are awarded a rating of one star
- trusts that have shown the poorest levels of measured performance or little progress in implementing clinical governance receive a rating of zero stars

NHS trusts were penalised or rewarded based on their performance relative to their benchmarks (Beverley and Haynes, 2005). Organisations that failed against key targets were zero-rated and named and shamed as failing, with their chief executives at risk of losing their job. Those that received three stars were rewarded by being publicly celebrated for being high performing and given earned autonomy, which provided them with more operational freedom and reduced monitoring (Bevan and Hood, 2006). However, failing to take into account local variations in the socio-economic composition of the populations served by the trusts created problems in assessing the comparative performance of trusts. For instance, four key targets in the focus area improving health for primary care trusts (PCTs) were; mortality rates from circulatory diseases, accidents, cancer and rates of teenage pregnancy (Bevan, 2006). Assessing performance based upon raw outcomes would mean that PCTs in affluent areas would have fared better than poorer areas. Star Ratings sought to reduce the influence of socio-economic effects on performance by examining the change in outcomes across these four targets. This seemed to have little effect on ranking though, with all PCTs in areas of high material deprivation all awarded one star (Commission for Health Improvement, 2004).

4.2 **Negotiated Benchmarking Model**

A negotiated benchmarking system is where performance targets are negotiated between the government and each individual provider, taking into account differences in economic conditions and the characteristics of the participants served by the program. There exist a number of advantages to a negotiated benchmark. First, providers know how well they have to perform in order to have the contract renewed. Second, it allows for labour market factors, job seeker characteristics and expected economic conditions to be taken into account when determining the performance standard. Third, it allows for shared accountability, in which the provider is aware of how its performance target was set and can take greater ownership of the performance required as they agreed to it. Finally, it allows for setting a target based on best practice, which would encourage collaboration amongst providers to share information that can improve each other’s performance in order to achieve the targets. The case study below of the US Workforce Investment Act provides a detailed account of the negotiation process involved in setting performance standards.
In 1998, the US Congress passed the Workforce Investment Act (WIA) to unify a fragmented employment and training system and to better serve job seekers and employers. The Act required states and localities to bring together most federally funded employment and training services into a single system, called the one-stop centre system. Three of the programs providing services to adults, dislocated workers and youth, replace those previously funded under the Job Training Partnership Act (JTPA). Performance indicators are required to be produced by states and localities to show the effectiveness of the three programs. States are held accountable for performance, possibly suffering financial sanctions should they fail to meet their expected performance level, or receive incentive grants (additional funds) should they meet or exceed their benchmark. WIA is designed to provide for greater accountability than the accountability provided for under JTPA. The performance data collected from the states in support of the measures are intended to be comparable across states in order to objectively determine incentives and sanctions. They are also intended to provide information to support the U.S. Department of Labor’s (DOL) performance goals under the Government Performance and Results Act (GPRA). GPRA was intended to focus government decision-making, management, and accountability on the results and outcomes achieved by federal programs. These performance indicators measure program results in the areas of job placement, employment retention and earnings change, as well as skill attainment and customer satisfaction. The States are required to calculate 20 performance measures. Table outlines the measures for each of the three programs.

**Table 1: Performance Measures for the Three WIA-Funded Programs**

<table>
<thead>
<tr>
<th>WIA funding stream</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>1. Entered employment rate</td>
</tr>
<tr>
<td></td>
<td>2. Employment retention rate</td>
</tr>
<tr>
<td></td>
<td>3. Average six month earnings</td>
</tr>
<tr>
<td></td>
<td>4. Employment and credential rate</td>
</tr>
<tr>
<td>Dislocated worker</td>
<td>1. Entered employment rate</td>
</tr>
<tr>
<td></td>
<td>2. Employment retention rate</td>
</tr>
<tr>
<td></td>
<td>3. Average six month earnings</td>
</tr>
<tr>
<td></td>
<td>4. Employment and credential rate</td>
</tr>
<tr>
<td>Older youth (age 19-21)</td>
<td>9. Older youth entered employment rate</td>
</tr>
<tr>
<td></td>
<td>10 Older youth employment retention rate</td>
</tr>
<tr>
<td></td>
<td>11. Older youth earnings Change</td>
</tr>
<tr>
<td></td>
<td>12. Older youth credential rate</td>
</tr>
<tr>
<td>Younger youth (age 14-18)</td>
<td>13. Younger youth skill attainment rate</td>
</tr>
<tr>
<td></td>
<td>14. Younger youth diploma or equivalent rate</td>
</tr>
<tr>
<td></td>
<td>15. Younger youth retention rate</td>
</tr>
<tr>
<td>Customer satisfaction</td>
<td>16. Participant American customer satisfaction index</td>
</tr>
<tr>
<td></td>
<td>17. Employer American customer satisfaction index</td>
</tr>
<tr>
<td>Wagner-Peyser</td>
<td>Performance Measure</td>
</tr>
<tr>
<td></td>
<td>Entered employment rate for total exiters</td>
</tr>
<tr>
<td></td>
<td>Employment retention rate for total exiters</td>
</tr>
</tbody>
</table>


---

5 The Government Performance and Results Act (GPRA) of 1993 mandated the development of outcomes-based performance measurement systems in federal agencies, including annual performance plans specifying quantitatively measurable goals and levels of performance to be achieved and annual reports comparing actual performance with goals.
Furthermore, Section 136 of the WIA requires that specific core indicators of performance and the levels of performance for these indicators be set for both states and local areas through a negotiation process. States, in turn, must negotiate performance levels with each local area. According to the regulations, these negotiations entail (U.S. DOL, 2012):

- State performance targets must be expressed in an “objective, quantifiable, and measurable” form and show the progress of the state toward continuously improving performance.
- Negotiations must take into account five factors: how the benchmark compares with the performance benchmarks of other states, differences in economic conditions, participant characteristics, services provided and how the benchmark promotes continuous improvement.
- In determining performance targets, there are a number of steps in the negotiation process to determine performance targets. The U.S. DOL recommend a number of tools and resources states should review prior to negotiations to ensure these factors have been considered in determining their proposed level and that there is a sound rationale for their proposed levels of performance.
- The U.S. DOL advises states to use their historical, annual performance information to inform projected levels of performance.
- In addition, the states should be aware of the GPRA goals of the U.S. DOL. These are the performance goals established at the national level across all 20 measures, based upon the estimated effect of unemployment rates on program performance outcomes and the Office of Management and Budget’s (OMB) assumptions about future unemployment rates and the Consumer Price Index-Urban Consumers.
- The WIA encourages states to serve those individuals with barriers to employment and individuals more at-risk of not connecting to the labour market, including those who were formerly incarcerated, the homeless, Veterans, individuals with disabilities and out-of-school youth. For instance, Section 112 (b)(17)(A)(iv) requires that the State Plan describe how the state will serve the employment and training needs of dislocated workers, low-income individuals, individuals training for non-traditional employment and other individuals with multiple barriers to employment (including older individuals and individuals with disabilities). Section 134(d)(4)(E) requires that priority must be given to recipients of public assistance and other low-income individuals for intensive services and training services. Performance targets can be adjusted to accommodate states serving a high number of individuals with significant barriers to employment who need greater servicing to achieve a positive outcome. During the negotiation process, states must provide data to support a target adjustment based on the number of individuals with barriers currently being served.
- The next step in the process of reaching an agreement on state performance levels is for the State to submit the proposed levels to the Regional Administrator serving the state. The state is expected to provide the methodology used for developing the proposed levels of performance and how the target will promote continuous improvement in state performance. The adjustments made by the state to its proposed performance standard can be highly subjective and dependent on the regional office accepting the adjustment procedure. The adjustment methodologies employed are also not consistent across states.
- The regional office reviews the analyses used by the state to develop the proposed performance levels and works with the state to set mutually agreed upon levels of performance. Regional offices will take into account the environmental factors addressed by the state, including current and future economic conditions. The regional office will consider the proposed levels in light of previously negotiated goals, past performance results, and the national GPRA goals.
The negotiations focus on whether each performance level appears appropriate and the adequacy of any information the state offers to substantiate each level. If regional offices determine through their analysis that a state could increase its proposed performance levels to more fully support continuous improvement and customer satisfaction strategies, they will negotiate with the state to obtain higher mutually agreed-upon performance levels.

### 4.2.1 Target Setting

Estimating a negotiated benchmark faces similar difficulties to setting absolute benchmarks (see section 4.1.1) in determining the base level of performance from which the negotiation process begins. That is, the level of performance that would occur under a competitive level of effort not subject to external factors outside the control of providers. These factors include determining whether to use the past performance of each provider as a baseline performance level, the average performance of all providers, the average performance of the top providers and the performance of similar labour market programs. A further complication under a negotiated standard is that there may be a significant level of discretion in the negotiation process as to what performance information should be applied in setting the base level of performance.

### Case Study on Negotiated Benchmarking: US Workforce Investment Act

Under the WIA, the negotiation process between states and localities saw the process for setting performance standards using past performance information vary widely. Among the majority of states using baseline performance data to determine performance standards, there was considerable difference in the source of data used (Heinrich, 2004). This included using the projected national averages for the negotiated standards provided by the DOL (based upon the experiences of seven early implementation States), federal baseline numbers (available in the federal performance tracking system), unemployment insurance data and the state’s own performance data taken from previous program years. For example, the state of Georgia used their 1998 state performance records and the projected national averages to determine their performance targets. Indiana used its 1999 performance data in setting its standards. In contrast, other states such as New Hampshire and Ohio employed unemployment insurance data from 1994-97 and performance data from DOL to determine their performance standards. Wisconsin reported using its 1997 performance data and the seven-state projected averages in negotiations to set the standard. Further, when Wisconsin’s base level performance was above the projected national averages, the state employed the national average as the benchmark. When their baseline performance was below the national average, the baseline values were set as the benchmark. Washington, Nebraska, South Carolina and others followed a similar process. Texas, Maryland and the District of Columbia reported using statistical models to determine their performance standard. Arkansas set its standard one percentage point above the national averages in the first year. Three states set their performance standards below the national goals. North Carolina for example used its 1997 baseline data in determining its performance standards, with all standards being below the state’s baseline data and national goals. Heinrich (2004) reports descriptive information on the level of negotiated performance standards. The author reports significant variation in the performance standards, indicated by the significant standard deviation of the performance standards and the large difference between the minimum and maximum levels of performance reported across states.
4.2.2 External Factors

A negotiated benchmark enables the department and provider to adjust the performance standard for the provider’s particular client mix and labour market they operate in. For example, if a provider operates in an area with weak labour market conditions a lower benchmark could be negotiated to take this factor into account. An issue with this approach is how accurate the adjustment process is. Individually negotiated targets would by definition be subjective and not directly comparable between providers. Without using a regression-based approach, it would be difficult to take into account the multitude of complex factors that would inform the setting of a negotiated target. A related issue is how to test the validity of claims by providers that the targets suggested by the department are too high or low. This would be particularly important once the initial negotiated target has been set, as changes to local economic conditions and in the composition of the provider’s caseload may be used by providers to argue that the initial target is no longer appropriate.

Case Study on Negotiated Benchmarking: US Workforce Investment Act

The U.S. Department of Labor in its quarterly reports of GPRA goals and results states that “since program performance goals are nationwide goals, they may not apply in every instance to individual states or grantees where different economic or other circumstances may demand an adjustment in performance expectations.” The negotiation process is designed to take these differences into account when determining performance levels. Analysing WIA program performance across the States, Heinrich (2004) assesses the relationship between labour market factors and participant characteristics to measured performance. This is to examine whether the negotiation process in adjusting performance standards accurately accounts for these factors. For instance, States with higher unemployment rates may negotiate to lower their benchmark as a higher unemployment level indicates fewer employment opportunities. The author examines this by estimating a multivariate regression using the performance level and differential between the State’s actual performance level and the negotiated benchmark as the dependent variables. Ordinary least squares regressions estimated separately for each performance standard show significant relationships between the negotiated performance levels and economic and participant characteristics. Factors affecting performance levels include education level, work history, unemployment rates and proportion of Hispanics or participants with limited English proficiency. The evidence suggests that adjustments were being made during negotiations to take account of economic and participant characteristics.

Subsequent analysis examined how accurately negotiations accounted for these factors. It is expected that states which accurately adjust for these factors during the negotiation process will see fewer or weaker relationships between the performance differentials and labour market factors and participant characteristics. Race, work history, education and unemployment rates continued to be statistically significant predictors of the performance differential, suggesting that adjustments for differences in characteristics were inadequate.

The results of Heinrich (2004) are further supported by qualitative evidence. Interviewing state officials, Barnow and King (2009) report that most officials indicated that the WIA’s performance management system was a step backwards from the adjusted benchmark model of the JTPA. They criticised the negotiation process as not having an appropriate procedure to adjust for participant characteristics and local economic conditions, resulting in states and local areas not being placed on

---

6 Workforce System Results, January 1-March 31, 2010, Employment and Training Administration, United States Department of Labor, p. 8.
a level playing field. A 2002 study by the U.S Government Accountability Office (GAO) also report
that many states felt the negotiation process did not sufficiently account for participant
characteristics and local economic conditions.

4.2.3 Forecasting

In addition to accounting for factors known at the time that the benchmark is established, it is also
important to adjust the benchmark to offset future risks (i.e. random economic shocks) to
performance that are outside the control of providers. While exceptional performance of a provider
is still an unbiased estimator of excess effort even in the presence of a random shock, this shock
introduces noise in the measurement of performance and the ability of the provider to meet the
required benchmark (Courty, Heinrich and Marschke, 2011). There are two challenges that arise
from adjusting for future economic conditions in a negotiated model. The first is that future
economic conditions are difficult to forecast, with actual economic performance differing
significantly from forecasts. This issue is exacerbated the longer the forecast required. For example,
forecasting future unemployment is subject to less accuracy three years in the future relative to
two months. The second challenge is how well the negotiated model takes into account economic
forecasts and how quickly performance standards change in response to changes in the economic
environment.

Case Study on Negotiated Benchmarking: US Workforce Investment Act

The negotiation process needs to make adjustments to the benchmarks to offset future unknown
risks outside the control of public managers. State performance benchmarks under WIA were
established using baseline performance data from previous years and also built in continuous
improvements in performance. For the first three years of the WIA, the negotiated benchmarks
increased an annual 1-2.5 per cent. Furthermore, the increase in performance between PY 2001 and
PY 2002 was greater than the increase between PY 2000 and PY 2001. For example, the mean
employment rate standard across all states was 66.44 per cent in 2000, 69.17 per cent in 2001 and
70.94 per cent in 2002. After the implementation of the WIA, there was a significant change in
economic conditions. Unemployment rates declined over the 1998-1999 period at a median decline
of 0.2 per cent. This improvement in employment condition was widespread, with 75 per cent of
states reporting a decline. This pattern continued over the 1999-2000 period. The trend in the
unemployment rate reversed over 2000 and 2001 with a median increase of 0.7 per cent with 75 per
cent of states experiencing an increase. The increase in the unemployment rate was even greater
between 2001 (4.59 per cent) and 2002 (5.35 per cent) with all States except one experiencing an
increase. Thus while the performance standards were increasing over this period, economic
conditions were worsening and creating an adverse labour market environment.

The difficulty in accurately estimating future economic conditions and incorporating these
expectations into the negotiated benchmark can be seen in the number of states failing to achieve
the negotiated performance standards. Making appropriate adjustments are important, as states are
required to achieve 80 per cent of the negotiated performance benchmark across all performance
measures in order to be eligible for an incentive grant. In addition, states may be subject to a 5 per
cent reduction in their WIA grant for not meeting their performance goals two years in a row.
Heinrich (2004) calculates the difference between a state’s performance level attained and the
performance target for that particular program year for each performance measure. A negative
differential indicates that states were on average performing below their benchmark. The results
show a negative turn in performance differentials from PY 2001 to the first quarter of PY 2002, in
line with the downturn in the general economy. Further evidence of the impact of economic
conditions on performance is shown in the 21 per cent decrease in the proportion of states meeting
their older youth entered unemployment rates.

The creaming and parking of participants can be exacerbated by a negotiated benchmark model. When external factors have not been properly accounted for in the performance standards, meeting targets is harder with harder to serve clients and in areas with poor labour markets. Adjusting the benchmark for differences in labour market conditions and participant characteristics offsets the incentives to cream and park participants. For example, a provider may increase their employment outcomes by focusing on those with a higher completed level of education relative to those with less (where those with a lower level having lower employment outcomes), leading to the parking of these participants. However, a negotiated model could increase the performance standard for providers who have a high proportion of male clients, offsetting the advantage of focussing on males. Providers should therefore be indifferent between the participants. This however is dependent on a negotiated benchmark accurately adjusting for differences in client characteristics and labour market conditions, which the above evidence indicates is difficult to achieve.

Case Study: The U.S. Workforce Investment Act

The evidence indicates that the negotiated benchmark model has contributed to the creaming and parking of participants. Since the introduction of the WIA, the share of adults receiving training who are low-income or disadvantaged has declined. Frank and Minoff (2005) report that the share of low-income adults receiving training declined 14 per cent over 2000 to 2003. An internal (U.S. DOL, 2002) study corroborates these findings. A survey of states by the U.S. DOL’s Chicago Regional Office indicated that registrations occurred at half the rate of enrolment of its predecessor program the Job Training Partnership Act. The survey found that some local areas were basing their decision on whether to register a person based on their likelihood of success as opposed to their level of need. The study suggests that this was driven by concerns over meeting performance levels. The 2002 General Accountability Office study also found that the need to meet performance levels contributed to the creaming and parking of program participants:

‘Many states reported that the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services at the local level. All the states we visited told us that local areas are not registering many WIA participants, largely attributing the low number of WIA participants to concerns by local staff about meeting performance levels. Local staff are reluctant to provide WIA-funded services to job seekers who may be less likely to get and keep a job. One state official described how local areas were carefully screening potential participants and holding meetings to decide whether to register them. As a result, individuals who are eligible for and may benefit from WIA-funded services may not be receiving services that are tracked under WIA.’

The use of creaming to meet performance standards arises from the inability of the negotiation process to sufficiently account for variations in economic conditions or populations served (GAO, 2002).

4.2.4 Resourcing Impost

A particular challenge in negotiating performance targets is the significant resource requirements this entails. Under a negotiated model, the Department would need to negotiate each performance measure across every site in Australia. There are currently 11 performance measures in JSA and 9 in DES. In JSA, the performance measures also apply to four different streams. According to the March 2013 Start Ratings release there were 555 contracts and 2,240 sites awarded a Star Rating in JSA. For DES there were 1000 contracts and 2,494 sites according to the December 2012 Star Ratings release. Undertaking negotiations for every single contract or site would be an enormous workload for each
of the providers and the department, which would involve setting tens of thousands of individual performance measure benchmarks. For example, the Department may provide five benchmarks for each performance indicator showing the performance a provider needs to achieve to obtain a particular Star Rating. With 31 indicators for JSA, this would equate to 86,025 individual benchmarks that would need to be negotiated by the Department at the contract level and 347,200 benchmarks at the site level. With 10 indicators for DES, this equates to 50,000 benchmarks at the contract level and 124,700 benchmarks at the site level. Furthermore, a negotiated model would require the ability to renegotiate benchmarks during the performance period to take into account unexpected changes in economic conditions. As discussed by Tulip and Wallace (2012), forecasts of Australia’s key macroeconomic variables are subjected to significant uncertainty. Performance standards will need to be renegotiated when actual economic conditions differ to that forecasted.

Case Study on Negotiated Benchmarking: US Workforce Investment Act

In Austria’s Public Employment Services (PES) a negotiated model is used for setting performance targets. The Austrian PES had eight headline targets as of 2011 (Ecorys, 2011).

- Transitions into employment of older persons within six months of unemployment;
- Transitions of youth within six months of unemployment into long-term-unemployment;
- Transitions into employment of people with less than two months employment in the last year;
- Number of women re-entering the labour market after childbirth in vocational training or in employment after training;
- Transitions into employment after training;
- Number of girls/women in higher technical vocational training;
- Filled vacancies;
- Registered high-skill vacancies.

Performance standards for these measures are set by the respective superior organisational levels but arrived at through negotiation. A number of steps are involved in setting the performance standard:

1. Targets are set following an extensive discussion between the federal government, social partners and the federal and national organisations belonging to the national public employment service.
2. Following this coordinating process, the administrative board of the Austrian PES (Arbeitsmarktservice, AMS) comprised of a three-tier governance structure including national headquarters (located in Vienna), regional offices (one in each of the nine federal states or Länder), and local PES offices sets the target values at the state and federal level.
3. State organisations set the target values for the subordinate regional organisations of the PES. These targets are negotiated on an annual basis.
4. Once the annual target is set, a sub-annual verification of performance and revision of targets is conducted. This examines the extent to which the target values will be achieved by the end of the year. This is completed by setting sub-annual values taking into account anticipate seasonal fluctuations and comparing these estimates to actual performance.
5. Targets are revised if the annual targets are unlikely to be achieved based upon the sub-annual results.

This negotiation process for setting annual target values of labour market indicators and reassessing them during the year is a costly and time consuming process. It consumes resources that are
subsequently not available for the AMS to achieve these targets (Kaltenborn, 2009).

4.3 Adjusted Benchmarking Model

An adjusted benchmarking system is where performance targets are adjusted for a set of unanticipated or uncontrollable factors that affect the achievement of the performance target using formal statistical methods. The first step involves determining the base level of performance providers are subject to. There are many options to determine the base level of performance, as discussed in section 4.1.1. The second step in adjusting performance involves identifying factors that are outside the organisation’s control (Rubenstein, Schwartz, and Stiefel, 1999). The third step is to estimate a regression equation in which the performance measure of interest is regressed against this set of variables and the pooling data on a set of similar organisations. This method measures the performance of each individual organisation while accounting for differences in client characteristics and local economic conditions. To show this, let us imagine a hypothetical adjustment model used to assess provider performance, with the performance measure in question being the ‘13 Week Full Outcome Rate’. The technique employed involves adjusting the 13 Week Full Outcome Rate target \((P_s)\) to which an individual or organisation’s measured performance \((P_m)\) is compared. The minimum 13 Week Full Outcome target for providers is established by using a regression model to adjust for client demographic characteristics \((X)\) and economic conditions \((Z)\) that may influence performance (e.g. local area unemployment rates). Typically, baseline data and/or data on past performance \((P_0)\) and the vectors \(X\) and \(Z\) of factors influencing performance are pooled across units and used to estimate a model, such as:

\[
P_0 = \alpha + \beta_1 X + \beta_2 Z + \epsilon
\]

where the estimates of \(\beta_1\) and \(\beta_2\) (vectors) are subsequently used as weights for the influence of these factors in adjusting the common standard \((P_s)\) to derive unit-specific performance targets for a given performance measure. Performance is then judged not by comparing actual performance \((P_s)\) across units and/or time, but by comparing the differential between a unit’s target \((P_{st})\) and its measured performance \((P_{mi})\). By assigning high expected performance to providers with favourable external factors and lower expected performance to providers with less favourable external factors, the statistical adjustment reduces the bias of the raw outcome measures and will provide relatively accurate information about provider performance.

Case Study on Adjusted Benchmarking: The Workforce Investment Act

The U.S Department of Labor (DOL) adopted a regression-adjusted methodology to set state and local area performance targets for the Workforce Investment Act (WIA) in 2011. These targets were used in the negotiation process for setting the PY 2011 performance standards at both the state and Workforce Investment Board (WIB) local area levels for the Adult, Dislocated Worker and Youth programs. This extended the regression-adjusted approach adopted by the US DOL for setting national targets for these three programs. The approach estimates the performance targets based upon what was achieved in the last observable program year and then adjusts this target based upon how changes in participant characteristics and unemployment rates are expected to affect future outcomes. Eberts, Huang and Cai (2012) provide a detailed overview of the regression methodology employed to estimate these targets, which is discussed below.

The reference point in making adjustments is the past performance of each entity. This differs to the approach used in the prior Job Training Partnership Act (JTPA) that used the national average performance level as the reference point. The next step adjusts for changes in personal characteristics. Estimates of the effect of participant characteristics on outcomes are obtained by
regressing each performance measure on all personal characteristics recorded in the Workforce Investment Act Standard Reporting Database (WISARD) (as is done in the Star Ratings). These factors include gender, age, ethnic background, education level, disability, income level, work history, homelessness and English proficiency, veteran and ex-offender and single parent status. The estimated coefficients are used as weights to calculate the adjustment factor resulting from the change in personal characteristics over time. The third step in developing performance targets is to adjust for forecasted changes in unemployment rates. Other macroeconomic factors such as GDP growth are not considered. In estimating the effect of unemployment rates on performance outcomes for each state and WIB, the residuals generated from the previous step are used in place of the actual outcomes recorded. In other words, the difference between the actual outcomes for each performance measure and the estimated outcomes from the first regression is used for estimating the effect of unemployment rates. This approach enables an ‘identical’ person to be placed in each local labour market to observe the effect of that area’s unemployment rate on that ‘identical’ person’s performance outcomes. The residuals from the regression are first aggregated to the WIB level. The estimated effect of unemployment rates are obtained by regressing the aggregated residuals on the WIB unemployment rates, with state estimates obtained by pooling the WIBs in each state. To estimate the effect of future changes of unemployment rates on performance, OMB forecasts of the change in the national unemployment rate are used. The forecasted change in national unemployment rates multiplied by the estimated effect of unemployment rates on each performance measure provides the statistical adjustment for different economic conditions across states and WIBs.

After the adjustments are made, the final step involves calculating the performance targets. The target is calculated by adding the adjustment from personal characteristics and the adjustment from changes in unemployment. The sum of these two adjustments are added to the past performance of each entity to obtain the target.

### 4.3.1 External Factors

Adjustment models allow for the performance standards applied to each individual provider to be adjusted to account for external factors using statistical regression. This is similar to the approach of the Star Ratings model, with the difference being that the adjustment is done ex-ante (i.e. at the start) to the performance period, not ex-post. Studies examining performance adjustments indicate that they are an improvement over unadjusted measures leading to improved equity in the system. Using data on elementary schools in Georgia U.S. over 1996-1997, Rubenstein, Stiefel and Schwartz (2003) find that ‘raw’ rankings of school performance change considerably after adjusting for external factors affecting educational performance such as poverty. Examining school outcomes, Fiedler, Wight and Schimdt (1999) analyse expenditures of public hospitals over a 12-year period in El Salvador. National (tertiary) hospitals received larger funding relative to lower-level (district) hospitals under the belief that national hospitals treat persons with more difficult illnesses or who have more severe illnesses relative to district hospitals. The authors test this assumption by estimating the required funding the two hospital types should receive controlling for patient morbidity, outputs and other characteristics. Results show that district hospitals were significantly underfunded once differences in client mix were adjusted for, highlighting significant funding inequity resulting from a lack of statistical adjustment to the funding methodology.

An analysis of the US 1996 Personal Responsibility and Work Opportunities Reconciliation Act also shows the importance of adjusting performance measures to ensure the performance framework aligns with policy goals (Barnow and Heinrich, 2010). Incentivising States to encourage parental responsibility, $20 million was awarded to a total of five States for the largest reductions of out-of-
wedlock births to total births. In awarding the bonus, the proportion of out-of-wedlock births to total births for the prior two years was compared to that of the preceding two-year period. Offner (2001) stated that state winners of the bonus in 1999 (District of Columbia, California, Michigan, Alabama and Massachusetts) all had large African American and Hispanic populations. The author reported that the non-marital birth-rates for these two groups fell twice as fast as white non-marital birth rates, suggesting that demographic characteristics and not state policies were responsible for changes in measured performance. This conclusion was supported by the U.S. Department of Health and Human Services (2000) who said ‘more evidence is needed to fully understand the range of factors contributing to a decrease in the proportion of out-of-wedlock births in these particular States. The failure to adjust performance measures for external factors affecting the measure undermined the purpose of the performance measurement system that sought to promote accountability, improve performance incentives and increase policy effectiveness. The downside of adjustment procedures though is that adjustments made prior to the performance period can lead to inaccurate adjustments if the characteristics of the provider’s caseload over the performance period differ to what occurred in the past.

Case Study: The U.S. Job Training Partnership Act

The U.S. Job Training Partnership Act (JTPA) highlights the problems that arise when a provider’s current client mix is extrapolated into the future. The JTPA’s performance standard adjustment procedure involved employing a regression model to adjust the performance indicators for differences in local conditions and participant characteristics. Adjustment weights from the regression model were applied to the following performance measures: follow-up adult entered-employment, follow-up weekly earnings and entered-employment rate. In a regression model, some factors may affect the value of the above performance measures and others may not. For example, whether a person is female or male may be significantly related to follow-up adult entered-employment, whereas being a single parent may not be significantly related.

Under the JTPA’s performance framework, the regression model used to make adjustment for the following program year is based on data from the prior program year. Only factors that were significantly related to the performance measures and were of the expected sign and a reasonable magnitude were included (Barnow, 1992). For example, if the coefficient for the proportion of single parents is not statistically significant or positive then that variable is removed for that year. The issue that arises is that although the variable may not be significant for the prior year, this may change for the year coming. By excluding this variable, the adjustment procedure has not accurately accounted for the effect of participant characteristics on performance. This leads to significant measurement errors. Examining the effect of JTPA on people with disabilities, Barnow (1996) found that not including all relevant factors for the following year lead to inconsistent treatment of people with disabilities. This JTPA case study highlights a broad problem with an adjusted benchmark. A provider’s client mix is not static but changes over time. As an adjusted benchmark is estimated at the beginning of the performance period, changes in client mix will affect the applicability of the estimated benchmark. The resulting adjustments to provider performance will only be accurate at the beginning of the performance period and will become less accurate over time.

4.3.2 Forecasting

One of the advantages of an adjusted model is that is can take into account forecasts of future economic conditions and changes in the client mix of participants when adjusting the benchmarks of each individual provider. However, because the benchmark is adjusted using the forecasted point estimates of different economic variables, it is important to understand how this is derived and the resulting issues in using them. The forecasting technology employed involves the use of an econometric model, with the use of judgement supplied by in-depth analysis of major sectors of the
economy (Treasury, 2012). Schuh (2001) provides a brief overview of an econometric model used to forecast a particular economic variable. Let \( y_t \) be a variable, such as output growth, to be forecast. The forecasting model is

\[
y_t = f(A(L)y_{t-1}, B(L)X_{t;\delta}) + \epsilon_t
\]

where \( A \) and \( B \) are unknown parameters, \( L \) is the lag operator; \( X_t \) is a set of explanatory variables, \( \delta \) is a set of parameters, \( f(\cdot) \) is a function, \( t \) denotes time, and \( \epsilon_t \) is a stochastic error. Forecasts are obtained as follows. Using data on \( y_t \) and \( X_t \), forecasters calculate econometric estimates of \( \delta \). Using a variety of methods, including auxiliary forecasting models, judgmental extrapolation, exogenous policy assumptions and forecasts from other forecasters, the forecaster obtains forecasts of the explanatory variables, \( \overline{X}_{t+k} \), \( k \) periods into the future. Then the \( k \)-step ahead forecast (made at the beginning of period \( t \)) is

\[
\overline{y}_{t+k} = f(A(L)y_{t+k-1}, B(L)\overline{X}_{t+k};\overline{\delta}) + \epsilon_t
\]

A forecast should be unbiased, efficient and have uncorrelated errors. Unbiased means that the forecasts are on average equal to the actual data. Efficient means that forecasts come from accurate models of economic behaviour using all relevant information available. Uncorrelated errors refer to forecast errors not being correlated with past errors. Achieving accurate unbiased forecasts are very difficult to achieve. As the Treasury (Australia, 2012, p. 1) states,

‘Forecasting is an inherently difficult exercise. It is challenging to capture the salient features of a modern complex economy in a framework simple enough to be tractable. Unlike the physical sciences, the inability to conduct repeated experiments means that it is difficult to use economic data to disentangle quantitatively the impacts of different influences on the economy, particularly in the face of continual structural change. Human behaviour, in particular the influence of animal spirits, has elements that are inherently unpredictable. Economies are buffeted by shocks, which by their very nature are not foreseeable, for example economic disasters, such as droughts, and technological advances. Lags in data collection and survey error mean that forecasters are reliant upon their own informed assessment about the current state of the economy.’

As a result, violations of these basic forecasting principles can occur due to inaccurate parameter estimates, omitted explanatory variables and erroneous models of the economy. For employment services, the ability to forecast extreme economic movements and structural shifts is of high importance given the significant impact these have on employment and outcomes achieved. The evidence indicates that forecasters do a poor job predicting both outcomes. The identification and prediction of major economic turning points in the business cycle has been a failure of most forecasters globally (Gruen and Stephan, 2010). For example, Treasury budget forecasts exhibited insignificant forecast errors over the period 1990-91 to 2011-2012. Forecast errors though were correlated with the economic cycle, underestimating economic growth during upswings and overestimating growth during the downswings. This is because models are highly dependent on what has happened in the past, which isn’t necessarily predictive of the future. For example, Stevens (2004) notes that average professional forecast accuracy for CPI inflation and GDP growth improved over 1984-2005. However, a naïve forecast predicting that the future value is equal to the current value also improved over the period, indicating that the improvement was due to reduced volatility in the economy as opposed to improved forecasting ability (Simon, 2001).

Forecasting issues are not limited to Australian organisations. Surveying US and UK forecasting evaluations, Fildes and Stekler (2002) conclude that forecasters generally did not predict recessions
and made systematic errors. Growth rates were underestimated in expansionary periods and overestimated during recessions. Inflation was also underestimated when it was accelerating and overestimated when inflation was decelerating. Oller and Barot (2000) find that forecasts made by European national institutes and the OECD did not generally predict cyclical downturns, with some forecasting groups never forecasting negative real growth despite significant declines in growth in the countries concerned. Examining private sector growth forecasts for a large sample of countries, Loungani (2001) finds that forecasters did not predict recessions in advance. These results suggest that the use of forecasted point estimates would do a poor job in accurately adjusting benchmarks for future changes in economic conditions. A further issue with using these forecasts is that benchmarks apply at the site and contract level. Conversely, the point estimates of different macroeconomic variables apply at the national level. This means that forecasts of economic conditions at the national level may not apply to the particular local area that the provider is situated in.

**Case Study: U.S. Workforce Investment Act**

An analysis of the Workforce Investment Act shows that forecast inaccuracy relating to future economic conditions reduces the suitability of the adjusted benchmark that providers are subject to, Bartik, Eberts and Huang (2009) find a significant and negative relationship between the WIA’s performance indicators for employment and unemployment rates. The results indicate the need for differences in economic conditions to be taken into account when determining performance standards. However, because performance standards are set for the future, economic conditions need to be forecasted before an adjusted performance standard can be estimated. Eberts, Huang and Cai (2013) calculate adjusted performance standards for States and Workforce Investment Board (WIB) local are levels for the Adult, Dislocated Worker and Youth programs. In determining the effect of unemployment rates on performance indicators, they use unemployment rate forecasts of the OMB.

Over the periods 2007 to 2009, the authors compared the adjusted forecasts across States to the actual performance each state achieved for the three years for Adult Entered Employment Rates. For each year, there was a difference in performance of 4.0 per cent (2007), 4.7 per cent (2008) and 6.0 per cent (2009). Performance will differ to that expected across states as a result of States strategies in regards to their management of the WIA program. The strategies of some States will generate significant performance improvements, leading them to exceed their performance standard and vice versa. It is also expected that performance will differ when future economic conditions diverge from what was forecasted. This makes it difficult to determine what effect inaccurate economic forecasts have on the performance standard. Its effect can be seen by examining the standard deviation of the performance figures for each year. The standard deviation figures reveal significant dispersion in performance difference across States, being 4.6 per cent (2007), 3.7 per cent (2008) and 6.8 per cent (2009). Importantly, the deviation of performance increased significantly in 2009. U.S. economic conditions worsened significantly over that year, with the unemployment rate rising significantly from 5.8 per cent to 9.3 per cent. This unemployment rate was significantly higher than forecasted, with the U.S Federal reserve estimating the unemployment rate of 2009 to be 7.6 per cent as of October 2008.

---

7 [http://data.bls.gov/timeseries/LNS14000000](http://data.bls.gov/timeseries/LNS14000000)
5. Discussion

The Star Ratings model offers a number of advantages in assessing provider performance. This includes effective and comprehensive adjusting for external factors, using ex-post performance data in assessing providers, and reducing the incidence of creaming. The benchmarking models were assessed against these advantages within the public policy paradigm and the evidence indicated that does not offer the same benefits. However, there is a view among some stakeholders that the Star Ratings model also has disadvantages that benchmarking models can overcome. This section examines the disadvantages of the Star Ratings model raised by stakeholders and analyses whether benchmarking models offer a solution to these concerns.

The following lists a number of challenges stakeholders believe the Star Ratings model poses and examines if benchmarks present a solution. These issues include:

- Providers do not have certainty over expected performance levels;
- Lack of transparency over the calculations used to assess provider performance;
- Regression models are subject to measurement error;
- Adjustments for economic conditions does not provide for equity in outcomes;
- A high rating is not indicative of high quality of service;
- Does not promote continuous improvement in performance;
- For one provider to ‘succeed’ another must ‘fail’, creating competition leading to opportunistic behaviour and risk selection; and
- Reduced scope for collaboration and cooperation among providers.

5.1 Certainty of performance

Under a benchmark model, the department would relieve providers of the risks involved in forecasting future expected performance standards, by providing certainty to providers about the level of performance they are required to achieve. This is in contrast to the Star Ratings model because as it is a relative model, it cannot provide an expected performance that the provider needs to achieve. It is not possible to provide this information in a simple form as the stars are based on the ratio between predicted (i.e. expected) and actual performance. If a provider performed 20 per cent better than predicted then they would have an actual to predicted ratio of 1.2. So, to take a simple example, if a provider was predicted to have an outcome rate of 20 per cent and they had an actual outcome rate of 24 per cent their ratio would be 1.2 (24 per cent/20 per cent) for that specific indicator. One issue is that the predicted value is not known in advance because participant and local labour market characteristics are not known prior to the period. In a survey of specialist JSA providers, Flentje, Cull and Giuliani (2010) found that the key challenge providers highlighted was the disconnect between the performance data that providers are able to access through the ESS and the Star Ratings that they receive once the department has applied the regression model. This disconnect between performance data and the stars makes it difficult for providers to work towards improving their Star Rating.

They quote a provider stating that:

‘The fundamental challenge is that we have various measures for assessing performance, including outcome rates and percentile scores etc. as given in the ESS performance reports but the Star Ratings which we eventually receive following the application of the ‘regression model’ will often bare little resemblance to these scores! We regularly have a situation

9 The overall ratio is a weighted average of each of the actual versus predicted ratios from the regressions of the individual outcome components.
whereby performance appears to have improved or declined using the pre-regression model data, only to find that the Star Rating moves in the opposite direction!

This centres on concerns that, as the system currently operates, providers may maintain or even improve their absolute level of measured performance and have a lower Star Rating. Therefore relative performance measures reward providers that improve their absolute performance by more than others. Further, a benchmark would also not provide complete certainty. There are 31 performance measures for JSA and 10 for DES. Providers would be given benchmarks for each performance indicator indicating the level of performance they need to achieve to receive a 5-Star Rating. However, the interaction of all of these measures will go into determining the final Star Rating. Performance of x on performance measure y does not guarantee a 5-Star because it is dependent on the other performance measures.

In addition, an effective performance management system is one where sanctions are placed on poor performing providers. Though benchmarks may provide performance certainty, it reduces the ability of the department to accurately determine which providers are performing unsatisfactorily. This is because it does not necessarily accurately take into account differences in participant characteristics and labour market conditions. This reduces the effect of using business reallocations to improve performance.

Besley, Bevan and Burchardi (2009) show that in the context of hospitals, the ability to sanction poorly performing providers is imperative for performance management to have an impact on improving performance. In 2001, England introduced a ranking system of performance measurement dominated by targets for waiting times for National Health Service (NHS) trusts. Hospitals that failed against key targets and were zero-rated were named and shamed as failing, with their chief executives at risk of losing their job. The authors evaluate the effect of naming and shaming on performance through the natural experiment offered by the regime applied to waiting lists for hospital admissions in Wales. Though targets were introduced into the NHS in Wales, this did not include naming and shaming hospitals. Given that the Welsh and English NHS are otherwise organisationally similar and subjected to the same funding, the Welsh NHS is a suitable control group for evaluating the ‘treatment’ of the English NHS. Results show that policy of naming and shaming hospitals did reduce waiting times in England as compared to Wales. Propper, Sutton, Whitnall and Windmeijer (2008) find similar results for between England and Scotland.

5.2 Transparency

The Star Ratings has the advantage that it adjusts for differences in external factors that affect the provider performance. These benefits though may be contingent on providers understanding and having confidence in the adjustment mechanisms. This advantage therefore is also argued to be a negative aspect of the performance framework. The technically rigorous approach of statistical adjustments to performance measures may be said to be too complicated, leaving performance analysis out of reach of providers and the public (Shulock, 1999). This limits the transparency of results and hence how providers and the program itself are performing (Struyven and Steurs, 2004). Past consultation with providers has revealed some concern about the current performance monitoring system for this very reason.

The issue relating to benchmarked performance is that providers wish to have guidance in advance about the absolute performance level required to be awarded a high Star Rating. For example, Flentje, Cull and Giuliani (2010) conducted a survey of specialist JSA providers and reported that half of the respondents argued for the need for greater transparency and reduced complexity of how the Star Ratings is calculated. This is not explicitly feasible under the current system, though it is said that such a feature would improve the transparency of the model. If transparency is a driver of
performance improvement, it may be in all stakeholders’ interests to explore other potential models that do not rely on a relative measure.

Although Star Ratings are used to ensure fairness in provider ratings, this may be less important to providers than transparency. To illustrate, the US Child Support Enforcement (CSE) program established under Title IV-D of the Social Security Act links incentives to States performance in five areas: 1) paternity establishment, 2) child support orders, 3) collections on current support due, 4) cases with collections on arrears (past support due) and 5) cost-effectiveness. These performance rankings across states are not adjusted for differences in economic and demographic characteristics. Tapogna et al. (2002) found that using a regression adjustment model for the CSE performance rankings would increase the fairness of the rankings, finding that the adjustment models would lead in some instances to major changes in the rankings of the States. States though opposed the use of performance adjustments, even though they agreed that the CSE program operates in different socioeconomic environments. This was because of the added complexity and uncertainty this would add to the performance rankings and the difficulty in reaching a consensus on adjustment factors (Gardiner et al. 2004). However, not adjusting the performance standards for the preference of providers needs to be weighed against the needs of job seekers and the goals of the Department. The department seeks to promote continuous improvement in the performance of Australia’s employment services through allocating business to the best performance providers. It becomes more difficult to determine which are the best performing providers using unadjusted outcome rates.

Further, a lack of transparency isn’t confined to the Star Ratings model. Benchmarking models also have transparency issues because a benchmarking model still requires a methodology for determining the performance standard. With an adjusted benchmarking model the different benchmarks for each provider are developed with regression analysis, creating many similar issues about transparency that some providers dislike about the Star Ratings model. The approach of Germany’s PES demonstrates this. In examining the performance of the local public employment offices, cluster analysis is conducted at the local organisational units and districts of their performance management system called Management by Objectives. This was done for the purpose of grouping into clusters local organisational units which are similar in terms of exogenous factors to setting target values differentiated by region. To do this, regression analysis is conducted to determine relevant exogenous factors using a target indicator (in the unemployment insurance system: degree of integrations; in the welfare system; integration rate) as the dependent variable. Clustering was then carried out on the basis of the exogenous factors. This methodologically demanding process is difficult to understand by practitioners on the ground (Kaltenborn, 2009).

The absolute benchmark model employed in the UK Work Programme suggests that an absolute benchmark can suffer from a lack transparency also. In setting the absolute benchmark, the DWP determines the non-intervention level of performance that would occur in the absence of the WP. This refers to what the employment outcome rate would be if there was no employment program existed. The DWP then adds a target performance onto this level for the providers to achieve. Providers subject to the DWPs performance targets are unaware how the non-intervention level is set. The DWP states that it is based on an historical analysis of the labour market and forecasts of future employment but has not stated how the analysis was conducted to arrive at their non-intervention figures.

It could be argued that the negotiated model provides greater levels of transparency. Factors outside the provider’s control that affected program outcomes might be more easily conveyed and weighed in negotiations. This was the goal of the negotiated model used in the WIA by the U.S. DOL. Similarly, because the provider is actively involved in the negotiations, the provider will have a deep
understanding of the factors that were employed to determine the performance standard the provider is subject to. Improved transparency under this model however only extends to the department and each individual provider. It can lead to significant issues of transparency for the performance of providers as a whole. The negotiation process would lead to questions as to how performance standards were set across providers, having been based on a qualitative discussion rather than actual data. Further, performance targets become a reflection of the relative strengths and effectiveness of negotiation skills as opposed to appropriate accounting of exogenous factors.

5.3 Adjustment for external factors

Another concern with providing certainty of performance is that if the performance standard doesn’t accurately adjust for external factors, providers will challenge the performance benchmarks given to them. It may be unrealistic to believe that variations in provider quality can be isolated by adjusting for participant characteristics and labour market conditions. Regression-based performance adjustment models have been criticized for having low explanatory power and flawed specifications, suggesting that sometimes adjustments may be biased or unreliable (Barnow and Heinrich, 2010). In this situation, adjustments to provider performance provide a false sense that residual differences among providers reflect differences in quality. Such arguments have been levelled at risk adjustment approaches in the health industry, where risk adjustment is used to determine variation in hospital performance (Lilford et al., 2004). For instance, Lilford et al. (2004, p.1147) states that ‘case-mix [i.e. risk] adjustment can lead to the erroneous conclusion that an unbiased comparison between providers follows. We term this the case-mix fallacy.’

An example of this problem is illustrated when the U.S. Medicare agency released the hospital mortality rates to its beneficiaries in 1986. The mortality list showed the percentage of Medicare patients who died relative to the proportion expected based upon national statistics. For each hospital, the expected mortality rate of hospitals was adjusted for 89 variables that could affect performance. Results found that 142 hospitals had significantly higher death rates than predicted and 127 had significantly lower death rates. Of concern, for the facility with the most aberrant death rates, 87.6 per cent of Medicare patients died relative to a predicted 22.5 per cent. This facility however cared for terminally ill patients. The government’s risk adjusted model did not adequately adjust for patient risks. These results lead to significant backlash from the medical industry (Brinkley, 1986). That adjustments are not perfect however does not imply that benchmarking is a superior alternative. The argument that there is noise in the regression model does not mean it isn’t useful. The estimates of specific explanatory variables can still be highly significant, implying that there are factors having an important effect on predicted performance and should be taken into account. As the Productivity Commission has argued:

... the goal of the Star Rating model is not perfection (which is impossible), but rather to capture the most salient differences between local labour markets and client mixes. There will be prediction errors, but the counterfactual of using raw performance outcomes as the basis for assessment would introduce far greater problems.

5.4 Quality

The quality of Australia’s employment services is defined as the degree to which employment services for individuals increase the likelihood of desired employment outcomes. It is imperative that job seekers receive high quality services from providers and that they are rewarded for doing so. A

---

valid measure of provider quality of service should be able to distinguish accurately between providers that are delivering a service of acceptable quality and those delivering a poor-quality service. Variations in employment outcomes have been taken to indicate variations in the quality of the service provided. This is because providers who are implementing better practices to help job seekers will generate better outcomes. Outcome rates are adjusted for characteristics of the provider’s caseload and labour market conditions to separate the contribution of provider and non-provider effects. This is used to calculate a residual unexplained variation that may implicate the quality of service provided. It could be the case though that this residual variation does not reflect differences in underlying quality of service but random noise. If this is true, then differences in performance between providers do not reflect differences in quality but differences in unexplained factors.

The relationship between outcomes and quality of service can be examined by testing the correlation between a provider’s outcome rates and their adherence to certain standards of service that are evidenced to improve outcomes (Ben-Tovim et al., 2009). For example, if increasing the time spent with a job seeker at the initial interview improves outcomes, then differences in this process measure across providers should explain differences in provider performance as measured by Star Ratings. If process-based measures have no correlation to the rating of providers, then it might be assumed that Star Ratings do not convey differences in quality. Though conceptually an applicable test of quality, it is fraught with difficulties. First, there are currently no evidenced-based practices known to improve outcomes that apply across the range of participants that use employment services. Second, process-measures are not immune from bias as more disadvantaged participants are more challenging than less disadvantaged participants, tilting the playing field against those who help harder to serve job seekers. Finally, quality measures may be associated with improved outcomes for two reasons. First, measured activities may directly improve care. Second, success on these measures may be a marker for other unmeasured aspects of high quality care. If the latter is true, observed differences in outcomes will be larger than expected differences as measured by process-indicators (Werner, Bradlow and Asch, 2008). This would suggest that the Star Ratings model does not accurately measure differences in quality when it is in reality the failure of process-measures to accurately do so. Finally, as benchmarks are also applied to outcomes, challenges in measuring quality will still be applicable under a benchmarking model.

5.5 Equity

The Star Ratings model seeks to provide equity in terms of provider performance by taking into account factors affecting performance external to the provider. An alternative conception is the need to focus on equity in outcomes. It may be argued that if it is important to place highly disadvantaged job seekers into employment, then the performance standard should be the same across all organisations, regardless of the challenges this poses to some organisation to achieve minimum levels of performance. This is because a job seeker with high barriers to employment should be allowed to have the same access to opportunity regardless of where they live.

Employing an absolute benchmarking model can achieve this aim through setting the same standard for all employment providers. This was the approach taken in the U.S. No Child Left Behind (NCLB) Act, which required States to set standards for reading and mathematics proficiency and ensure that all children meet these minimum requirements within a specified timeframe. The system promotes standards-based education; a system where high standards are set and measurable goals are defined to achieve better individual education outcomes. The No Child Left Behind Act of 2001 (NCLB) mandates that 100 per cent of American students will be proficient in English Language Arts and mathematics by the year 2014 (Rosenberg, 2004). Rather than implementing a nation-wide performance framework, NCLB is administered by individual states which develop annual...
measurable objectives (AMOs) to determine whether or not a school, district or entire State is making adequate yearly progress (AYP) toward the 2014 projection (Cronin et al, 2007). This performance requirement was irrespective of the child’s backgrounds, special needs, their school resources etc. This forced State’s to find innovative strategies to meet these performance standards. Texas and California were funding incentive awards for school districts, schools and principals where student progress was demonstrated. Incentive awards were also given to individual teachers based upon their contribution to improved outcomes measured using value-added statistical models (House Research Organisation, Texas House of Representatives, 2004).

Unfortunately, the requirement that all students achieve the minimum performance level lead to schools attempting to ‘game’ what was viewed as an unfair system (Barnow and Heinrich, 2010). Radin (2006) uncovered intentional underreporting of high school dropout numbers. Jacob (2005) found significant increases in reading and math test scores in Chicago Public schools that were not consistent with prior trends but were consistent with the NCLB incentives. Evidence indicated that resources were reallocated across subjects and increases in student special education, which was particularly pertinent in low-income schools. Jacob and Levitt (2003) empirically document cheating by teachers who systematically altered student performance to increase test scores in order to meet minimum performance standards. This example highlights the problems that may arise in Australia’s employment services if an absolute benchmark is implemented. Providers with hard to serve clients or in poor labour market conditions may seek to game the system by parking more difficult participants (depending on their mix of participants) to improve performance in order to meet the benchmark.

As another illustration of seeking to achieve equity in outcomes, Iezzoni (2010) discusses the problems that can arise in the medical industry from pursuing equity goals. The author uses an example of health-care administrators decreeing that all older women should undergo a mammography, regardless of socio-demographic background. Providers who care for women that are less apt to obtain a mammogram should be held to the same standards regardless. Adjusting for participant characteristics therefore becomes irrelevant. Although this has the goal of achieving equity across subgroups, the consequence is that providers may neglect other issues as resources are diverted to meet this performance requirement.

5.6 Continuous Improvement

Continuous improvement refers to the performance of providers increasing over time. Outcomes data can provide a powerful insight in to how providers are performing and where to target quality improvement efforts, leading to improved provider performance. Rosenthal et al. (1998) documents the benefits of measuring performance data regarding quality improvements. The authors examined the regional initiative choice Cleveland Health Quality Choice of 1989, a severity adjustment system that includes diagnosis-specific models for medical, surgical, and obstetrical patients which are based on clinical data abstracted from patients' medical records.

Since 1992, semi-annual reports were disseminated profiling hospital mortality rates, length of stay and caesarean section rates using the severity adjustment system. Hospitals receive tabular and graphical representations of hospital outcomes and patient level data files to further examine outcomes in clinical subgroups. Results from case studies show that the dissemination of outcome data led to the development of successful hospital programs to decrease lengths of stay, caesarean section rates and hospital mortality rates. The Star Ratings model seeks to contribute to the effect of performance measurement on actual performance by incentivising providers to maximise performance to achieve a 4 or 5-Star. This is because failure to maximise their performance may lead to their business being reallocated to another provider/s or reduced opportunity to be awarded.
additional business. This is illustrated in Figures 2 and 3 showing shows that the number of 13 week outcomes is positively correlated with the release of Star Ratings. Examination of Figures 2 and 3 reveals definite peaks in outcomes claimed in the months of December and June, the months immediately prior to Star Ratings releases (JSA Star Ratings are now publicly released every quarter, as of August 2011).

Figure 2:

![13-week Full Outcomes by Month - JSA All Streams](image)

Figure 3:

![13-week Full Outcomes by Month - DES DMS and ESS](image)

Conversely, it has been argued that because the star-rating system benchmarks to average performance as opposed to fixed benchmarks of expected performance, this allows all services to drift towards low outcomes because they can still maintain their Star Rating. The need to simply achieve an outcome rate above other providers creates a disincentive for providers to strive for continuous improvement in performance if they know they are already performing better than the average. In contrast, a benchmark can be set at a level of expected performance that would force all providers to strive for improvement in their performance. Currently, provider performance in JSA and DES is converging to the average, as indicated by a growing proportion of providers being rated 3-Stars. However, this is what we would expect from a system that promotes continuous
improvement in performance. In the Productivity Commission’s 2002 review of Job Network, this results was anticipated as discussed in footnote 24.

Stars are ordinal not cardinal measures of performance. Currently, the underlying cardinal difference in performance between one and five star performers is large (chapter 10). However, as best practice diffuses throughout the industry and poorer performers are exited, then the difference may become much smaller.

The perverse incentives viewed by some to arise from the Star Ratings model can also be argued against a benchmarking model. If the targets are set either too high or low it can introduce unintended consequences into the market. A benchmark can be set at a level of expected performance that would force all providers to strive for improvement in their performance. However, once some providers had met their absolute targets, the incentive effects of the performance management system may be weakened. That is, they would no longer invest as much time and effort in lifting their performance (as it would no longer boost rewards). Alternatively, if the targets are set too low then a lot of providers may be rewarded with a 4 or 5-Star but where the performance of those higher rating providers may not necessarily reflect superior performance and service delivery. To the extent that there are improvements over time, the government’s ability to ensure job seekers receive the assistance that is most likely to get them quickly into sustainable quality jobs by driving continuous improvements would then be more limited. While the absolute benchmarks could be adjusted upwards there would be transitional issues as the benchmarks catch up to the outcome levels.

An alternative to the argument that Star Ratings may not drive continuous improvement is that even if they do, it is not as effective as the use of benchmarking. A relative model introduces competition between providers, forcing them to maximise their performance. Benchmarks offer the advantage of setting high but achievable performance targets. The empirical evidence indicates that relative performance models generate better performance relative to benchmarking models. Matsumura and Shin (2006) examine the effect of introducing relative-performance evaluation on the financial performance of 214 post offices managed by Korea Post. Up until late 1997, employees pay was tied to the performance of the store where the employees worked. Korea Post then introduced a Relative Performance Evaluation (RPE) based incentive plan which rewarded stores for their performance relative to an appropriate reference group. Using cluster analysis, stores were clustered into nine reference groups based on similar business environments. Results indicate that introducing competition between stores as a result of the RPE incentive plan is positively associated with store-level financial performance. This suggests that an RPE-based incentive plan can induce agents to exert more effort, leading to higher financial performance than with an incentive contract with metrics based on absolute performance.

Benchmarking may lead to greater performance over the long run because benchmarking reduces the level of competition in the market, allowing for greater levels of collaboration and innovation and hence improved performance. Whether the evidence indicates this to be the case is examined in detail in the next two subsections.

5.7 Competition and Performance

It is argued that competition should lead to improved provider performance as pressure from competitors incentivises providers to reduce costs, improve quality, achieve job outcomes, meet the needs of individual clients and develop and implement innovative practices (Productivity Commission, 2002). The possible effect of provider competition on performance can be examined through an analysis of the introduction of competition in other public sector services. Belfield and Levin (2002) review U.S. evidence from cross-sectional research on educational outcomes when
schools must compete with each other. The authors identify over 41 empirical studies testing the effects of competition. Competition is measured using either the Herfindahl Index or the enrolment rate at an alternative choice (e.g. private school). Outcomes are separated into those relating to academic test scores, graduation/attainment, expenditures/efficiency, teacher quality, wages, and house prices. A sizable majority of these studies report beneficial effects of competition across all outcomes, with many reporting statistically significant correlations. The evidence shows reasonably consistent evidence of a positive link between competition (choice) and education quality.

In regards to the medical industry, Cooper et al. (2011) examines whether or not hospital competition in a market with fixed reimbursement prices can prompt improvements in clinical quality. In January 2006, the British Government introduced a major extension of their market-based reforms to the English National Health Service (NHS). From January 2006 onwards, every patient in England could choose their hospital for secondary care and hospitals had to compete with each other to attract patients to secure their revenue. One of the central aims of this policy was to create financial incentives for providers to improve their clinical performance. They test the effectiveness of this policy using the natural experiment provided by the choice-based reforms, which created sharper financial incentives for hospitals in markets where choice is geographically feasible and that prior to 2006, in the absence of patient choice, hospitals had no direct financial incentive to improve performance in order to attract more patients. Using AMI mortality as a quality indicator, results show that mortality fell more quickly (i.e. quality improved) for patients living in more competitive markets after the introduction of hospital competition in January 2006. A further study by Cooper et al. (2010) showed that the introduction of competition also made hospitals more efficient. The authors examined a patient’s length of stay in hospital for an elective hip replacement using two measures: the time from a patient’s admission until their surgery that is a direct function of hospital efficiency; and the time from surgery to discharge. Results show that hospitals facing greater competition reduced their pre-surgery length of stay relative to monopoly providers but not their post-surgery length of stay.

5.8 Innovation

An issue with competition in employment services is that any reduction in collaboration between providers may hinder the level of innovation in the sector and hence the long-term performance of the sector as a whole. Innovation within the employment services market encompasses two aspects. The first concerns the production of something new or significantly improved, including product and process innovations. Product innovations involve a new type of service offered to job seekers, such as mental health, counselling or housing services and so on through established partnerships with external providers. Process innovations involve production and delivery techniques such as meeting with job seekers online. The second concerns the adoption of innovations by providers. For example, a provider implementing an established practice will be considered innovative as it is an improvement on their old business practices. An employment services system should promote both types of innovation, encouraging the developments of new products and processes and the sharing of these ideas across providers.

The incentive to innovate is the difference in profit that a firm can earn if it invests in research and development (R&D) compared to what it would earn if it did not invest (Gilbert, 2006). Market competition and the profit motive are suggested to encourage providers to innovate under the credible threat that otherwise their business will be reallocated to competing firms. In a principal-agent relationship, the agent may not supply the effort necessary to maximise performance, which may involve examining and creating innovative services and processes. Higher competition incentivises the agent to maximise performance (Leibenstein, 2006). Second, competition means firms can be exited from the market. Innovation reduces this risk by making the company more
efficient. Greater levels of competition encourage greater levels of innovation as competitive firms are subject to a higher probability of being exited (Aghion, 1999).

The Star Ratings model encourages competition though business reallocations and the entrance of new providers into the employment services market. Reinganum (1983) develop a model on the effect of new entrants on innovation where an incumbent firm and a challenger engage in a game of innovation. The author finds that the incumbent firm has a lower marginal incentive to invest in R&D than does the challenger. Scherer (1980) empirically supports this conclusion, finding that potential new entrants play a crucial role in stimulating technological progress in which they contribute a disproportionately high share of all revolutionary products and processes. Cockburn and Henderson (1995) show that competition promotes innovation as an innovative discovery by one firm increases the innovative efforts of competing firms. The authors examine R&D expenditures by US pharmaceutical firms at the therapeutical program level. Results show that when a firm makes a discovery on a drug, rival firms R&D levels tend to increase. This is because discovery is cumulative and doesn’t stop new products being invented by rival firms. In other words, discoveries by one company expand the technological opportunities of other firms, thereby stimulating new innovation. Examining a sample of US firms, Blundell, Griffith and Reenen (1999) find that less competitive industries (based upon lower import penetration and higher concentration levels) have fewer aggregate innovations. Within industries, increased product competition in the industry stimulates innovative activity though higher market share firms tend to commercialise more innovations.

Competition within a purchaser-provider model of employment services however may stifle innovation and limit the sharing of best practice (Bowman & Horn, 2010; Considine, 2003). Any meaningful collaborative effort involves the sharing of how a service is delivered. It is unlikely that organisations who are seeking to achieve 5-Stars are prepared to share their “trade secrets” with each other. A number of studies point towards the beneficial effects of collaboration on innovation, showing that R&D intensity and the level of technological sophistication is positively correlated with the intensity and number of alliances in those sectors (Freeman, 1991; Hagedoorn, 1995). There are two possible reasons for firms deciding to cooperate. The first is a strategic decisions by firms, where the choice to collaborate occurs when the benefits exceed the costs. Firms will collaborate when they cannot produce the required skills and resources internally and that the risks of collaboration are at an acceptable level (Williamson, 1991). The second is that the formation of new ideas occurs from the relationships between firms, universities, research institutions, suppliers and customers. This is because knowledge is broadly distributed across this network. The degree to which firms source new innovations are a function of their participation within these groups (Levinthal and March, 1994; Powell, Koput and Smith-Doerr, 1996).

The knowledge required to place job seekers into employment is highly distributed across Australia’s employment services sector and other industries. For example, knowledge about how to best assist a person with mental health issues will require input from the mental health services sector. This knowledge will help assist the provider in placing the job seeker into employment, providing them with a competitive advantage relative to other providers. Inter-organisational collaboration can occur with a variety of partners, including suppliers and customers (Shaw, 1994; Von Hippel, 1988), universities and research centres (Gerwin, Kumar and Pal, 1992; Santoro, 2000; Tidd, Bessant and Pavitt, 2002) and even potential or existing industry competitors (Dodgson, 1993; Hamel, 1991). However, it is not the number of collaborative relationships per se that matter, which can lead to duplication of information, but the diversity of relationships. Analysing a sample of Belgian firms, Faems, Van Looy and Debackere (2005) find that the more a firm uses different collaboration possibilities, the more likely they are to innovate.
The type of partners that firms partner with though matter. Analysing a sample of Spanish manufacturing firms, Nieto and Santamaria (2007) find that collaborative networks are important for generating product innovations but that this is dependent on the composition of that network. Collaboration with suppliers, clients and research organisations have a positive impact on innovation but collaboration with competitors has a negative impact. These findings are corroborated by Fitjar and Rodriguez-Pose (2013) who examine the sources of product and process innovation in Norway through a survey of 1604 firms. The author’s find collaboration using Jensen’s et al. (2007) ‘Doing, Using and Interacting’ mode of interaction has a positive effect on innovation with regard to suppliers and customers but has a negative effect with regards to interaction between competitors. With respect to employment services, it could be argued collaboration is likely most beneficial between providers and external organisations, not necessarily between competitors. As the Star Ratings model promotes competition between providers but not between providers and outside organisations, it may be that the current performance framework does not have a negative impact on the type of collaboration necessary for innovative practices to occur.

6. Conclusion

This paper examines the applicability of benchmarking models in assessing the performance of providers in Australia’s employment services. This paper reviews the international literature on the use of benchmarking systems in the performance management of areas of public policy. The review identified three types of benchmarking models; absolute, negotiated and adjusted benchmarks.

Absolute benchmarks would provide a consistent performance standard that applies to all providers across Australia. Negotiated benchmarks allow for providers to negotiate with department a performance standard they believe is fair. Adjusted benchmarks allow the performance standard to be statistically adjusted to take into account the particular client mix of each provider and the labour market conditions they operate in.

The benchmarking models were evaluated against the advantages and disadvantages that are said to be associated with the current Star Ratings model. The Star Ratings model is used by the Department as a result of the numerous advantages it offers in measuring provider performance. These benefits include statistically adjusting for external factors affecting provider performance, the use of ex-post data, reducing the incentive to cream and park participants and using the performance rankings to inform business reallocation decisions.

The review of all three benchmarking models and the examples of international public programs that use these performance measurement systems suggest that all three models struggle to meet any of the advantages of the Star Ratings model. This creates insurmountable problems in trying to accurately assess provider performance in Australia’s employment services. However, stakeholders advise that the suggested problems generated by the Star Ratings model more than offsets these advantages. These drawbacks include; providers not having certainty over expected performance levels; lack of transparency; measurement error in the regression model used; mismatches between ratings and quality of service provided; not promoting continuous improvement; supporting competition that generates opportunistic behaviour and risk selection; and hindering collaboration that may reduce innovation. The research though reveals that these criticisms are either not accurate representations of the Star Ratings model or that the identified benchmarking models cannot easily overcome the problems stakeholders suggest the Star Ratings model possesses.
References


AIHW: Ben-Tovim, D., R. Woodman, J.E. Harrison, S. Pointer, P. Hakendorf and G Henley, 2009, ‘Measuring and reporting mortality in hospital patients,’ Cat. no. HSE 69, Canberra: AIHW.


Bartik, T.J., R.W. Eberts, and K. Kline, 2009, ‘Estimating a performance standards adjustment model for workforce programs that provides timely feedback and uses data from only one state,’ Upjohn Institute for Employment Research, 09-144.


Frank, A. and E. Minoff, 2005, ‘Declining share of adults receiving training under WIA are low-income or disadvantaged,’ Centre for Law and Social Policy, Washington, DC.


Gerwin, D., V. Kumar and S. Pal, 1992, ‘Transfer of advanced manufacturing technology from Canadian universities to industry,’ Technology Transfer, 12, 57-67.


PBS Newshour 2010, PBS Newshour, United States, 14th April 2013.
http://www.pbs.org/newshour/indepth_coverage/education/no_child/map.php?state=nv


Shulock, N.B., 1999, ‘The paradox of policy analysis: If it is not used, why do we produce so much of it?’ Journal of 'Policy Analysis and Management, 18, 226-44.


Treasury (Australia), 2012, ‘Review of Treasury macroeconomic and revenue forecasting.’


Werner, R.M., E.T. Bradlow and D.A. Asch, 2008, ‘Does hospital performance on process measures directly measure high quality care or is it a marker of unmeasured care?’ Health Services Research.


Appendix A

The six main steps in calculating Star Ratings are as follows:

Step 1: Calculate Actual Rate
- For the outcome rate (e.g. 13 week full outcomes) performance measures, the model calculates the actual outcome rates which have been achieved.
- For the time to outcome (e.g. time to job placement) performance measures, the model calculates the average number of days taken to achieve the outcomes.

Step 2: Calculate Expected Rate
By statistical regression, which takes account of variable job seeker and labour market characteristics, the model calculates:
- expected outcome rates for the outcome rate measures, and
- expected average number of days for the time to outcome measures.

Step 3: Calculate Ratio of Actual Rate to Expected Rate
- For the outcome rate performance measures, the ratio of actual rate to expected rate is calculated.
- For the time to outcome performance measures, the ratio is reversed, to give higher ratios to sites which are achieving their outcomes more quickly.

Step 4: Transform the Ratios to Performance Measure Scores Between 0 and 4
- The model transforms the ratios to performance measure scores between 0 and 4. This ensures that the performance measure weightings are applied correctly.

Step 5: Calculate the Performance Score
- Each performance measure score for a site is multiplied by the weighting, to give a weighted score. All weighted scores are then aggregated to calculate the Stream performance score.

Step 6: Calculate the Star Percentage and Star Rating
- From the performance scores for all sites across Australia, the national average site performance score is calculated. The example below shows a national average site performance score of 1.90. This average score is subject to change from release to release and is variable across programs/streets.
- Each site’s Star Percentage is then calculated by comparing their performance scores with the national average site performance score. The decimal components of the final calculated percentages are always rounded down. For example, a Star Percentage of 19.82 is rounded down to ‘+ 19’, and a Star Percentage of -19.82 is rounded down to ‘- 19’.
- Based on the Star Rating bands, a Star Percentage of -19 receives a Star Rating of 3-Stars.