

A multi source measurement approach to the assessment of higher order competencies

Justin Connally (The University of Melbourne)
Ken Jorgensen (Department of Defence)
Shelley Gillis (The University of Melbourne)
Patrick Griffin (The University of Melbourne)

Paper presented at the New Zealand Association for Research in Education (NZARE)
Australian Association for Research in Education (AARE) Joint Conference

Auckland, December 2003

Abstract

This paper presents findings from a study investigating the application of a multi-source measurement approach to the assessment of higher order competencies in the public service industry. The aim of the project was to develop and validate strategies to synthesise multiple sources of evidence to inform judgements of workplace competence. The methodology adopted integrated developments in two fields of study, performance appraisals and psychometrics. Seventy-five candidates were assessed using a combination of self-assessment, observer reports, interview and portfolio. Rasch analysis techniques proved to be useful for synthesising evidence gathered from a range of sources, and can be used to inform the decision making process.

Introduction

Competency based assessment

Competency based assessment (CBA) has been in use in Australian industry for several years. It is considered central to the National Training Framework that is expected to improve Australia's economic competitiveness within the Asian Pacific region (Keating, 1995). In Australia's vocational education and training (VET) context, competency is the specification of knowledge and skill, and the application of that knowledge and skill, to the standard of performance expected in the workplace (ANTA, 2002). Competence can be thought of as the ability to use and integrate a variety of skills and knowledge to solve real workplace problems (McCurry, 1994). Officially, competency was defined by the National Training Board (NTB, 1992) as consisting of five components: performing tasks, managing a set of tasks, incorporating task skills into the overall job role, handling contingencies, and transferring skills and knowledge to new and different contexts and situations.

CBA must therefore focus on the complex combination of knowledge and skills that are required for successful performance in the workplace. This often requires the collection of evidence from multiple sources, using multiple assessment methods across a period of time. Despite this, methods for combining evidence from multiple sources to reach an on balance judgement of competence have been too difficult to implement and not cost efficient (Griffin & Gillis, 2000).

CBA is the purposeful process of gathering appropriate and sufficient evidence of competence, and the interpretation of that information against industry competency standards. As part of this process, results are recorded and communicated to stakeholders (Griffin, 1995). The CBA movement claims to adhere to criterion referencing, as CBA measures performance against a set of pre-specified criteria. These performance criteria are industry defined and endorsed competency standards (Hager, Athanasou, & Gonczi, 1994). A criterion referenced interpretation requires comparisons to be made with predetermined standards of behaviour. Glaser (1981) clarified the definition of criterion referencing to include that it should "*encourage the development of procedures whereby assessments of proficiency could be referred to stages along progressions of increasing competence*" (Glaser, 1981, p935). In a criterion referenced framework tasks or competencies can be arranged along a progression or continuum of development, and individuals of varying competence can be positioned along this continuum.

Assessing management and higher order competencies

CBA at higher levels refers to the assessment of covert, higher order competencies required for successful performance in professional and skilled work. While these higher order competencies are not confined to jobs at the higher levels of the Australian Qualifications Framework (AQF), their importance certainly increases at these levels (Hager & Gillis, 1995). The assessment of management competencies, such as decision-making, problem solving, leadership, conflict resolution, negotiation and strategic planning skills have traditionally involved the sole use of supervisor reports. These competencies are inherently difficult to assess as there is greater independence of action, less supervision, and the impact of decisions are often difficult to attribute to the person responsible due to the time lapse between the action and its consequences (Edmonds & Stuart, 1992). Thus, the importance of integrating evidence from multiple sources for

the assessment of these competencies has been extensively documented in CBA literature (Griffin & Gillis, 2000).

Higher order competencies, such as those at the advanced diploma level within the AQF are complex, with a strong focus on the contingency and transferability skill dimensions. These dimensions are difficult to assess using direct observation and other methods that are typically used at lower AQF levels. An integrated approach to competency assessment is needed that assesses underpinning knowledge and understanding, problem solving and technical skills, attitudes, values, ethics, and the need for reflective practice. Assessing attitudes and values is particularly important at higher AQF levels, as individuals are likely to be responsible for the well being of others and compliance with codes of conduct, ethics and legislation (Hager & Gillis, 1995).

Multi-source assessment

Within a CBA framework, multi-source assessment refers to the use of a number of evidence gathering strategies or methods. While direct observation of performance is used frequently for assessments at the lower levels of the AQF, this technique is not appropriate for assessing higher order competencies that are covert in nature. The 360-degree feedback approach, used widely in performance appraisals, offers an alternative.

The central concept in 360-degree feedback is that performance ratings are obtained on an individual from a range of observers such as supervisors, peers, subordinates and the individual themselves (Griffin & Gillis, 2000). Critical to the success of the process is that observers have a high degree of familiarity with the individual being rated, interact with them regularly and have exposure to a considerable amount of their workplace performance (Hurley, 1998).

The 360-degree approach is thought to offer a number of advantages over traditional assessment techniques. Based in part on the assumptions of measurement theory, information obtained from multiple sources is thought to produce more reliable and valid results (Hurley, 1998). It is suggested that 360-degree feedback is more objective as multiple observer provide a fairer and less biased view of performance (Fletcher, Baldry, & Cunningham-Snell, 1998). Another proposed benefit of 360-degree feedback is that different observers may provide unique information about the individual because they interact with them in different capacities (Goudy, 1998).

Often referred to as third party or observer reports within a CBA context, this approach appears well suited to the assessment of higher level management competencies, given their inherent complexity (Brutus, Fleenor, & London, 1998). Importantly, this strategy has the advantage of allowing for real time, on the job assessment of performance with minimal disruption to workplace activities (Griffin & Gillis, 2000), and hold benefits for assessment candidates, who would likely find feedback from a variety of sources as fairer and more accurate than any single evaluation (Bozeman, 1997). When used in conjunction with the traditional CBA methodologies of portfolio and interview, this approach allows for multiple sources of evidence to be gathered across a range of contexts and covering an extended period of time.

Research aims

The implementation of this approach for the assessment of higher order competencies is in line with the extensive CBA literature that argues the importance of holistic assessments and the integration of evidence from multiple sources (Griffin & Gillis, 2000). The use of such techniques for the assessment of management competencies is invaluable given that they are particularly difficult to assess (Gregarus & Robie, 1998). Unfortunately, as is the case with CBA, research into multi-source assessment has not progressed at the same rate as its implementation in the workplace, with limited studies conducted in organisational settings using appropriate samples (Gregarus & Robie, 1998; Hurley, 1998).

Further, a method is needed for synthesising evidence from multiple sources to formulate an overall judgement of the competence level of candidates. An aim of CBA is to determine the competence of candidates regardless of what evidence is used or which observers participate in the assessment process (Griffin & Gillis, 2000). This is the fundamental reliability concern in CBA, whether the placement of candidates in one category or another (e.g. competent or not yet competent) is consistent across assessment methods, times and contexts (Masters, 1993; Jaeger, 1989). This is because the purpose of assessment is to infer candidate competence beyond the sample of tasks used to estimate competence (Lunz & Wright, 1997).

This study investigated the application of a multi-source measurement approach to the assessment of higher order competencies within the public services industry. The primary aim of the investigation was to develop and validate strategies to synthesise multiple sources of evidence to inform judgements of workplace competence. Rasch calibration techniques (Rasch, 1961), such as the partial credit model (Masters, 1982) and the multi-faceted Rasch model (Linacre, 1994), were evaluated as tools for synthesising multiple sources of evidence into a single measure of competence for each candidate.

Method

Sample

Participants were 75 candidates from within the Department of Defence and other public service organisations. The sample consisted primarily of human resource and specialist managers (Australian Public Services Level 6), who were often the most senior person in their business unit or work area.

Unit of competency

The unit of competency used assessed in this study was entitled *Facilitate People Management* (PSPMNGT603A) from the Public Services Training Package (PSP99). The unit was related to the management working area and covered the implementation of people management strategies, plans and processes within the business unit in cooperation with specialist human resource personnel. This unit contained five elements and 23 Performance Criteria. The Elements contained within the unit of competency were undertake human resource planning, manage the performance of individuals, manage grievance procedures, counsel employees, and manage employee rehabilitation. The critical aspects of evidence for the unit included an integrated demonstration of effective people management strategies, which were expected to facilitate the attainment

of business unit objectives. The unit contributed to awards at the Advanced Diploma level.

Instrument development

Based largely on the 360-degree feedback model used widely in performance appraisals, the *Facilitate People Management Assessment Instrument (FPM-AI)* was developed to gather structured performance information from workplace observers such as supervisors, peers and subordinates (observer reports). A self-assessment version of the *FPM-AI* was also developed for use by candidates during self-assessment. Interview questions were written to assess the underpinning skills and knowledge required for competent performance against the unit of competency, and were designed to assess candidates' ability to evaluate and reflect on their performance, and identify areas for improvement.

As recommended, rubrics for the *FPM-AI* and interview questions were developed by a group of subject matter experts (SME) drawn from a cross-section of workplaces, thus representing a variety of perspectives (Bennett, 1998). Rubrics are "a set of scoring guidelines that describe the characteristics of the different levels of performance used in scoring or judging a performance" (Gronlund, 1998, p. 225). Thus the central feature of rubrics are the ordered categories or levels of performance that comprise a description of the cognitive, affective and psychomotor skills embedded in competent performance (Griffin, 2000; Waltman, 1997). Underpinning the concept of rubrics is the criterion referenced interpretation in which an individual's achievement or competence is described in terms of the tasks that they can perform (Glaser, 1981). The use of criterion referenced definitions for rating scales convey far greater information about the quality of performance, discriminates more accurately between individuals, and allows for candidates to be given more diagnostic feedback, feedback that they will likely perceive as more constructive and valid (Bondy, 1983).

Selection of assessment methods

Candidates were required to select, in conjunction with their assessor, the most appropriate assessment methods for their context. All candidates were required to complete a self-assessment using the *FPM-AI* prior to selecting their assessment methods, as self-assessment assisted in identifying where evidence of competence could be gathered. It was recommended that candidates selected at least two of the three available methodologies (observer reports, interview and portfolio) in addition to completing a self-assessment to ensure that adequate evidence could be gathered for a decision of competence. Only nine candidates completed portfolios, while the most popular combination of assessments methods was observer reports and interview ($N=59$).

Candidates who selected observer reports as one of their assessment methods were required to negotiate with their assessor who would act as their workplace observers. All 75 candidates were assessed using observer reports, with a total of 71 supervisors, 105 peers and 98 subordinates completing the *FPM-AI*. All candidates also completed a self-assessment using the *FPM-AI*. On average, each candidate was rated by 3.66 observers in addition to their self-assessment. Observers were selected if they were familiar with the skills and knowledge required to manage within the public service industry, had the opportunity to observe the candidate applying their skills and knowledge in the workplace, and understood the nature of the candidate's role (Thorndike, 1997).

Rasch Analysis

Data analysis was undertaken using the Rasch partial credit model (Masters, 1982) and the multi-faceted Rasch model (Linacre, 1983), both extensions of the simple logistic Rasch model (Rasch, 1960 and revised 1980). The simple logistic Rasch model states that the probability of a person answering a dichotomous item correctly is dependent only upon the ability (or competence) of the person (θ_n) and the difficulty of the item (δ_i). Extending on this, the partial credit model (Masters, 1982) allows for scoring one or more intermediate levels on an item, and to award partial credit for reaching one of these levels. The partial credit model (shown in equation 1) can be applied to situations where ordered response alternatives vary in number and structure across items (Linacre, 1994; Masters, 1982), and thresholds ($\tau_{i1}, \tau_{i2}, \dots, \tau_{im}$) are estimated for each response alternative (Wright & Masters, 1983).

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = \theta_n - \delta_i + \tau_{ik} \quad (1)$$

In equation 1, P_{nik} is the probability of candidate n receiving a rating of k on item i ; P_{nik-1} is the probability of candidate n receiving a rating of $k-1$ on item i ; θ_n is the competence of candidate n ; δ_i is the difficulty of item i ; and τ_{ik} is the difficulty of receiving a rating of k rather than $k-1$ for each item i separately.

The Rasch model can also be extended to include additional *facets* of the assessment context, such as observer group severity, by the addition of an additional (severity) parameter (γ_g). As can be seen in equation 2, it is not only the competence of the candidate (θ_n) and the difficulty of the item (δ_i) that governs the probability of a particular rating, but also the severity of the observer group (γ_g) making the judgement.

$$\log\left(\frac{P_{nigk}}{P_{nigk-1}}\right) = \theta_n - \delta_i + \gamma_g + \tau_{ik} \quad (2)$$

In equation 2, P_{nigk} is the probability of candidate n receiving a rating of k from observer group g on item i ; P_{nigk-1} is the probability of candidate n receiving a rating of $k-1$ from observer group g on item i ; θ_n is the competence of candidate n ; δ_i is the difficulty of item i ; γ_g is the severity of observer group g and τ_{ik} is the difficulty of receiving a rating of k rather than $k-1$ averaged across all observer groups g for each item i separately.

Results and Discussion

Multi-faceted rasch analysis

The FPM-AI ratings obtained from observer reports and self-assessments were calibrated using the multi-faceted Rasch model, allowing for multiple sets of ratings for each candidate to be combined into a single measure of competence. This analysis also allowed for an investigation of observer group (supervisor, peer, subordinate and self-

ratings) severity. Multi-faceted Rasch analysis was performed using the FACETS software program (Linacre, 1999).

The multi-faceted Rasch analysis produced a relatively normal distribution of candidate competence measures. The distribution of FPM-*AI* item difficulty estimates displayed a slight positive skew. When taken with the mean competence estimate for the sample (0.23 logits), these observations suggested that the FPM-*AI* was well matched to the competence level of the candidates assessed. The 30-item FPM-*AI* displayed very good internal consistency ($\alpha=.91$, $N=346$), and successfully separated candidates of varying levels of competence (separation reliability=.93) providing evidence of criterion validity (Wright & Masters, 1982). Similarly, the very high item separation index (0.98) provided evidence of construct validity (Wright & Masters, 1982).

All FPM-*AI* items displayed acceptable fit. As no infit values were below 0.7 it was concluded that there were no redundant items. Similarly, as no values were greater than 1.3 there was no evidence of psychometric multidimensionality, that is, a single measure of competence had been obtained for all candidates (Myford & Wolfe, 2000; McNamara, 1996). The fit statistics produced for each candidate proved a useful tool for evaluating individual assessments. While very few candidates displayed fit values outside of the acceptable range of 0.7 to 1.3 (Adams & Khoo, 1995), these statistics provided an indication of which assessments required review or clarification, and could assist assessors in determining appropriate courses of action for assessments.

With the exception of self-ratings, which displayed a degree of leniency (-0.29 logits), all observer groups tended to cluster around the same severity level (0.07 to 0.13 logits). All observer groups displayed infit values well within the acceptable range, with all values between 0.9 and 1.1. The overall difference between the observer groups was significant, $\chi^2(3)=233.0$, $p<.01$, with a very high separation reliability (0.98). While this indicated actual differences in observer group severity, it should be noted that the range of severity estimates was more than 5 times smaller than the range of candidate competence estimates, and almost 4 times smaller than the range of item difficulty estimates. Further, the only ratings to display some variation were self-ratings, and as all candidates completed a self-assessment this apparent leniency in self-ratings did not advantage any candidates. This is illustrated in Figure 1, which shows a comparison of competence estimates (in logits) and (average) raw scores for each candidate. As can be seen in Figure 1, the candidate competence estimates and average raw scores were very similar ($r=0.99$, $p<.001$, $N=75$), indicating that observer group severity had almost no impact on candidate competence estimates.

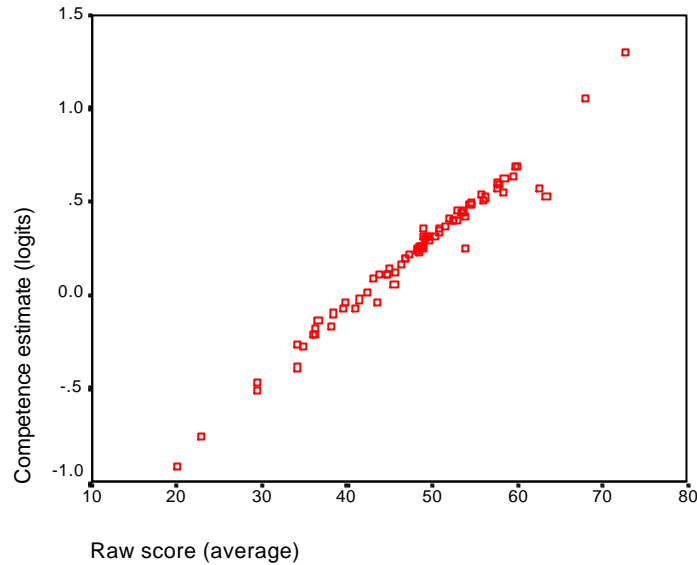


Figure 1: Plot of candidate competence estimates versus raw scores (average)

Further, the degree of leniency observed in self-ratings likely resulted from the level of non response present in the ratings provided by other observers. In an additional multi-faceted Rasch analysis, in which non response was treated as missing data, self-ratings did not display the same pattern of leniency. In fact, the difference between the severity estimates of self-ratings and the ratings provided by all other observers was very small, with a much smaller separation reliability (.91) and a chi square value ($\chi^2(3)=38.3$) approaching insignificance. This is an important finding, as any inferences regarding candidate competence should not be constrained to any specifics of the assessment situation, such as which observers provided ratings (Myford & Wolfe, 2000).

Rasch partial credit analysis

An aim of the present study was to simultaneously calibrate assessment methods on the same underlying measurement scale, resulting in a single, overall measure of competence for each candidate. As insufficient data was gathered using the portfolio method of assessment ($N=9$), this method could not be included in the calibration. Self-assessments and observer reports completed using the FPM-AI were calibrated simultaneously with interview ratings using the Rasch partial credit model (Masters, 1982). The Quest computer program (Adams & Khoo, 1995) was used to perform the Rasch partial credit analysis.

As the Rasch partial credit model requires a single rating for each item, the modal (most frequently occurring) rating for each of the FPM-AI items was used in the calibration procedure. Rasch analysis was used to explore the utility of using the modal rating in this instance. A standard Rasch partial credit model was used to generate candidate competence estimates based on the modal rating for each of the FPM-AI items, and these estimates were compared to competence estimates based on all ratings obtained using the multi-faceted Rasch model. The correlation between these two competence estimates was very high ($r=.87, p<.001, N=75$), indicating that the modal rating for each FPM-AI item was illustrative of the complete set of ratings.

Figure 2 displays graphically the calibration of the FPM-*AI* items and interview questions. The first column provides the linear measurement scale (*logits*), with positive and higher values indicate more of the latent construct being measured. The second column displays the distribution of candidate competence estimates, with each candidate represented by an *x*. The third and fourth columns display the distribution of thresholds for the FPM-*AI* items (third column) and interview questions (fourth column), with positive and higher values indicating more difficult thresholds. The *x.y* notation in Figure 2 indicates item number followed by threshold (rating). For example, 12.3 represents a rating of three on item 12.

As can be seen in Figure 2, the majority of candidates had competence estimates above zero on the measurement scale ($M=0.58$ logits). Competence estimates ranged from -1.14 logits to 1.64 logits, and the candidate separation reliability was .91. While candidate competence estimates were distributed over 2.78 logits, thresholds were spread over a range of 5.53 logits. The FPM-*AI* thresholds were distributed relatively normally across the measurement scale, while the distribution of interview thresholds displayed a slight positive skew.

Eighteen candidates displayed infit values outside of the range of 0.7 to 1.3 (Adams & Khoo, 1993). When evaluating candidate fit against the less restrictive range of 0.6 to 1.5 (as recommended by Englehard, 1994), only eight candidates displayed a degree of misfit. Again, candidate fit statistics provided a framework within which individual assessments could be reviewed. For example, the large infit values displayed by some candidates occurred because they received high performance ratings from their coworkers (observer reports) but were unable to demonstrate a comparable level of underpinning knowledge and understanding during the interview. These assessments would require review before a decision of competence could be supported.

Measure	Candidates (more competent)	FPM-AI items (more difficult)	Interview questions (more difficult)
3.0		30.4	
		19.2	
			33.5
		22.3	
		4.3	
2.0		27.4	
		20.3	
			43.3
		28.2	
		14.3	45.3
			34.5 40.4 46.3
	X	9.3	
	XXX	29.2	47.3
	X		37.4
	X	2.3 18.3	41.4
X	27.3		
1.0	XX		32.3
	XX	10.3 13.3 16.3 30.3	31.3
	XX		34.4 36.3
	XXXX	6.4 12.4 17.3 22.2	
	XXX	3.4 25.2 29.1 30.2	44.3
	XXXX	9.2 26.4 27.1 30.1	
	XXXXXXXX		39.3
	XXXXXX	5.2 12.3 20.2	33.4
	XX	28.1	38.4 40.3
	XX	3.3	
XX	3.2 11.3 12.2 15.3		
XX	1.3 13.2 14.2 17.2 19.1		
XXX	16.2 21.2	41.3	
X	24.2	33.3	
.0		6.3 7.3 10.2	
	XX	2.2 11.2 23.2 26.2	47.2
	XXX		34.3 37.3
	XX		35.3
		6.2 8 26.1	39.2 46.2
-1.0		18.2 20.1 21.1	32.2 33.2
		24.1	34.2
	X	4.2 11.1 23.1	40.2 41.2
	X	15.2 16.1	36.2 41.1
		2.1 7.2	31.2 42.2 47.1
			43.2 44.2 45.2
		1.2 9.1	33.1 39.1
	X		
	X	1.1	34.1
		22.1	38.1 38.2
-2.0			40.1
		18.1	32.1 35.2
		13.1	
		15.1	42.1 46.1
		10.1	31.1 35.1 36.1
			43.1
		12.1 14.1	
		25.1	
		17.1	44.1
			37.1
	5.1		
	4.1 6.1		
	(less competent)	(less difficult)	(less difficult)

Figure 2: Distribution of candidate competence and threshold difficulty

Variable interpretation

As Rasch analysis was undertaken using the partial credit model, the variable interpretation involved an examination of response thresholds. The variable interpretation process involved a content analysis of clusters of thresholds at approximately the same difficulty level along the developmental continuum to determine if a common, substantive interpretation was possible (Giffin, 2000). For successful variable interpretation, thresholds must have substantive meaning, order and magnitude (Griffin, 2003).

The variable interpretation process, undertaken by SMEs, yielded three clearly distinguishable levels of competence, each characterised by a common set of underpinning skills and knowledge. The cut points for the three levels of competence are indicated on Figure 2 by the position of the horizontal lines. The first cut point, separating the lower and middle band levels, was located at -0.31 logits. This cut point was determined by SMEs to be the cut point on the scale for decisions of competence. While no decisions of competence were made in this study, it was however an aim of the investigation to determine cut points for levels of competence on the constructed measurement scale. The second cut point, separating the middle and upper band levels was located at 0.56 logits. This point was determined to be the cut point on the measurement scale for competent/highly competent decisions.

A content analysis of the thresholds located within the three levels allowed for the development of *profile* descriptions for each level. This process is central to the criterion referenced interpretation at the heart of CBA as it allows for the monitoring of progress along the continuum of developing competence (Glaser, 1981). The profile descriptions for the three levels of competence identified are presented in Table 1. Significantly, even candidates who were determined to be not yet competent could be provided with a description of the skills and knowledge that they possess and a framework within which they could work towards gaining competence. A further aim of this investigation was to determine to what extent the empirically derived variable interpretation would correspond to the hypothesised Facilitate People Management construct. Both the hypothesised construct and derived variable were characterised by three distinct levels of competence, and the congruence between the two variables provided evidence of the construct validity of the measure (Griffin, 2003).

Table 1
Profile descriptions for Facilitate People Management

Profile description	Percentage of candidates
<p>Highly competent Demonstrates lateral and strategic thinking in all facets of human resource planning. Introduces effective feedback mechanisms and provides opportunities for staff development of transferable skills. Develops staff capability to evaluate, self-appraise and improve performance. Improves performance management criteria against organisational performance. Considers long term solutions to grievances, and reviews grievance procedures. Anticipates the need for counselling support. Implements, monitors and evaluates the return to work program considering resource implications. Fosters social integration within the business unit.</p>	31%
<p>Competent Considers current and future staffing needs. Justifies the preferred human resource plan which links to higher level corporate planning. Applies strategies to improve performance management processes, and negotiates staff agreement to performance management criteria. Applies systematic preventative approaches to potential complaints and grievances. Applies a range of counselling techniques, and where needed refers staff to appropriate agencies ensuring consistently successful counselling outcomes. Determines the impact, including resource implications, of the return to work program. Maintains a supportive workplace.</p>	62%
<p>Not yet competent Analyses current human resource needs using available planning data. Prioritises resources within budgetary guidance. Informs staff of performance management criteria, gives effective feedback to staff and applies strategies for performance improvement. Seeks resolution of actual and potential grievances. Identifies problems requiring counselling and offers appropriate counselling services, including referrals, to facilitate performance and well-being.</p>	7%

Competence decision making

As reported, no empirical judgments of competence were made in the present study. Assessors made all decisions of candidate competence prior to any data analysis. It was however possible to make a hypothetical decision of competence for each of the candidates who were assessed using observer reports and interview based on their Rasch estimated level of competence, and to compare these empirical decisions to the assessor judgements made at the conclusion of the assessment process. As assessors simply judged candidates as not yet competent or competent, only these decisions could be compared (the competent and highly competent band levels were combined).

Only four candidates had competence estimates located in the lowest band level and as such would have been empirically determined to be not yet competent. These four candidates were also determined to be not yet competent by their assessors. Nine candidates estimated to be competent based on the Rasch analysis were judged as not yet competent by their assessors. Even when considering the lower competence estimates (based on 95 percent confidence intervals) for each of these candidates, there were still seven candidates who were estimated to be above the competent cut point who were judged by their assessors to be not yet competent.

One possible explanation for this disparity is that assessors applied different decision making techniques. Despite attempts to encourage professional judgement in the decision making process, at least in some cases assessors appeared reluctant to make any inference of competence, relying instead on the application of a check list approach that required every performance criteria to be demonstrated across each method. Hagar et al (1994) suggested that when making decisions of competence, successful performance on difficult elements is likely evidence of competence on a wider range of other elements, and may pre-suppose the ability to perform competently on a range of easier elements. A degree of inference is critical to the success of the assessment process, and is particularly important when assessing higher order competencies as underlying competence cannot be directly observed but must rather be inferred from performances on a range of tasks (Masters, 1994).

It is also possible that the standards applied by different assessors varied. The official definition of competence is “*the application of skills and knowledge to the standard expected in the workplace*” (ANTA, 2002). While this *standard* should be consistent across workplaces, and even across industries, the reality is that different organisations have different conceptualisations of what constitutes competent performance. In other words, some organisations demand more of their employees. This presets a challenge for the VET sector that is not easily addressed but has significant implications for the portability of skills and knowledge.

Conclusions

Multi-faceted Rasch analysis proved to be a useful tool for synthesising ratings gathered from observer reports and self-assessments. This analysis not only produced a single measure of competence for each candidate, adjusted for any differences in observer group severity, but also a variety of statistics for evaluating the precision of these measures (e.g. standard error, fit statistics). Similarly, the simultaneous calibration of FPM-*AI* items and interview questions provided valuable information regarding the appropriateness of observer reports and interview methodologies for assessing units of

competency at the Advanced Diploma level. If applied within a CBA system, such approaches to evidence synthesis could reduce much of the workload currently placed on assessors, and would provide candidates with more interpretable and usable feedback.

The variable interpretation process yielded a Facilitate People Management variable with three clearly distinguishable levels of competence. This analysis allowed for hypothetical decisions of competence to be made to explore the utility of a Rasch based approach to decision making. Rasch analysis can be used successfully to inform the decision making process, and the development of profile descriptions for each level of competence has significant implications for reporting and the construction of staff training and development strategies.

References

- Adams, R., & Khoo, S. T. (1995). *Quest: an Interactive Item Analysis Program*. Melbourne: Australian Council for Educational Research.
- ANTA. (2002). *Australian quality training framework: Guidelines for course developers: A guide to developing VET courses for accreditation under the Australian quality training framework*. Melbourne.
- Bennett, J. L. (1998). *A procedure for equating curriculum-based public examinations using professional judgement informed by the psychometric analysis of response data and student scripts.*, University of New South Wales.
- Bondy, K. N. (1983). Criterion-referenced definitions for rating scales in clinical evaluation. *Journal of Nursing Education*, 22(9), 376-382.
- Bozeman, D. P. (1997). Interrater agreement in multi-source performance appraisal: a commentary. *Journal of Organizational Behavior*, 18, 313-316.
- Brutus, S., Fleenor, J. W., & London, M. (1998). Does 360 feedback work in different industries: a between-industry comparison of the reliability and validity of multi-source performance ratings. *Journal of Management*, 17(3), 177-190.
- Edmonds, T., & Stuart, D. (1992). Assessing competence at higher levels: the management experience, *Competence and Assessment*
- Englehard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement*, 31(2), 93-112.
- Fletcher, C., Baldry, C., & Cunningham-Snell, N. (1998). The psychometric properties of 360 degree feedback: and empirical study and a cautionary tale. *International Journal of Selection and Assessment*, 6(1), 19-34.
- Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36, 923-936.
- Goudy, K. L. (1998). *The measurement equivalence of 360-degree feedback ratings across rater populations*. Unpublished Doctor of Philosophy, DePaul, Chicago, Illinois.
- Gregarus, G. J., & Robie, C. (1998). A new look at within source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, 83(6), 960-968.
- Griffin, P. (1995). Competency assessment: avoiding the pitfalls of the past. *Australian and New Zealand Journal of Vocational Education Research*, 3(2), 33-59.
- Griffin, P. (2000, 28 April). *Competency Based Assessment of Higher Order Competencies*. Paper presented at the NSW ACEA State Conference, Mudgee.
- Griffin, P., & Gillis, S. (2000). *An integrated approach to competency assessment*. Paper presented at the British Education Research Association, Cardiff, Wales.
- Gronlund, N. E. (1998). *Assessment of student achievement* (sixth ed.). Needham Heights, MA: Allyn and Bacon.
- Hager, P., Athanasou, J., & Gonczi, A. (1994). *Assessment Technical Manual*. Canberra: AGPS.
- Hager, P., & Gillis, S. (1995). *Assessment at higher levels*. Paper presented at the NCVER, Adelaide.
- Hurley, S. (1998). Application of team-based 360 degree feedback systems. *Team Performance Management*, 4(5), 202-210.
- Jaeger, R. M. (1989). Certification of student competence. In R.L.Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). New York: Macmillan.
- Keating, J. (1995). *Australian Training Reform: Implications for Schools*. Melbourne: Curriculum Corporation.
- Linacre, J. M. (1994). *Many-facet Rasch Measurement*. Chicago: MESA Press.
- Linacre, J. M. (1999). *Facets*. Chicago: MESA Press.

- Lunz, M. E., & Wright, B. D. (1997). Latent trait models for performance examinations. In J. Roost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. Munster: Waxmann.
- Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Masters, G. (1993, April). *Certainty and Probability in Assessment of Competence*. Paper presented at the VEETAC National Assessment Research Forum on Competency Based Assessment Issues.
- McCurry, D. (1994). Assessing standards of competence. In A. Gonczi (Ed.), *Developing a Competent Workforce: Adult Learning Strategies for vocational Educators and Trainers*. Adelaide: NCVER.
- McNamara, T. F. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman.
- Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the test of spoken english assessment system* (65). Princeton, New Jersey: Educational testing service.
- NTB. (1992). *National Competency Standards Policies and Guidelines* (2nd ed.). Canberra.
- Rasch, G. (1960 and revised 1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Danmarks Paedagogiske Institute and Chicage: University of Chicago Press.
- Thorndike, R. M. (1997). *Measurement and Evaluation in Psychology and Education*: Prebtice-Hall.
- Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement*, 34(2), 101-121.
- Wright, B., & Masters, G. (1983). *Rating Scale Analysis*. Chicago: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA.