

**The appropriateness of professional judgement to determine
performance rubrics in a graded competency based assessment
framework**

Andrea Bateman, The University of Ballarat
Patrick Griffin, The University of Melbourne
Australia, 2003

Abstract

With the implementation of competency based assessment within the Australian vocational education and training (VET) sector the focus has been on valid and reliable assessments to ensure that there are consistent outcomes across training providers. Underpinning this has been the notion of providing assessment judgements within a dichotomous reporting framework; that is competent or not yet competent.

This study investigated the appropriateness of professional expertise in developing performance rubrics for competency as defined by the *Public Services Training Package*. Levels of performance were identified along a continuum for interpretive purposes and competency decision making. Groups of judges estimated the relative difficulty of each of the rubrics. Item response theory calibrated the rubrics. A comparison of judges' estimates of difficulty and interpretation of developmental continuum, was compared to the outcomes of item response analysis.

The findings indicated that the specialists who developed the items and their relative difficulty levels were accurate in their judgments. The internal consistency measure was high indicating that the assessment instrument was a reliable measure of the construct. The criterion validity measure (person separation index) was high. There was room for improvement in terms of the construct validity (item separation index) of the instrument.

The study concluded that standard setting using subject matter experts proved adequate for developing performance rubrics.

Introduction

This study represented an attempt to apply in a competency based assessment context the procedures for standard setting by subject matter experts. It was motivated by the increasing interest in determining levels of performance and also in an interest in utilising multi-rater or third party assessment to determine competence. This study was derived from a broader research project undertaken within the public safety and public service Training Packages that involved an investigation of how multiple sources of data could be synthesised into a single score and used for interpretation as a competency decision. This research project evolved out of a need to design, develop and validate strategies to synthesise multiple sources of evidence to make judgements of competence associated with higher order competencies that cannot be directly observed nor simulated.

The focus of this study was the investigation of the appropriateness of professional judgement to predict or determine performance rubrics within a graded competency based assessment and reporting framework. It also investigated the criterion and construct validity of the scoring rubrics developed by the subject matter experts.

Assessing levels of performance in a competency based context

When competency based training and assessment was introduced into the VET sector in Australia in the early 1990s, there were no clear policy guidelines about whether learners' levels of performance should be assessed and reported or whether competency based assessment should be conducted and reported using a dichotomous reporting system of competent/not yet competent (Williams & Bateman 2002). Debate has raged about whether the principles, underpinning competency based training and assessment, imply only one standard of performance, or whether multiple standards of performance were possible within a criterion-referenced assessment system.

Recent research in performance assessment has been stimulated by the renewed interest in recognising levels of performance in criterion referenced assessment (of which competency based assessment is a form) has occurred following such reports as McGaw (1996); NSW Government (1997); Griffin & Gillis (2001) and Griffin, Gillis, Keating & Fennessy (2001).

Standards referencing is considered a subset of criterion referencing where levels (or bands) of performance are defined along a continuum of increasing competence (Glaser, 1963) and used for interpretive purposes to infer a competency decision. The standards referenced framework allows for

reporting of results in a range of ways including the dichotomous competent/not yet competent, grades or differentiating scores (Griffin & Gillis, 2001) and is said to address the requirements of both the VET and school sector (Griffin et al 2001).

Methodology

The unit of competency under consideration in this study was *Facilitate People Management* (PSPMNGT603A) from the *Public Services Training Package* (ANTA, 1999). The unit of competency was within the management area and contained 5 elements, 23 performance criteria:

1. Undertake human resource planning (4 Performance Criteria)
2. Manage the performance of individuals (8 Performance Criteria)
3. Manage grievance procedures (3 Performance Criteria)
4. Counsel employees (5 Performance Criteria)
5. Manage employee rehabilitation (3 Performance Criteria).

The context of the research investigation was in the field of public service industry, with the focus of the study being on the assessment of the skills and knowledge required to facilitate people management in senior management positions.

Selection of candidates and raters

This study used data gathered in the main by the Department of Defence personnel who volunteered to be participants. Candidates were required to undertake a self-assessment and provide assessments of their performance undertaken by peers, subordinates, supervisors and clients. In some instances the subject matter experts were also raters. Third party raters (n=142) were treated equally for the calibration of the scales. In the development of the scale it was assumed that responses by the raters to the subset of items (behavioural descriptors) were dependent on the candidate's position on the latent variable.

Selection and training of subject matter experts

The current research study used a multi-stage approach to standard setting similar to that of Bennett (1998b). As with the Bennett's model (1998b), the selected subject experts within this study were responsible for the development of the rubrics. Subject matter experts who were identified and recruited from the Department of Defence were assessor trained and considered expert in the field under review. They undertook additional training and participated in a workshop to develop the initial pool of items. An iterative process (Jaeger 1989) was used in this study with subject matter experts seeking feedback and considering the opinions of other subject matter experts and reaching a consensus on item development, predicting the level of difficulty of each item's behavioural descriptors, determining the cut-off scores as well as the development of the levels of band descriptors of performance. Throughout the development and trialing stages the sharing of opinions between the subject matter experts (n=10) was critical to the development and refinement of the items and the criterion referenced band level descriptors.

Item development

The development of behaviourally anchored rating scales is similar to the development of criterion referenced rating scale descriptors. The development of descriptions of performance provided raters and candidates with clear indicators of levels of performance. To best reflect workplace realities the development of the initial item pool involved a detailed analysis of the unit of competency by the subject matter experts involving a review of all dimensions of competency, the range of variables and the key competencies, so that the rubrics would best reflect the requirements. The dimensions of competency include task skill, task management, contingency management, incorporating the task(s) into the job role environment and the ability to transfer the skills and knowledge to other contexts and situations. The range of variables is the range of contexts in which the competency may be performed or from which the

evidence can be provided. The key competencies are considered generic competencies embedded within the principal unit of competency

The development of the items was based initially on the Performance Criteria within the unit of competency, however; the subject matter experts decided in a number of instances to combine a number of Performance Criteria into one item. There were 18 items in the initial pool relating to 23 Performance Criteria. The subject matter experts then developed in consultation the behavioural descriptors for each item. Refer to Table 1 for a sample item and its descriptors.

Table 1: Sample item and behavioural descriptors

Item	Behavioural descriptors		
	1	2	3
Review of action (grievance) and complaints are managed promptly and in a manner which optimises the likelihood of a positive outcome.	Identifies and seek resolution of actual/potential grievances and complaints.	Anticipates potential grievances and complaints and applies, systematic, preventative approaches.	Diagnoses systematic issues and executes judgement on possible long term solutions.

Throughout the development process a number of assumptions were made. It was assumed that behavioural descriptors were not of the same level of difficulty within or across items. The measure of difficulty for a particular descriptor of an item is taken from Rasch measurement terminology. In essence it measures how much ability or competence is required to achieve a rating in a certain score category, the more ability or competence that a person has the more likely that they will receive a higher rating.

No assumptions made about the amount of difference between the behavioural descriptors within an item. That is for example, the level of difficulty between category 1 and category 2 is not necessarily equal distance from category 2 and category 3. Therefore, using the Rasch model, all items and their descriptors are positioned somewhere on the continuum or scale. Hence items and their descriptors can be compared to each other based on their position on the scale.

Finally it was not always assumed that there were a specific number of levels, of performance (descriptors), for each item. The items and their descriptors have been labelled with the following sequence, for example *item 2 descriptor 3* appears in the tables and graphs as 2.3.

Throughout its development the subject matter experts reviewed and then piloted the instrument with managers in the field.

Development of holistic level descriptors

The final stage of the rubrics development involved a re-examination of the behavioural descriptors to determine their predicted level of difficulty. In order to understand and to represent differences in quality of performance, the codes were placed on a continuum that demonstrated how the performance reflected development. To begin the lowest quality descriptor (1) for the simplest item was placed on the bottom of the grid. Each placement for each behavioural descriptor was a judgement made by the group of subject matter experts as to the predicted level of difficulty and the quality of candidate performance. Each item descriptor was placed on the grid relative to the others. The number assigned in each grid was the item number followed by the behavioural descriptor code. Empty cells in the grid were there simply to illustrate the relative difference between predicted candidate performances.

The item descriptors' codes (i.e. 1, 2, 3, 4) were entered on a scale according to their predicted relative difficulty. Refer to the following table (Table 2) that illustrates this stage. Analysis of these predictions (Table 2) is ascertained later regarding the calibration of the scales using item response modelling.

Table 2: The hypothesised rubric for assessing the unit of competency: *Facilitate People Management*

Undertake HR planning								Manage performance of individuals								Manage grievance procedures		Counsel Employees						Manage employee rehabilitation								
3		4				3					4							3										4				
	3				4			3	3			3	3		3	3		2					2	4								
		3	3						3	3	3			3						3		2		3	3	2					4	
	2							2				2				2					2						2					
2					3	2			2	2	2			2	2		1			2		1			2	1			3			
			2	2			2						2			1	2							2							2	
		2			2				1		1			1					2	1												
	1						1		1		1			1				1			1			1	1							
1			1	1		1		1					1							1				1					1	1		
		1			1											1																
1.1	1.2	2.1	2.2	3.1	3.2	3.3	4.1	4.2	5.1	5.2	6.1	6.2	7.1	8.1	9.1	10.1	11.1	12.1	12.2	13.1	13.2	14.1	14.2	15.1	16.1	17.1	17.2	17.3	18.1			

Using the codes and relative position on the grid a holistic description of what the items have in common was developed. This meant reviewing the codes and determining what each cluster or group of codes meant when considered together. A common theme for each predicted level of performance was determined and developed into a brief description. The predicted cut-off between competent and not yet competent (the two lower levels of performance) was also determined. Table 3 illustrates the predicted band level descriptors for the unit of competency *Facilitate People Management*.

Table 3: The hypothesised performance level descriptors for the unit of competency: *Facilitate People Management*

Performance level descriptors
<p>High (Expert) Using an independent and proactive approach, can anticipate future HR planning requirements which link with the higher organisational plans. Implements continuous improvement strategies in all facets of people management activities. Is able to negotiate agreements and plans with staff and other relevant organisational parties. Embeds communication and feedback processes into work area practices and culture. Empowers staff to contribute to a supportive workplace environment.</p>
<p>Medium (Experienced) Under own initiative can align, develop, implement and review HR planning processes in accordance with budget and business plans, as well as organisational and legislative requirements for their work area/business unit. Has an in-depth understanding of a range of performance management processes, issues and strategies (including counselling). Can apply these when consulting with staff and/or dealing with staff issues. Able to assess and review resources required to establish action plans.</p>
<p>Low (competent) Under limited guidance, is able to identify, implement and modify HR planning processes in accordance with organisational and legislative requirements within their work area/business unit. Can inform and communicate with staff about performance and grievance related issues. Able to assess HR issues within their work environments. Develops and implements plans of action (eg return to work, performance and grievance issues).</p>
<p>Below (Searching comp) Has limited demonstrated ability to implement people management strategies, plans and processes within the business unit/work area.</p>

For easy reference a title was assigned to each level of performance. These titles and the concept of a continuum of competence are similar to that developed by Bondy (1983) and Benner (1984).

Data analyses procedures

Item response theory was used to calibrate and evaluate item descriptors, to determine difficulty estimate of performance descriptors and to help specialists to determine cut-off points for competence. The Rasch model is a statistical one that estimates the probability of a person demonstrating competence, to an item of known difficulty (Curtis & Denton 2002). Calibrating the items 'is a process that defines the level of difficulty of the items and establishes their accuracy as measuring devices' (Griffin, 1997).

The partial credit model enables the identification of one or more intermediate levels of performance on an item and awards partial credit for reaching these intermediate levels (Wright & Masters 1982) and can be extended to situations involving polytomous (i.e. divided into many parts) responses. It takes as its basic observation the number of steps that a person has made beyond the lowest performance level (Wilson & Iventosch 1988). In the partial credit model it is not assumed that every item has the same step structure or difficulty between levels (i.e. between categories or levels of the behavioural descriptors coded as 1, 2, 3 or 4) and is therefore useful for assessing performance of items or assessment tasks made up of sub-tasks. In the partial credit model the interaction between a person and an item is independent between the items (Wilson & Iventosch 1988).

This model can be used to describe any ordered sequence of dichotomous steps. By using an estimate of 'step difficulty' within each item of the assessment, the partial credit model positions a person on the underlying variable (which is assumed to underpin the unit of competency) considering the number of steps that the person has made above the lowest performance. Behavioural descriptors were developed with increasing levels of difficulty for each item rather than a right/wrong answer and therefore provided information about the person's progression and level of skills and knowledge of the unit of competency rather than the dichotomous reporting model of 'competent' or 'not yet

competent'. The point at which the likelihood of a higher level response becomes greater than that of a lower level response is termed the threshold.

The Rasch model is able to detect deviations from expected patterns of responses for both items and respondents. When all items fit the Rasch model it is considered that there is a predominantly single trait underlying all items (Waugh 2002). The accuracy of the scales or whether the variable meets the intentions of the assessment instrument developers can be determined using two key measures standard error and item fit.

The first is the standard error of measurement for each of the item difficulty estimates. This was calculated by determining the difference between the true (or modelled) item difficulty and the estimated item difficulty, using responses of all raters to that particular item (Wright & Stone, 1979).

The second is a measure of the extent to which the data fits the Rasch model. An analysis of the fit of the items to the Rasch model provides a means of determining how accurately the variable can be used to predict performance ability. The infit and outfit statistics indicate the degree to which individual persons and items fit the partial credit model and hence whether the data supports the construction of the linear scale (Wolfe & Miller 1997) and are based on the difference between expected and observed scores.

Both infit and outfit expect a value of 1.0 when the model fits the data (Wright & Masters 1982). It is generally considered that a 30 percent variation is acceptable in most cases, so that values between 0.7 and 1.3 are considered satisfactory (Wilczynski 1995). The Rasch perspective involves retaining only those items which are found to fit the model. Strictly speaking items that do not fit the model are examined to determine the cause of the misfit and may still be retained if it is believed that the misfit is due to a few large residuals (DeAyala, Dodd & Koch 1992, p. 3).

Items that underfit the model indicate either excessive randomness in the ratings or more likely a specific systematic problem causing the observed responses to differ from the expected responses (Wilczynski 1995), such as unexpected incorrect responses by high ability candidates or inherent inconsistencies within the item itself (Smith 1994). An item that underfits the model is an indication that the item is considered not to define the same construct as the rest of the items in the instrument or to be ambiguously defined (Lai, Haglund & Kielhofer 1999) and consideration should be given to excluding these items or reviewing the structure of the item and step ability level descriptors.

Items that overfit the model indicate predictability or redundancy, providing little additional information to the scale that is not already provided by other items (Curtis & Denton 2002, Lai et al 1999, Wilczynski 1995). Such items are still said to define the same construct as the rest of the items however, they don't improve the measurement qualities of the instrument (Lai et al 1999) and are considered of less concern than underfitting items (Curtis & Denton 2002).

Classical test analysis uses Cronbach alpha co-efficient of reliability to measure the internal consistency of the group of items on the assessment instrument. This measurement can vary from 0.00 to 1.00, however the closer the co-efficient is to 1.0 the more internally consistent is the instrument Griffin (1997). This means that a high alpha coefficient reflects item sampling adequacy (Gable & Ludlow 1990).

Using the Rasch model concurrent validity can be established by analysing the internal consistency of each person's responses (Wright & Masters 1982). In the Rasch model this is referred to as the person separation index, which is the extent to which the persons can be separated on the continuum of the variable. For example, if the cases (persons) are clustered at either end of the scale therefore the difficulty levels of the items do not match the ability measures of the persons.

The item separation index is an indicator of construct validity. Wright & Masters (1982) defined and explained construct validity in terms of the partial credit model. Construct validity could be affirmed if it provided adequate item separation that enabled the definition of several distinct levels (and therefore levels of complexity) of the variable.

Therefore the aim of the subject matter experts was to develop an assessment instrument that contained items that related only to the specified construct and which were sufficiently dispersed to

enable identification of levels of difficulty in accordance with the intentions of the instrument developers (Lunz, Wright & Linacre 1990).

Findings and Scale calibration

The *Facilitate People Management* scale was constructed to estimate the ability of managers in regards to the construct assumed to underpin the unit of competency. Both classical and Rasch analyses of the scale data were undertaken using the program *Quest* (Adams & Khoo 1993). Estimates were obtained from 30 items designed to examine the level of a candidate's skills and knowledge in the workplace related to facilitating people management.

The item difficulty estimates (d1) are an indication of the degree of difficulty or demand of the items. The range of item difficulty estimates provide an indication of the range of the ability levels that the assessment instrument is able to measure. The item difficulty estimates varied from +3.2 to -2.41 a range of 5.61 logits. Given the item logit range it appears as though the scale permits the measurement of the Facilitate People Management construct over a broad range of ability levels.

The Mean Squared Item Infit was 1.01, with a standard deviation of 0.19. These values illustrate that in general the items fit the Rasch model and that there is evidence of a dominant underlying construct in the variable being measured. The Mean Squared Case Infit was 1.06 with a standard deviation of 0.42. This indicates that there was generally a homogenous group with consistent fit.

The internal consistency was 0.91 which indicates that the assessment instrument was reasonably reliable as a measurement instrument. The reliability estimate of the item separation index was 0.67. This measure of reliability determines how sufficiently well separated each of the items were in terms of increasing intensity level within the latent construct and is referred to as construct validity. This estimate is a little low and indicates that the scale may not effectively measure the construct as would be desired. The reliability estimate for the cases or person separation index on the other hand was 0.9. This estimate of criterion validity indicates that the instrument was an effective measure in ensuring that it provided an effective distribution of the ability of the cases on the scale.

All items had an acceptable fit to the model except for the step level descriptors relating to items 2.1 and 6.1. Both these items had an infit value above the generally acceptable limit of 1.3 and are said to underfit the model. Items that underfit the Rasch model may be influenced by a 'factor that is not reflected in other items' or step level descriptors (Curtis & Denton 2002, p. 46). Consideration should be given to reviewing these items or for them to be removed from the scale. An analysis of these items and their step level descriptors indicate that the language may have been inconsistent or lacked applicability to the workplace context.

Given that there were only two items that were not fitting the Rasch model and given that the internal consistency was high it could be assumed that the subject matter experts were able to effectively develop and describe the levels of difficulty of the items that relate to the construct that underpins the unit of competency. Although the subject matter experts were able to develop an assessment instrument that enabled effective distribution of cases there was still room for further development to ensure improved construct validity (the separation of the items on the scale).

The difficulty of the items was plotted in decreasing order of difficulty. The set of items were then examined to identify specific clusters or groupings to determine cut-points for varying attitudinal levels on the variable. This required consideration of two criteria: the first to determine which items were clustering together according to similar difficulty levels; the second to perform a content analysis of the group of items to determine whether there was a common theme or interpretation of the underpinning construct. That is, was there a common set of underpinning skills and knowledge that were representative of the clustered items?

An interpretation of the items that were clustering together in terms of difficulty levels indicated that there could be as many as six band levels. A content analysis of these items sought to identify a commonality within the each cluster and the themes that emerged related to different interpretation frameworks. The result of this is presented in the table below.

Table 4: Interpretation of the *Facilitate People Management* levels from analyses of the scale

Performance level descriptors
<p>Expert Using an independent, proactive and innovative approach, can anticipate future HR planning requirements which link with the higher organisational plans. Formulates continuous improvement strategies in all facets of people management activities (including grievance and rehabilitation). Designs contingency plans as well as evaluates and modifies policy and procedures. Diagnoses systemic issues and takes a long-term perspective with the design of solutions.</p>
<p>Experienced Under own initiative can analyse, evaluate and improve HR planning and performance management processes. Has an in-depth understanding of a range of performance management processes, issues and contingency strategies. Promotes and empowers staff to develop self-awareness so to develop transferable skills and knowledge and to negotiate work performance plans. Applies consistently successful counselling and negotiation techniques. Values and fosters a supportive workforce based on equity principles.</p>
<p>Competent Is able to anticipate future trends and issues related to potential grievances and complaints. Can identify, select and implement performance improvement strategies as well as take responsibility for continuous improvement of these. Is able to select and apply appropriate strategies for long-term career planning that benefits individual and organisation. In consultation, establishes and maintains effective rehabilitation programs within the workforce. Able to identify appropriate approaches to counselling employees, intervention strategies.</p>
<p>Approaching competence Under guidance, is able to develop and justify HR plans linking with higher corporate planning and in accordance with legislative requirements within their work unit. Can develop and establish performance management processes and identify future needs of business unit. Able to assess HR issues within their work environments. Can inform and communicate with staff about performance and grievance related issues. Demonstrates effective communication, sound counselling and accurate referral skills within grievance and counselling processes.</p>
<p>Below (Searching competence) Under guidance and assistance can select and apply appropriate HR planning and performance processes within organisational and legislative requirements for their work unit. Can develop workforce plans and examine resource and policy implications. Understands procedural fairness principles and participative management as well as the benefits of a consultative process.</p>
<p>Below (Low competence) Has limited demonstrated ability to implement people management strategies, plans and processes within the business unit/work area. Can undertake short term planning within defined workplace boundaries and alternatives.</p>

Variable maps provide an axis on which the cases (Xs) and the items (numbers) can be plotted, showing the persons' ability or level of performance and the items' level of difficulty. Being a partial credit model the items included item numbers and their step values, e.g. Item 1.1 Step 2 is recorded as 1.1.2. The variable map is based on a logit chart and in this instance ranges from 4.0 to -0.3.

A variable map is usually illustrated with all items and their step values fully integrated however for easy interpretation of the item and their step value numbers were separated according to the five sub-tasks (or elements of the unit of competency). The variable map below, Figure 1, indicates the variations between the hypothesised levels of difficulty for each item and its step (level descriptor) and the empirically derived (or observed) item estimates and includes the hypothesised band levels.

The item numbers have been colour coded to provide an analysis of each item and its behavioural descriptor (step) in regards to the hypothesised level of difficulty and the observed level of difficulty. The hypothesised level of difficulty of each item and its step has been represented in Table 2. On the following variable map, red items were hypothesised to be within band level 4 (highest level titled *Expert*), blue items were hypothesised to be within band level 3 (medium level titled *Experienced*), green items were hypothesised to be in band level 2 (lower level titled *Competent*), and black represents items that were hypothesised to be within the lowest band level (level 1 titled *Searching competence*).

It could be assumed that the hypothesised item difficulty estimates should appear within their hypothesised band levels. On the variable map the coloured items should read within their bands from the lowest black, then green, then to blue and finally, at the highest level of difficulty, red.

An analysis of the variable map indicates that in general the progression of the hypothesised level of difficulty of the items reflected the observed. The lowest band level items are coded black, the next two band levels are coded predominantly green and blue respectively and finally the band with the highest difficulty level items is predominantly coded red. The band with the greatest lack of correlation between hypothesised and observed was the band level titled *Expert*.

It is of issue that the band levels do not accurately reflect the progression and cut points regarding the levels of performance on the continuum that underpins the unit of competency. Within the VET sector it is assumed that all elements within the unit of competency must be achieved before a person can be judged and reported as competent (this is referred to as the conjunctive approach). For this unit of competency, the bottom band level on this chart and its cut point should indicate the level between competent and not yet competent (or in this case *Searching competence*). It was estimated that this cut point is where the candidate in the workplace should have demonstrated all elements of competency. The final element on the variable map (*Manage employee rehabilitation*) has a high level of difficulty, and in a large number of cases, people were unable to demonstrate competence. To achieve competence at this element the odds were that the candidates would have had to be experts in facilitating people management to be recognised as competent. This is similar to the third element (*Manage grievance*), whereby the odds were that a candidate would have to demonstrate a high level of competence to be considered competent for judgement and reporting purposes. Therefore the estimated cut-off point predicted by the subject matter experts should have been much higher than anticipated.

The next variable map, Figure 2, represented below is similar to the above variable map but includes the observed band levels that were empirically derived from the item logits and an analysis of the item cluster content. To the far right, the chart includes the gist statements of the empirically derived bands, which have been derived from an analysis of the cluster of items as well as the percentage of cases that fall within these bands. In this variable map there are six levels, and more than one band level falls below the cut point for competence. It is estimated that there are three levels that are above the cut-off point for competence. Figure 2 is linked with Table 4 in the interpretation of the empirically derived bands.

Within the VET sector a conjunctive approach to assessment is assumed, that is all elements within the unit of competency must be achieved before a person can be judged and reported as competent. This approach has led to three bands levels of the continuum being below the level of estimated competence (or cut-off point). It was at the band level related to *Competent* that all elements have been demonstrated. The final element on the variable map (*Manage employee rehabilitation*) had a high level of difficulty, and in a large number of cases, people were unable to demonstrate competence. At least one behavioural indicator should be demonstrated before competence can be determined.

A review of the variable map (Figure 2) indicates that the elements within the units of competency demonstrate a progressive increase in difficulty across the chart. Clearly the elements have wide range of varying levels of difficulty. The most difficult element is *Manage employee rehabilitation*, followed by *Manage grievances* and then *Counsel employees*. The logit range of these three elements for all items and their step levels range from 3.2 to -0.53. The logit range of the previous two elements for all items and their step levels range from 1.49 to -2.41.

Logit	Persons	Items according to Elements of Competency										Gist statement	
4.0												Expert 16%	
3.0											18.1.4		
2.0	XX										12.1.2		
	XX										12.2.3	17.1.4	
	XX										13.2.3		
	XXXXX	2.2.3	4.2.3		7.1.3							17.1.3	
		1.2.3											
		3.2.4										17.2.2	
1.0	XXXX XXX XXXXXXX XXXXXX XXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXX				5.1.3	6.1.4	11.1.3					17.2.2 18.1.3 17.3.2	
					6.2.3					16.1.4			
					10.1.3								
		2.1.4								15.1.2		17.1.2 18.1.2	
					5.2.3	6.1.3	9.1.3		12.2.2				
	XXXXXXXXXXXXXXXX	4.2.2			8.1.3					13.2.2		17.1.1 17.2.1 17.3.1 18.1.1	
	XXXXXXXXXXXXXXXX	1.1.3	3.1.2							16.1.3		Competent 51%	
	XXXXXXXXXXXX				6.2.2	7.1.2				14.2.2			
.0	XXXXXX	2.1.3	3.3.3		5.2.2	6.1.2	10.1.2			13.1.2	14.1.2		
	XXXXXXXX	2.1.2	3.2.3		5.1.2	9.1.2			12.1.1	16.1.2			
	XXXXXXXX	1.2.2										Approaching competence 31%	
	XXXXXXXX	2.2.2			11.1.2								
	XXXXX	4.1.1			8.1.2				12.2.1	14.1.1	14.2.1	Below (Searching competence) 31%	
	XXXXX	3.2.2	4.2.1		5.2.1	9.1.1				13.1.1	16.1.1		
	X	3.3.2											
-1.0	XXXX	1.1.2			6.2.1								
	XXXX	1.2.1			8.1.1					13.2.1			
	XX	7.1.1											
	X												
	XXXX	2.2.1			6.1.1	11.1.1						Below (Low competence) 9%	
	X												
		1.1.1	3.2.1	3.3.1	5.1.1					15.1.1			
-2.0		3.1.1											
		2.1.1			10.1.1								
	X												
-3.0		Undertake human resource planning		Manage the performance of individuals			Manage grievance procedures		Counsel employees		Manage employee rehabilitation		

Figure 2: The *Facilitate People Management* scale: Empirically derived band levels

The number of candidates that were not observed performing items was high for both *Manage employee rehabilitation* (n=142) and *Manage grievance procedures* (n=98). However, this does not necessarily mean that the candidates were unable to demonstrate the item due to their lack of skills and knowledge (and hence the difficulty of the item). However, in a competency based framework the onus is on the individual to provide evidence and prove competence.

Performance assessments are dependent on the candidate being provided with opportunities to demonstrate their skills and knowledge, for these particular candidates there may have been no opportunities to do so. This lack of opportunity may be a limitation of the use of the third party assessment as an effective single method for collecting evidence. Supplementing the third party assessment method with other methods such as interview, simulation or case study analysis should be considered by the task developers to enable the candidates to demonstrate their level of performance.

Given the disparity in difficulty levels across the elements it may be pertinent to separate the first two elements (items 1.1 to 11.1) from the last three elements (items 12.1 to 18.1). This would enable a separate judgement and reporting framework to ensure that candidates' achievement would be more effectively recorded and reported. A content analysis of the elements of competency supports a separation as Elements 1 and 2 (*Undertake human resource planning* and *Manage performance of individuals*) relate more to the daily functions of human resource management whereas the final three elements (*Manage grievance procedures*, *Counsel employees* and *Manage employee rehabilitation*) relate more to additional functions, skills and knowledge that are required when issues arise within the workplace. This is especially so with elements 3 and 5, which relate to major concerns or issues within the workplace occurring either through grievances or the rehabilitation of employees.

Units of competency were developed for the industry by subject matter experts and in this instance a unit of competency has been developed that, although it has one underlying construct, contains elements that are too disparate in terms of difficulty levels.

Discussion

The results obtained in this study and the processes undertaken to develop an effective and efficient assessment instrument are encouraging. However, even though the results are satisfactory, a number of changes or modifications should be considered in any further application of the procedure.

Theoretical implications

The importance of using subject matter experts in the development of items and performance indicators as well as the determination of cut-off scores was endorsed by the results of this study. In this study a small team of subject matter experts (n = 10) were selected for their varied experience, skills and knowledge of the workplace context and the subject matter under review. Considering the outcomes of the study it is suggested that this team was sufficient for the required purpose.

Careful selection, training as well as the skills and knowledge of the experts was critical to the success of the process. The subject matter experts undertook training as a group, which enabled extensive discussion and reflection amongst the team. The training occurred immediately prior to the standard setting process and subject matter experts were essentially 'trained on the job'. Although this was a method supported in a number of reports in the literature it is suggested that the initial stages of the standard setting procedure, which included development of item and performance indicators, could have spanned a greater period of time. It is proposed that more time could have been spent determining the cut-off scores to enable greater discussion and reflection.

The standard setting procedure enabled a flexible determination of the number of iterations of the feedback cycle to promote the accuracy of the assessment tool developed as well as the determination of the cut-off scores.

A standards referenced framework allows for reporting of results in a range of ways including the dichotomous competent/not yet competent, grades or differentiating scores (Griffin & Gillis 2001) and provides assessment practitioners with a methodology that would enable the assessment and reporting of levels of performance to be applied in a consistent manner. It is suggested that further

iteration of the standard setting procedure combined with empirical feedback could provide greater accuracy of decision of the subject matter experts.

Policy implications

The assessments were conducted in the context of a competency based framework within the Australian VET sector and as such the policy framework for decision making is critical to determining cut-off scores. Kane (1998) considered that much of the arbitrariness in standard setting derives from the need to make policy decisions when developing performance standards (such as how good is good enough?). A conjunctive approach to decision making is accepted across the Australian VET sector; that is, to be determined competent it is expected that candidates will perform to a minimum standard across all items (elements or units of competency). This decision making policy has important repercussions for competency standards developers as well as assessment instrument developers and practitioners. Within this study greater attention to this aspect of decision making would have enabled the subject matter experts to make a more accurate determination of cut-off scores.

In addition, competency developers should consider the level of difficulty of sub-tasks (elements) and the policy framework (conjunctive approach to assessment) in which assessment is conducted in the VET sector when developing units of competency. Careful consideration of not only whether the elements have a similar underlying construct but also whether they have a similar level of difficulty is critical to the development process.

A key aspect of continuous improvement of competency standards is the review and validation process every three to five years of the 'life' of the Training Package. The use of item response theory and the Rasch partial credit model would provide valuable feedback to subject matter experts to assist the review of competency standards. The use of item response modelling and in particular the partial credit model would provide an effective process for the selection, review and refinement of polytomous items and the instrument overall. This is especially so if assessment practitioners are using a standards referenced framework to determine levels of performance.

Within this study the partial credit model was used to calibrate the instruments and identify items and performance indicators that did not fit the model. This model also enabled the determination of the difficulty of the items and their performance indicators. Subject matter experts were not provided with this information in this stage of the study as the information gained was used to evaluate their effectiveness in developing items and performance indicators and in determining cut-off scores. In any further study this stage of the development process would be incorporated into the iterative process and would result in an improvement to the overall instrument.

Practical implications

Given the limitations associated with third party assessments and with the unit of competency under review it is prudent to suggest that in any assessment multiple methods of assessment should be used. The use of multiple assessment methods has been advocated throughout the Australian competency based assessment literature. Supplementing the third party assessment method of evidence collection with other methods should be considered by task developers to promote the validity of the assessment judgements and to enable candidates to demonstrate their level of performance without being limited by the disparate level of difficulty of the elements.

The literature supports the view that when utilising third party assessments it is recommended that the behaviourally anchored rating scales be used (Smith & Kendall 1963) and that by increasing the number of raters and then averaging (or synthesising) the ratings is a method for negating subjectivity (and hence inaccuracy) (Fletcher, Baldry, & Cunningham-Snell 1998). Hence to synthesise multiple sources into an overall estimate of competence level as opposed to reporting just discrepancies in ratings of various observer groups, is a method of overcoming a number of limitations.

Further research implications

Both the encouraging outcomes of the study and the evidence of some limitations suggest that there are opportunities for further research or extension of this work.

In terms of this particular study, the use of item response theory to evaluate indicators of performance and to determine cut-off scores have provided information that should be fed back into the continuous improvement of the assessment instrument. In addition, the process of the development of the instrument suggests that the standard setting procedures using subject matter experts could be used for the development of other instruments used to assess levels of performance.

Of interest in this study is the use of third party assessments to determine workplace competence. Traditionally assessments of higher order competencies have centred on the use of portfolio evidence. The wider study of this research project also investigated the use of multiple raters with an assessor making an holistic judgement given the ratings of multiple observers (raters). The synthesis of multiple sources into an overall estimate of competence level as opposed to reporting just discrepancies in ratings of various observer groups, is a method of overcoming a number of limitations of third party assessment.

Concluding remarks

The results obtained in this study support the contention that a standard setting multi-stage methodology using subject matter experts proved adequate for the purpose of predicting or determining performance rubrics within a graded competency based assessment and reporting framework. The methodology chosen had a sound theoretical base and is sufficiently flexible to be used with a variety of different industry and educational contexts.

There are several observations to make about the data and about the specialist judges. The first is the rank order of the judges' placements agreed substantially with the empirical evidence developed using the Rasch model. The second is that the assumption of equal difficulty of the elements was not supported by the data. There were subtle but important differences in the element difficulties although they all fitted the model and were therefore mapping the same underlying construct of competence.

This has important implications for the development of standards referenced frameworks in vocational education and training. The purpose of the study was to investigate the agreement between the judgements made by specialists and an empirical analysis of data in developing a standards referenced framework. The amount of agreement of the two procedures indicates that a panel of specialists using a consensus moderation approach to the placement of quality indicators on an underlying continuum has potential for accurate framework development.

Consequently, it means that large-scale data collection may be obviated by the use of specialist panels which will lead to large savings in time and money in the development and roll out of units for Training Packages that develop assessments this way. There is also a considerable professional development component built into the approach.

The iterative feedback to specialists may also be important. However it needs to be noted that this study focussed on one unit of competency. The panels were also 'learning on the job'. Once the panel became experienced in this task they could become far more proficient and able to place the criteria on the continuum with a great deal more alacrity, thus increasing the efficiency and effectiveness of this approach to profile or standards reference development.

It would also be interesting to encourage the specialist panels to allow for differences in task or element difficulty within a unit of competency. The evidence in this study supports the contention that there are differences in the difficulty levels of the elements and that the assumption of equality may be unfounded.

Of final importance are the assessment of the candidates and the effectiveness of the approach for identifying training needs of all candidates. It has yet to be demonstrated but the implication of the approach is that training to improve performance quality can be offered to all candidates. In the conventional application of competency based assessment those workers assessed as not yet competent are offered training opportunities. However the logic of standards referenced framework leads us to the conclusion that everyone can be offered training to improve performance and the artificial dichotomy of competence may be anachronistic in a quality based workforce.

References

- Adams, R. J. & Khoo, S-T. 1993, *Quest: The interactive test analysis system*, Australian Council for Educational Research, Hawthorn.
- ANTA 1999, Public Services Training Package, ANTA, Melbourne.
- Bennett, J. 1998b, 'Setting standards and applying them across different administrations of large-scale, high-stakes, curriculum-based public examinations', Occasional paper, Board of Studies, NSW.
- Benner, P. 1984, *From novice to expert: excellence and power in clinical nursing practice*, Addison-Wesley, Menlo Park, pp. 13-34. Retrieved July 2002 via the world wide web at www.sonoma.edu/users/n/nola/n312/benner.htm
- Bondy, K. N. 1983, 'Criterion-referenced definitions for rating scales in clinical evaluation', *Journal of nursing*, vol. 22, no. 9, November 1983, pp. 376 – 382.
- Curtis, D. D. & Denton, R. 2002 draft, The Authentic performance-based assessment of problem-solving, NCVER, Adelaide.
- Fletcher, C., Baldry, C. & Cunningham-Snell, N. 1998, 'The psychometric properties of 360 degree feedback: An empirical study and a cautionary tale', *International Journal of selection and assessment*, Vol 6, No 1, January 1998, pp. 19-34.
- Gable, R. K. & Ludlow, L. H. 1990, 'The use of classical and Rasch latent trait models to enhance the validity of affective measures', *Educational and psychological measurement*, Winter90, vol. 50, issue 4, pp. 860-879.
- Glaser, R. 1963, 'Instructional Technology and the measurement of learning outcomes: some questions', *American psychologist*, vol. 18, pp. 519-521.
- Griffin, P. 1997, *An Introduction to the Rasch model: Measuring achievement using sub tests from a common item pool*, unpublished draft, Assessment Research Centre, University of Melbourne, Melbourne.
- Griffin, P. & Gillis, S. 2001, *Competence and quality: Can we assess both?*, Paper presented at the Upgrading Assessment conference, Melbourne.
- Griffin, P., Gillis, S., Keating, J. & Fennessy, D. 2001, *Assessment and reporting of VET courses within senior secondary certificates*, NSW Board of Vocational Education and Training, Sydney.
- Jaeger R. M. 1989, 'Certification of student competence', in R L Linn (ed), *Educational measurement*, 3rd edn, American Council on Education and MacMillan, New York, pp. 485-514.
- Kane, M. 1998, "Choosing between examinee-centred and test-centred standard setting methods", *Educational assessment*, vol. 5, no. 3, pp. 129-145.
- Lai J-S., Haglund, L. & Kielhofner, G. 1999, 'Occupational case analysis interview and rating scale', *Scand J Caring Sci*, vol 13, pp. 267-273.
- Lunz, M. E., Wright, B. D. & Linacre, J. M. 1990, 'Measuring the impact of judge severity on examination scores', *Applied measurement in education*, vol. 3, no. 4, pp. 331-345.
- McGaw, B. 1996, *Their future: Options for reform of the Higher School Certificate*, Department of Training and Education Co-ordination, Sydney, New South Wales. Retrieved July 2001 via the world wide web at www.boardofstudies.nsw.edu.au
- NSW Government 1997, *Securing their future: The New South Wales government's reforms for the Higher School Certificate*, NSW Government Printer, Sydney, New South Wales. Retrieved December 2001 via the world wide web at www.boardofstudies.nsw.edu.au
- Smith, R. M. 1991, 'The distributional properties of Rasch item fit statistics', *Educational and psychological measurement*, Fall94, vol. 51, issue 3, pp. 541-566.
- Smith, R. M. 1994, 'A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances', *Educational and psychological measurement*, Spring 94, vol. 54, issue 1, pp. 42-54.
- Smith, P. C. & Kendall, L. M. 1963, 'Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales', *Journal of applied psychology*, vol. 47, pp. 149-155.
- Wagh, R. F. 2002, 'Creating a scale to measure motivation to achieve academically: Linking attitudes and behaviours using Rasch measurement', *British journal of educational psychology*, vol 72, pp.65-86.
- Wilczenski, F. L. 1995, 'Development of a scale to measure attitudes towards inclusive education', *Educational and psychological measurement*, April 1995, vol. 55, issue 2, pp. 291-300.
- Williams, M. & Bateman, A. 2002, Graded assessment in VET: An analysis of national practice, drivers and areas for policy development, NCVER, Adelaide.
- Wilson, M. & Iventosch, L. 1988, 'Using the partial credit model to investigate responses to structured subtest', *Applied measurement in education*, vol. 1, issue 4, pp. 391-334.
- Wolfe, E. W. & Miller, T. R. 1997, 'Barriers to the implementation of portfolio assessment in secondary education', *Applied measurement in education*, vol. 10, no. 3, pp. 235-251.
- Wright, B. J. & Masters, G. N. 1982, *Rating scale analysis*, Mesa Press, Chicago.
- Wright B. D. & Stone M. H. 1979, *Best test design*, Mesa Press, Chicago.