An integrated 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 Australian Association for Research in Education Annual Conference

Brisbane, Australia, December 2002

Abstract

This paper presents preliminary findings from a project 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 a strategy to synthesise multiple sources of evidence to inform judgements of workplace competence. The methodology adopted integrates developments in two fields of study, performance appraisals and psychometrics. The method encapsulates features of a 360-degree feedback model that is widely used in performance appraisals, in which ratings are obtained from a supervisor, peers, subordinates, clients, and the assessment candidate (self-assessment). Item response (Rasch) modelling techniques are employed that allow for the identification of the competence of the candidate, the difficulty of the assessment task, and the influence of the source of evidence (e.g. supervisor, peer or subordinate ratings). Instrument development and preliminary data analysis are discussed.

Introduction

This paper presents preliminary findings from an Australian Research Council (ARC) funded project investigating the application of a multi source measurement approach to the assessment of higher order competencies in the public service and public safety industries. The SPIRT (Strategic Partnerships with Industry in Research and Training) project involves a collaboration between The University of Melbourne, Department of Defence (Education and Training Policy), and the Emergency Management Australia Institute (EMAI). The aim of the project was to develop and validate a strategy to synthesize multiple sources of evidence to inform holistic judgements of workplace competence. This paper presents the initial analysis of data obtained from the trial of assessment materials within the Department of Defence and other public service organizations (*N*=46 assessment candidates).

The SPIRT project methodology integrates developments in two fields of study, performance appraisals and psychometrics. The method encapsulates features of a 360-degree feedback model that is widely used in performance appraisals, in which ratings are obtained from a supervisor, peers, subordinates, clients, and the assessment candidate (self-assessment). The methodology expands the traditional observation basis of this methodology by allowing

for the inclusion of other sources of evidence, such as a portfolio or interview. Current advances in psychometrics are also utilised. Item Response (Rasch) models allow for the identification of the competence of the candidate and the difficulty of the assessment task or method encountered. A multi-faceted Rasch model can also identify and control for the influence of the source of evidence (e.g. supervisor, peer or subordinate ratings) on estimates of competence.

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 reform agenda (now known as the National Training Framework) that is expected to improve Australia's economic competitiveness within the Asian Pacific region . 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 . Competence can be thought of as the ability to use and integrate a variety of skills and knowledge to solve real workplace problems . Officially, competency was defined by the National Training Board 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 . 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 . 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"* . In a criterion referenced framework tasks or competencies can be arranged along a progression or continuum of development, and individuals of varying ability can be positioned along this continuum.

Within a CBA framework, an assessment conducted for the purpose of national recognition under the Australian Quality Training Framework (AQTF) must be undertaken by a qualified assessor who satisfies the assessor requirements of the industry training package. The assessor is responsible for planning, conducting and reviewing the assessment. The assessor must have relevant industry experience and demonstrated technical competence at least to the level at which they are assessing. Alternatively, if the assessor does not posses the relevant technical competence, they must work in conjunction with a subject matter expert (SME) during the assessment.

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. These competencies may include establishing rapport with a client, critical thinking, creativity, and reflection on performance. 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 .

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 . Thus, the importance of integrating evidence from multiple sources for the assessment of these competencies has been extensively documented in CBA literature .

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 . As a result, there has been an increase in the use of 360-degree feedback for the assessment of management and other higher order competencies , but this approach has yet to be used for certification or licensure purposes. According to Church and Bracken , information obtained using this process represents the next standard in performance evaluation and the assessment of managerial competence.

Multi-source assessment

The terms multi-source appraisal, multi-rater appraisal and 360-degree feedback are all used interchangeably in the literature. While the terms appear to vary slightly in their definitions, the central concept is that performance ratings are obtained on an individual from a range of sources such as supervisors, peers, subordinates and the individual themselves . For the purposes of this project, the term 360-degree assessment was adopted to distinguish the process from 360-feedback, in which an assessment decision is not necessarily made, but rather feedback is provided to the individual on their strengths and weaknesses. Critical to the success of the process is that raters 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 .

This method of assessment 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 raters is thought to produce more reliable and valid results . It is suggested that 360-degree assessment offers more objective assessments as multiple raters provide a fairer and less biased view of performance . Another proposed benefit of 360-degree assessment is that different raters may provide unique information about the individual because they interact with them in different capacities . In other words, these different constituents may all contribute uniquely to the evaluation. Thus, 360-degree assessment appears well suited to the assessment of higher level management competencies, given their inherent complexity . Importantly, 360-degree assessment also has the advantage of allowing for real time, on the job assessment of performance with

minimal disruption to workplace activities , and holds benefits for assessment candidates, who would likely find feedback from a variety of sources as fairer and more accurate than any single evaluation .

Within a CBA framework, multi-source assessment can be distinguished from 360-degree assessment through the inclusion of additional sources of evidence, such as portfolio or interview, to compliment the workplace observer ratings. A portfolio is a selection of annotated and validated information that illustrates the candidate's depth and breadth of skills and knowledge. It allows for a broad range of evidence to be gathered and is particularly useful in the management context as management projects and initiatives often occur over an extended period of time. Portfolios are seen as authentic, in that they refer to collections of performances in naturalistic settings, and for that reason they are held to have advantages over other forms of assessment . Importantly, portfolio pieces require explanatory narrative to clarify the context of each document and make it more comprehensible to the assessor. Thus the portfolio contains two distinct types of material; evidence and claims in which the candidate argues that he or she has satisfied the criteria .

The interview method of assessment is ideal for assessing underpinning knowledge and understanding. It is appropriate for examining a candidate's ability to synthesise and organise ideas, and is useful for assessing application of knowledge in new or unfamiliar situations. Interview questions can come in simple (e.g. clarification of processes) or extended form (e.g. verification of knowledge and understanding, examining attitudes and values). A major advantage of this form of assessment is the interview can be conducted following a review of portfolio or other evidence to clarify areas not covered adequately by these other methods.

The implementation of multi source assessment for the assessment of competencies is in line with the extensive CBA literature that argues the importance of holistic assessments and the integration of evidence from multiple sources . The use of such techniques for the assessment of management competencies is invaluable given that they are particularly difficult to assess . 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 . A method is needed for synthesising evidence from multiple sources to formulate an overall judgement of the competence of a candidate. An aim of CBA is to determine the competence of a candidate 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 . This is because the purpose of assessment is to infer candidate competence beyond the sample of tasks used to estimate competence .

Research aims

This project is investigating an innovative approach to the assessment of higher order competencies in the public service and public safety industries. The primary research question being investigated is *"To what extent can multiple sources of evidence be synthesised to inform the judgement of higher order competencies?*" Also of interest is an investigation of the extent to which candidate competence vary within industries, and the extent to which different rater groups vary in judgement stringency. Further, the extent to which a developmental progression of competency acquisition can be defined is of particular interest. If a meaningful variable can be constructed, it is of interest to explore if there are discernable stages along this variable, and to what extent an empirically derived interpretation corresponds to a qualitatively proposed definition.

Method

Sample

At present, 46 volunteer candidates have been assessed as part of the project, including 36 human resource and specialist managers from Defence, and six managers from other public service organisations (Centrelink, *N*=6). All 46 candidates participated in 360-degree assessment. In addition to the 360-degree assessment, 37 candidates also participated in an interview, while three candidates also completed portfolios.

Of the 46 candidates who undertook 360-degree assessment, observer record forms (360-degree assessment instrument) were completed by 43 supervisors (*M*=0.93, *SD*=0.71), 68 peers (*M*=1.46, *SD*=0.94), 55 subordinates (*M*=1.2, *SD*=1.29), and 3 clients, all candidates also completed a self-assessment. On average each candidate was rated by 3.65 (*SD*=1.54) raters 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 candidates role ; .

Unit of Competency

The unit of competency used in this project is entitled Facilitate People Management (PSPMNGT603A) from the Public Services Training Package (PSTP, 1999). The unit is related to the management working area and covers the implementation of people management strategies, plans and processes within the business unit in cooperation with specialist human resource personnel. This unit contains five elements and 23 Performance Criteria. The Elements contained within the unit of competency are 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 include an integrated demonstration of effective people management strategies, which were expected to facilitate the attainment of business unit objectives. The unit contributes to awards at the Advanced Diploma level.

360-degree instrument

As recommended, the rubrics for the 360-degree instrument were developed by a group of subject matter experts (SME) drawn from a cross-section of Defence workplaces, thus representing a variety of perspectives . Rubrics are *"a set of scoring guidelines that describe the characteristics of the different levels of performance used in scoring or judging a performance"* , 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; . 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 . 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 .

The 360-degree assessment instrument consists of 30 items designed to assess all aspects of the unit of competency to ensure adequate coverage of all the components of competency. Each item consists of a prompt statement (based on the performance criteria) followed by a series of behavioural descriptors. These descriptors detail typical managerial behaviours. The descriptors were arranged in order of increasing levels of competence, or

the level of skills and knowledge required for performance , and were coded numerically from 1 (representing the lowest quality of performance). The descriptors are distinguished by the degree of strategic and lateral thinking, and intellectual application involved in the management processes, the degree of autonomy with which the manager functioned, and the amount of insightfulness, leadership and intuitiveness demonstrated. An example item (item 17) is presented in Table 1. The items differed in the number of behavioural descriptors as each item was conceptually different, however no more than four descriptors were written for each item as a large number of descriptors can make rating and interpretation difficult (Griffin, 2001).

The descriptors were written, where possible, to be directly observable to members of the candidate's workplace (supervisor, peers, subordinates), as is the case with behaviourally anchored rating scales (BARS, . When items are based on overt behaviours ratings are more likely to be accurate, disagreement between raters becomes less likely, unfavourable ratings are generally easier for a candidate to accept, and there is a more substantial basis for the development of a training program for the candidate (Griffin & Gillis, 2001). The behavioural descriptions were also designed to be highly relevant to the job role function of managers, to allow differentiation between managers at different levels of competence, and to be amenable to change (growth) through training and development.

The raters (or observers) were required to select the behavioural descriptor that best described the skills and knowledge demonstrated by the candidate. Observers were also presented with the option of "have not observed the candidate applying these skills and knowledge" for items for which they had little or no knowledge of the candidate's performance. This phrase was chosen to be consistent with the current competency assessment terminology with which the participants would be familiar. For the self-assessment instrument the not observed option was replaced with "have not yet had the opportunity to apply these skills and knowledge" as this most closely reflected the response category on the observer record form.

Rasch Analysis

Rasch modelling techniques are used to calibrate items or tasks along an underlying construct or trait , in this project a continuum of people management competence. The calibrated items are considered indicators of the construct or trait being measured. The process of calibration establishes the difficulty of the items and the ability (or competence) of the assessment candidates, and positions them both on the same measurement scale (*logit* scale). This allows for a direct comparison of candidates and items. Calibration also establishes the accuracy of items as measuring devices .

The simple logistic Rasch model states that the probability of a person (*n*) answering an item (*i*) correctly is dependent only upon the ability (or competence) of the person (θ_n) and the difficulty of item (δ_i), and can be expressed as in equation 1.

$$\Pr(X_{ni}=1;\theta_n,\delta_i)=\frac{\exp(\theta_n-\delta_i)}{1+\exp(\theta_n-\delta_i)} \quad (1)$$

The simple logistic Rasch model can be extended to include additional facets of the assessment context, typically rater (or judge) severity, by the addition of an additional

(severity) parameter to equation 1. A multi-faceted Rasch model for dichotomous data (scored as correct or incorrect) can be expressed as in equation 2.

$$\Pr(X_{ni}=1;\theta_n,\delta_i)=\frac{\exp(\theta_n-\delta_i-\rho_r)}{1+\exp(\theta_n-\delta_i-\rho_r)} \quad (2)$$

In this model, 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 correct response, but also the severity of the rater (ρ_r) making the judgement. This could be considered equivalent to an adjustment in item difficulty. While these Rasch models are applicable only for dichotomous data, extensions of these models allow for data obtained using rating scales , or for partial credit to be awarded .

Multi-faceted Rasch models allow for the identification of the influence of the rater on the outcome of the assessment, and offer procedures for correcting assessment outcomes to account for this influence. Differences between raters can occur in a number of ways, but principally raters differ in their overall level of harshness or leniency .

The goal of Rasch models is the measurement of latent traits such as candidate ability or rater severity, and the establishment of these traits on a common linear (logit) scale . Multi-faceted Rasch models obtain from each candidate a non-linear raw score, which is then converted into a linear measure corrected for the particular rater, item or task encountered . Thus, in a three facet model each rating can be viewed as resulting from an interaction of the three facets, the competence of the candidate, the difficulty of the item and the severity of the rater. The multi-faceted Rasch equation states that the probability of a candidate being assigned a rating (*k)* on a particular item, from a particular rater can be predicted mathematically from these competence, difficulty and severity estimates .

The Partial Credit model ; allows for scoring one or more intermediate levels on an item, and to award partial credit for reaching one of these levels. Step difficulties are estimated for each movement from a score category (*k*-1) to the next highest score category (*k*) independently for each item. An item can thus be thought of as having one less step function than there are response categories . The Partial Credit model can be applied to situations where ordered response alternatives vary in number and structure across items ; , as in this project.

With most raters evaluating only one candidate the connectedness requirement of a multi-faceted Rasch model (Linacre, 1994) could not be satisfied if each rater was modelled to have a unique severity. As such, only rater group severity could be modelled. That is, each rater was modelled as belonging to a particular rater group based on their working relationship with the candidate. Rater group severity was modelled for supervisors, peers, subordinates and also self-ratings. This is similar to the group model described by Linacre (1994), in which raters can be conceptually grouped together when differences within groups are considered to be random, whereas differences between groups are thought to be systematic.

The multi-faceted Rasch model used for initial data analysis in this project is shown in equation 3.

$$\log\left(\frac{P_{nirk}}{P_{nirk-1}}\right) = \theta_n - \delta_i - \rho_r - \tau_{ik} \quad (3)$$

where

P_{nirk} is the probability of candidate *n* receiving a rating of *k* from rater group *r* on item *i*;

P_{nirk-1} is the probability of candidate *n* receiving a rating of *k*-1 from rater group *r* on item *i;*

θ_n is the competence of candidate *n*;

δ_i is the difficulty of item *i*;

ρ_r is the severity of rater group *r*; and

τ_{ik} is the difficulty of receiving a rating of *k* averaged across all raters groups for each item *i* separately.

Preliminary data analysis

Preliminary data analysis was performed using the FACETS software package . Parameter estimates are obtained using the unconditional maximum likelihood (UNCON), or joint maximum likelihood formula . Each facet of the assessment situation is modeled as an independent parameter (candidate competence, item difficulty, rater group severity), and candidate measures are calculated after the influence of the other facets has been removed . A logit value is produced for each candidate, item and rater group, and this is accompanied by a standard error statistic. Response thresholds ($\tau_1, \tau_2, \ldots, \tau_m$) are also estimated for the each rating scale step . Item difficulty is calibrated at the point on the continuum where the highest and lowest categories are equally likely to be awarded .

All parameter estimates are subject to error, and FACETS provides a variety of statistics to evaluate the amount of this error. The amount of error is largely dependent on the amount of data available . FACETS provides statistics regarding the extent to which the model *fits* the data, that is, whether the model is an appropriate basis for estimating the observed scores. General lack of fit of the data indicates that the model does not add up to one variable in which greater candidate competence results in higher ratings .

A review of the FACETS iteration report shows that the subset connection requirement was met, indicating that all measures could be estimated in one, unambiguous frame of reference . Similarly, the measurable data summary report shows a mean residual of zero, indicating that no estimation problems were encountered . Figure 1 displays graphically the calibration of the candidate, item and rater group facets. The first column provides the linear measurement scale (*logits*) on which each of the facets is estimated. The second column

displays the distribution of candidate competence while the third column displays the distribution of item difficulties. The distribution of candidate competence displays a slight negative skew, while the distribution of item difficulty displays a slight positive skew. There were five items with scale values well above the most competent candidate (items 19, 27, 28, 29 and 30), indicating that these are particularly difficult items. It can be seen that the majority of candidates are positioned between about -0.5 and 0.5 logits, with the exception of one candidate positioned at around -1 logits. The item difficulty estimates ranged from around -0.7 to 1 logit. The fourth column displays rater group severity estimates. With the exception of self-ratings, which appear to be more lenient (less severe), there is little difference between the severities of the other rater groups.

Candidates

The mean raw score for the sample was 216.3 (*SD*=81.9). This figure is not however very informative as candidates may have been rated by different numbers of raters. The mean candidate raw score per rater (candidate total raw score/*N* raters) was 46.25 (*SD*=8.92). In the present analysis the candidate facet was non-centered (centered facets are constrained to have a mean of zero). If more than one facet is non-centered then the frame of reference is not constrained sufficiently causing ambiguity . The mean competence estimate for the present sample was 0.15 logits (*SE*=0.09) with a standard deviation of 0.32 (*SE* =0.01). This value is close to zero, when taken with the distribution of candidate competence and item difficulties this suggests that the 360-degree instrument was reasonably well matched to the competence level of the candidates.

As can be seen in Table 2 candidate competence estimates ranged from -0.88 (candidate 18, the least competent) to 0.66 logits (candidate 17, the most competent), a range of only 1.54 logits. The standard errors for the candidate estimates were all acceptable (*M*=0.09,*SD*=0.01). Despite this relatively narrow spread, the overall difference between the competence level of the candidates was significant, $\chi^2(45)=472.3$, $p<.01$. The separation ratio (*G*) for the candidates was 3.35. This index compares the true spread of candidate competence measures with their measurement error, and indicates the spread of this sample of candidates . It is possible to determine the number of statistically distinct performance levels or discernible strata that are present in the data using the formula (4*G*+1)/3 (Fisher, 1992). Applying this formula, the candidates in the present study may be separated into five statistically different competence levels.

The candidate separation reliability was 0.92. This is a measure of the extent to which the instrument could successfully separate candidates of varying competence. The value is very similar to the Cronbach alpha for the 30 item scale (0.91). Like Cronbach alpha or the Kuder-Richardson 21 index of reliability, the coefficient represents the ration of variance attributable to the construct being measured (true score variance) to observed variance (true score variance and error variance) (McNamara, 1996). Values close to 1 suggest good reliability, while values less than 0.5 would indicate that differences between candidate competence estimates are attributable mostly to measurement error and not to actual differences in competence (Fisher, 1992).

Table 2 contains the infit mean square values, in their unstandardised and standardized form, for each candidate. This is a measure of the degree of *fit* between the observed ratings and the ratings expected by the model. The expected value for these mean square statistics is 1.0 when the model fits the data, and it has been suggested that an acceptable range is between 0.6 and 1.5 , or more conservatively between 0.7 and 1.3 . Higher values indicate more variability in the ratings than is expected, whereas values less than 1 indicate little variation in ratings, likely the result of identical or very similar ratings across all items.

Generally, infit mean square values greater than 1 are more problematic than values less than one.

With the exception of four candidates (candidates 2, 31, 32 and 39), all infit values are within the range of 0.7 to 1.3 (M=1.0, SD=0.2). Candidate 2 displayed the worst fit (infit mean square=1.5, standarised infit=4). The large infit value suggests more variation in the ratings assigned to candidate 2 than is expected by the model. An examination of the ratings assigned to candidate 2 reveals substantial differences in the ratings assigned by different raters. Candidate 2 was rated by four subordinates and completed a self-assessment. The self-assessment (M=1.87, SD=1.14) and the ratings assigned by one subordinate (subordinate 1, M=2.4, SD=1.35) were substantially higher than the ratings assigned by the remaining three subordinates (M=0.71, SD=1.18). The rating pattern of these other three subordinates shows frequent use of the "not observed" category (*M*=21.33, SD=3.06). This category was used much less frequently by subordinate 1 (*N*=4) and by the manager in their self-assessment (*N*=4). The mean ratings assigned by these other three subordinates based only on the items that they responded to (*M*=2.25, *SD*=0.68) resembles more closely the self-assessment and the ratings provided by subordinate 1. This indicates that these three subordinates only observed and responded to a small number of the items on the instrument, but rated the manager highly on these items.

This rating pattern could be the result of two scenarios. Firstly, the candidate and subordinate 1 may have rated accurately and it may be that the other three subordinates had not been exposed to a sufficient amount of the manager's workplace performance. Alternatively, it may be that the candidate and subordinate 1 have rated particularly leniently, and have perhaps provided ratings on some items for which they have only limited knowledge. In either situation an assessor would need to gather additional evidence before a decision of competence could be made. Interestingly, an examination of the length of time that each rater has worked with the candidate for reveals that subordinate 1 had worked with the candidate for 102 months, while the other subordinates have worked with the manager for an average of only 18.67 months (*SD*=13.31). This may suggest that the self-assessment and ratings from subordinate 1 are in fact representative of the candidate's competence, and that the other raters have not had sufficient exposure to the candidate's workplace performance to provide a complete evaluation. Such a finding has implications for selection of raters. Due to nature of human resource planning and people management, coworkers may need to work with a candidate for some time before they are in a position to provide a thorough evaluation of the candidate's management competence.

Items

As can be seen in Table 3, the item difficulties ranged from -0.74 logits (item 5, the easiest item) to 1.09 logits (item 19, the most difficult item), a range of only 1.83 logits. As can be seen in Figure 1, most items clustered at or around zero logits, with a group of six items at a lower level and five items spread out towards the top of the scale. Looking at the variable map it appears as if there are at least four clear levels of difficulty, based solely on statistical grounds. The standard errors for all items were acceptable (*M*=0.08, *SD*=0.02). Despite this somewhat restricted range, normally a range of -3.0 to 3.0 logits may reasonably be expected , the overall difference between the items was significant, $\chi^2(29)=1051.0$, $p<.01$, with a separation reliability of 0.97.

Examining the infit mean square values in Table 3, all items displayed acceptable fit, with all values between 0.9 and 1.3 (*M*=1.0,*SD*=0.01). As no infit values were below 0.7 it can be concluded that there are no redundant items in the instrument. Similarly, as no values were greater than 1.3 there is no evidence of pychometric multidimensionality, that is, a single measure of competence has been obtained (Myford & Wolfe, 2000; McNamara, 1996). If a

degree of multidimensionality was detected in the data then there may have been a need to report a profile of scores rather than an overall measure (Myford & Wolfe, 2000), perhaps at an element level.

Table 4 contains the rating scale statistics for the 30 items. As can be seen, the number of response categories for the items varies, as does the calibration of the rating scales across items. To examine if the response categories are appropriately ordered and clearly distinguishable, the average candidate competence measure for each response category should be examined . This value, labelled average measure (fit) in Table 4, is calculated by taking the average competence measure of all candidates receiving a rating in that particular category. If the rating scales are functioning correctly it is expected that the average candidate competence will increase with each rating category, suggesting that candidates receiving higher ratings posses a higher level of competence. Similarly, an oufit mean square index greater than 2 for any rating scale category suggests that ratings in that category for some candidates may not be contributing to meaningful measurement of the variable (Linacre, 1999c).

A review of the average measure for each score category reveals that only 13 categories (15 percent of the rating categories for the instrument) display a degree of misfit. Most of these appear to be the result of a lack of ratings (<15) in certain categories, as insufficient ratings in a score category result in the calibration for that category being unstable (Linacre, 1999), or appear insignificant (very small reduction in the average measure from one category to the next, and an oufit value close to 1). Only three rating categories displayed oufit values approaching 2 (item 3, category 3, oufit=1.9; item 12, category 4, oufit=1.7; item 26, category 1, oufit=1.8).

On examination of the ratings for item 3 it can be seen that there are only 15 ratings in category 2. This is the only item on the instrument that contains a middle rating category with a very small percentage of ratings. Clearly more data is needed to accurately estimate this item, however it is also possible that the descriptor for category 2 may be inappropriate given the small number of ratings. On examination, this descriptor (category 2) appears inconsistent with the other three descriptors. The descriptors for item 3 are presented in Table 5. As can be seen in Table 5, while the descriptors for categories 1, 3 and 4 all relate to staff competencies and the business unit, descriptor 2 refers to a business plan, something that has not previously been specified. Further, to understand the needs of the business unit (contained in descriptor 1), a candidate would require knowledge of the business plan. This item would require revision by SMEs and may require recoding (it may be appropriate to combine descriptors 1 and 2).

Category 4 from item 12 displays the lowest average measure of the valid rating categories for that item, and a relatively high outfit value (1.7). The descriptors for item 12 are shown in Table 6. The final descriptor (category 4) differs from the preceding three in that it does not require knowledge, or the application of performance management processes, schemes or strategies, but simply requires that staff are encouraged, through no formal processes, to improve their performance. Clearly the positioning and appropriatness of this descriptor would require revision.

The descriptors for item 26 would also require revision. The descriptors for item 26 are shown in Table 7. Category 1 displays a relatively high average measure (higher than categories 2 and 4) and a high oufit value. A review of these descriptors reveals that descriptor 2 extends on descriptor 1 by the addition of external agencies, but does not reiterate that this process is undertaken to facilitate employee performance and well being. These two descriptors should perhaps be combined, or at the very least descriptor 2 should be amended to reflect that referrals are undertaken to facilitate employee performance and

well being. The fourth descriptor for item 26, in contrast to the first three,does not require any specific management actions, processes or underpinning knowledge. This item would also require revision, and descriptor 4 may require deletion.

The range of item difficulties and thresholds indicates the range of competence levels that the instrument is able to accurately measure. Items must be sufficiently spaced along a variable to allow for an interpretation of a directional progression along the variable, that is, they must spread out in a way that demonstrates coherent and meaningful direction. If items are not spread out then all that has been defined is a position on a variable, not a variable . This can be investigated using the item separation ratio and also through an examination of the placement of items on the variable map . Item difficulties are displayed in Table 3 and thresholds are displayed in Table 4. Item difficulties only vary over 1.83 logits, while thresholds vary from -1.39 to 1.20 logits, a range of 2.59 logits. This indicates that the instrument, at present, is not measuring competence over a broad range, and more cases are required before a meaningful interpretation of the variable can undertaken. Given that the assessment trials are still in progress, and that the project is specifically targeting candidates across the full range of competence levels, it is expected that a sufficient range of competence will be measured.

Rater Groups

Table 8 shows the rater group severity estimates. As can be seen in Figure 1, with the exception of self-ratings, which display a degree of leniency (-0.31 logits), all rater groups tend to cluster around the same severity (0.07 to 0.13 logits). An examination of the fair average for each group also reveals that self-ratings are typically higher (1.89) than the other rater groups (1.45 to 1.52). The fair average is the mean rating for each rater group, adjusted for differences in candidate competence in each rater group's sample. The standard error of these estimates is satisfactory for all rater groups (0.2 or 0.3). The overall difference between the rater groups was significant, $\chi^2(3)=159.8, p<.01$, with a very high separation reliability (0.98). While this indicates actual differences in rater group severity, it should be noted that the range of severity estimates is more than 3 times smaller than the range of candidate competence or item difficulty estimates.

Further, the only ratings to display some variation are self-ratings, and as all candidates completed a self-assessment this apparent leniency in self-ratings should not advantage any candidates as might be the case if another rater group had displayed a similar pattern. Candidates only varied in which other raters evaluated them, and the difference between these other rater groups in severity is very small (0.06 logits). This is further illustrated in Figure 2, which shows a comparison of competence estimates (in logits) and (average) raw scores for each candidate. As can be seen, the candidate average raw score and competence estimate are very similar (r=0.99), indicating that rater group severity has little if any impact on candidate competence estimates.

An evaluation of infit for rater groups revealed that all raters had infit values within the acceptable range (all 0.9 to 1.1). The infit mean square value of 0.9 for self-ratings indicates a lack of variation in the pattern of ratings. This observation suggests the possibility of some degree of halo error in self-ratings. Halo error is the failure to discriminate between conceptually distinct aspects of performance . This finding is not however unexpected as research suggests that self-ratings of performance likely reflect a global evaluation of performance .

The observed leniency in self-ratings may in fact be the result of the treatment of not observed responses. When completing a self-assessment candidates would likely respond

to most items on the instrument, whereas other raters would likely select the not observed category on a number of items, depending on their degree of interaction with the candidate. As this not observed category is treated as a score of zero, the number of items that are scored in this way contributes to rater group severity estimates. The more not observed responses, the higher the severity estimate will be. An analysis of the use of this zero score category ("have not yet had the opportunity to apply these skills and knowledge" for self-assessments) by rater group reveals that this category was used quite regularly by supervisors (*M*=8.05, *SD*=5.93), peers (*M*=8, *SD*=7.19) and subordinates (*M*=7.39, *SD*=6.40), but much less frequently by candidates when conducting their self-assessments (*M*=3, *SD*=3.97).

To further explore this possibility, an additional multi-faceted analysis was performed in which the not observed category was treated as missing data. While this form of analysis is not appropriate for obtaining candidate competence estimates or variable interpretation, it does provide a comparison of rater group severity estimates. As can be seen in Table 9, in this secondary analysis self-ratings do not display the same pattern of leniency. In fact, the difference between the severity estimates of all rater groups is only 0.2 logits, with a much smaller separation reliability (0.79) and a chi square value ($\chi^2(3)=16$) approaching insignificance. This adds further support to the initial conclusion that there appears to be little difference between the severity of different rater groups, and as such this facet has little impact on assessment outcomes. 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 raters provide evaluations (Myford & Wolfe, 2000).

Conclusions and variable interpretation

In summary, a number of tentative conclusions are supported from this preliminary analysis, however further data is required to ensure stable and precise estimates of all parameters. Importantly, a broader range of candidates is required, as at this stage variation in candidate competence is only 1.54 logits.

The calibration of items using Rasch models allows for an investigation of the developmental progression of competency acquisition through an interpretation of the variable map. Masters discusses the notion of a *"progression of developing competence"*, with tasks or items calibrated along this progression. Using a variable map it is possible to identify varying levels of competence, and to identify the kinds of behaviours typically exhibited by individuals at these levels. This can be achieved through a content analysis of clusters of items at the same difficulty level .

Alternatively, this process can undertaken qualitatively by having subject matter experts (SME) position items (or descriptors) on a matrix according to their estimated difficulty (the level of skills and knowledge required). This process involves an iterative method outlined by Griffin (2000) in which the descriptors for one item are positioned, and then subsequent descriptors are compared with these initial placements, and positioned according to their relative difficulty. Using this approach it is possible to achieve qualitatively the same outcomes as item mapping procedures based on Rasch analysis. When clusters of items at the same difficulty level are examined, band level or profile descriptions can be developed. This interpretation is known as *standards referencing* whereby levels or bands are defined along the progression or continuum of competence for interpretive purposes (Griffin & Gillis, 2001). When item mapping is undertaken both qualitatively and empirically (Rasch based), the empirically derived profile descriptions can be used to validate the hypothesised construct. When further data has been gathered and stable threshold estimates can be obtained, a meaningful interpretation of the variable under consideration can be undertaken.

References

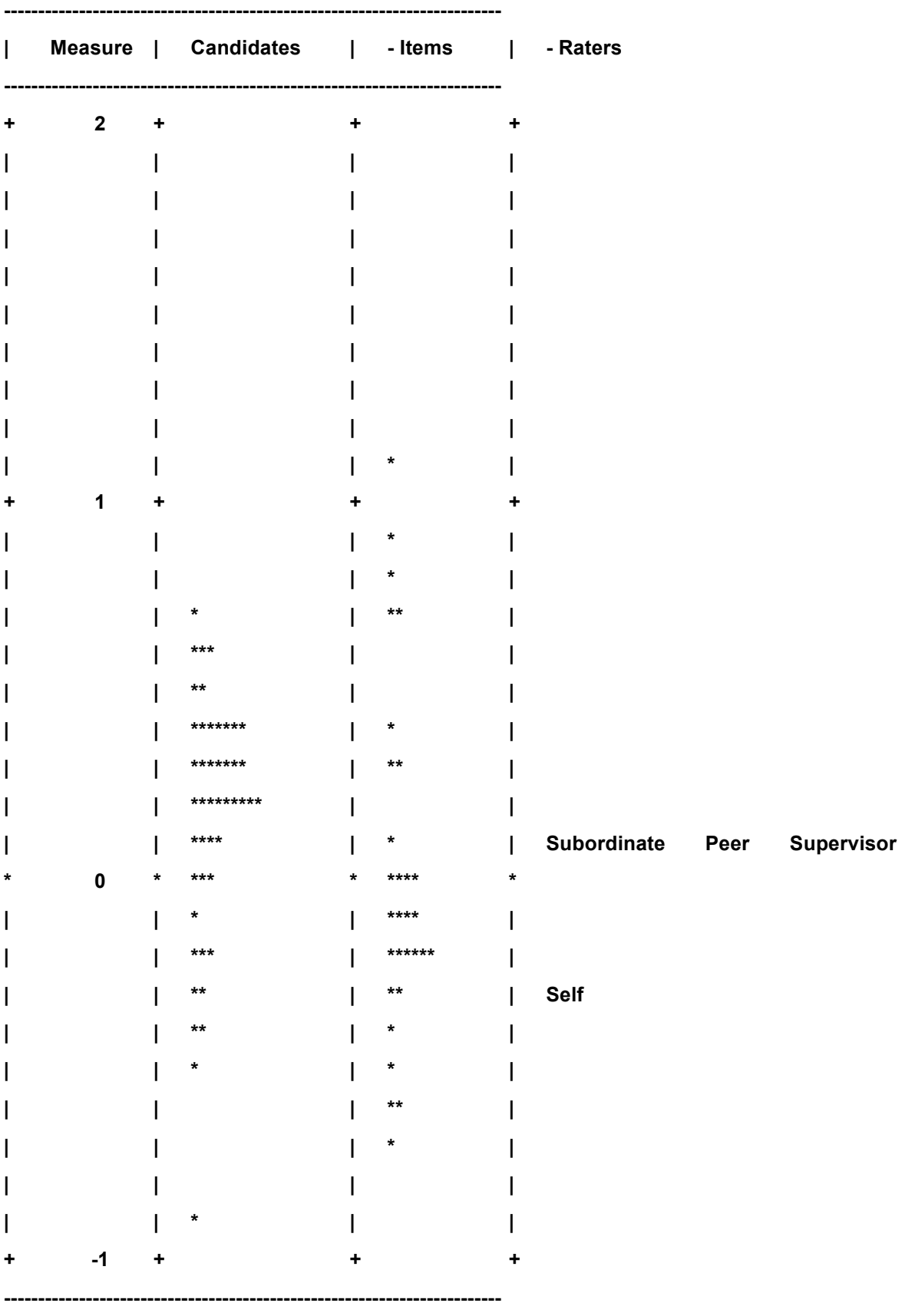

Figure 1: Distribution of Candidate Competence, Item Difficulty and Rater Severity

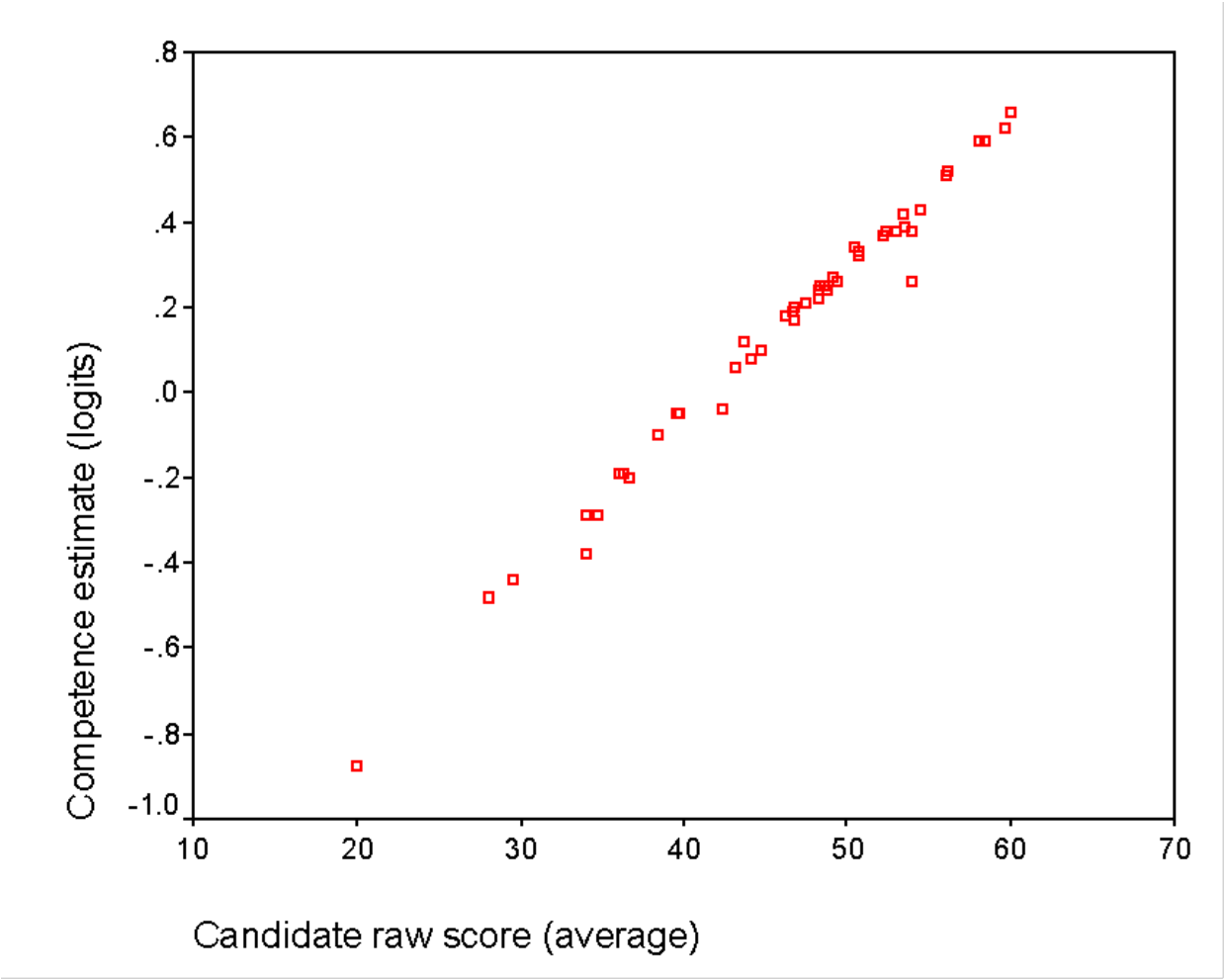

Figure 2: Plot of Candidate Raw Scores (Average) to Competence Estimates

Table 1

Example Item (Item 17) and Behavioural Descriptions

Item	**Behavioural descriptors**

	1	***2***	***3***
Information about training and development activities is made available to staff	Ensures information on training and development activities are available to staff	Customises and conveys information on training and development activities to individual staff against agreed training and development needs	Develops staff capability to seek out, evaluate and use information on training and development activities

Table 2

Candidate Competence Estimates, Standard Errors and Infit

Candidate	Competence Estimate	Error	Infit	Standardised Infit
18	-0.88	0.12	1.2	1
39	-0.48	0.09	1.4	3
27	-0.44	0.10	1.2	1
31	-0.38	0.14	1.4	2
3	-0.29	0.10	1.2	1
32	-0.29	0.11	1.4	2
2	-0.20	0.08	1.5	4
10	-0.19	0.10	1.2	1
41	-0.19	0.10	1.2	1
30	-0.10	0.08	0.8	-2

26	-0.05	0.08	1.1	1
38	-0.05	0.08	1.3	2
14	-0.04	0.11	1.0	0
11	0.06	0.08	0.9	-1
8	0.08	0.08	0.8	-1
37	0.10	0.09	0.9	0
40	0.12	0.07	1.0	0
45	0.17	0.09	0.7	-3
12	0.18	0.08	1.1	0
43	0.19	0.07	1.2	2
46	0.20	0.08	1.1	1
7	0.21	0.08	0.9	0
9	0.22	0.09	1.2	1
20	0.24	0.08	0.8	-2
35	0.24	0.09	1.2	1
13	0.25	0.08	1.0	0
22	0.25	0.08	0.8	-2
33	0.26	0.09	1.0	0
44	0.26	0.08	0.9	-1
42	0.27	0.08	1.2	2
24	0.32	0.08	1.0	0
4	0.33	0.08	1.0	0
28	0.34	0.05	1.0	0
16	0.37	0.08	1.1	0

5	0.38	0.08	0.9	-1
15	0.38	0.11	1.1	0
19	0.38	0.08	0.9	-1
6	0.39	0.10	0.9	-1
25	0.42	0.09	0.7	-3
23	0.43	0.10	1.0	0
1	0.51	0.10	1.3	2
36	0.52	0.10	0.7	-2
21	0.59	0.10	0.8	-1
29	0.59	0.10	1.0	0
34	0.62	0.12	1.1	0
17	0.66	0.10	0.9	0
Mean (N=46)	0.15	0.09	1.0	0.2
SD	0.32	0.01	0.2	1.8

Table 3

Item Difficulty Estimates, Standard Errors and Infit

Item	Difficulty Estimate *(logits)*	Error	Infit	Standardised Infit

5	-0.74	0.11	0.9	0
7	-0.64	0.08	0.9	0
1	-0.55	0.08	1.0	0
25	-0.53	0.11	0.9	0
8	-0.45	0.14	0.9	-1
17	-0.30	0.07	1.0	0
3	-0.28	0.05	1.2	3
15	-0.24	0.07	1.1	0
6	-0.23	0.06	1.1	1
10	-0.23	0.07	1.0	0
21	-0.22	0.09	0.9	-1
23	-0.21	0.08	0.9	-1
18	-0.20	0.07	1.0	0
11	-0.09	0.06	0.9	-1
12	-0.09	0.05	1.3	3
24	-0.08	0.08	1.0	0
4	-0.05	0.08	1.0	0
16	-0.05	0.06	0.9	0
2	-0.04	0.07	1.0	0
26	-0.02	0.05	1.2	2
13	0.00	0.07	1.0	0
14	0.06	0.07	1.0	0
9	0.31	0.07	1.0	0
22	0.33	0.08	1.0	0

20	0.40	0.07	1.0	0
28	0.67	0.09	0.9	-1
29	0.69	0.09	0.9	-1
27	0.76	0.05	1.0	0
30	0.92	0.06	0.9	0
19	1.09	0.11	1.0	0
Mean (N=30)	0.00	0.08	1.0	-0.01
SD	0.45	0.02	0.1	1.30

Table 4

Rating Scale Statistics for the 30 Items from 360-degree Instrument

Item	Score Category	Count	Average Measure (Fit)	Outfit	Step Difficulty	Step Difficulty SE	Thurstone Threshold
1	0	16	0.31	0.8			
	1	23	0.67	1.1	0.14	0.27	-0.64
	2	72	0.65*	1.0	-0.51	0.18	-0.14
	3	104	0.81	1.0	0.37	0.14	0.73

2	0	31	-0.04	0.9			
	1	50	0.14	1.2	-0.42	0.20	-0.92
	2	83	0.25	0.7	-0.34	0.15	-0.13
	3	51	0.26	1.1	0.76	0.16	1.02
3	0	14	0.09	0.9			
	1	67	0.45	1.5	-1.33	0.28	-1.39
	2	15	0.51	1.2	1.85	0.15	0.21
	3	52	0.41*	1.9	-0.80	0.15	0.33
	4	67	0.47	1.3	0.28	0.15	0.85
4	0	24	-0.06	0.9			
	1	44	0.19	1.2	-0.55	0.22	-1.10
	2	107	0.25	1.0	-0.71	0.15	-0.32
	3	40	0.22*	1.1	1.26	0.18	1.38
5	0	18	0.48	0.8			
	1	99	0.87	1.1	-0.92	0.25	-1.04
	2	98	0.97	1.0	0.92	0.14	1.04
6	0	16	0.03	0.9			
	1	36	0.29	1.2	-0.66	0.27	-0.99

	2	30	0.43	1.2	0.46	0.17	-0.17
	3	75	0.42*	0.9	-0.53	0.15	0.13
	4	58	0.44	1.2	0.73	0.16	1.02
7	0	14	0.50	0.9			
	1	33	0.55	0.8	-0.27	0.28	-0.71
	2	51	0.78	1.0	0.28	0.17	0.07
	3	117	0.89	0.9	-0.01	0.14	0.65
8	0	77	0.43	0.9			
	1	138	0.69	0.9			
9	0	58	-0.40	0.9			
	1	70	-0.10	1.4	-0.43	0.16	-0.82
	2	49	-0.06	0.8	0.22	0.14	0.07
	3	38	-0.04	1.0	0.21	0.18	0.76
10	0	21	0.07	0.9			
	1	71	0.37	1.1	-0.97	0.23	-1.17
	2	62	0.41	1.1	0.50	0.14	0.20
	3	61	0.48	1.0	0.48	0.16	0.99
11	0	46	-0.07	0.8			

	1	42	0.20	0.9	0.20	0.17	-0.47
	2	39	0.39	0.5	0.30	0.15	0.05
	3	88	0.36*	1.0	-0.50	0.15	0.43
12	0	22	-0.09	0.9			
	1	59	0.30	1.4	-0.92	0.23	-1.13
	2	39	0.30	1.2	0.60	0.15	0.00
	3	41	0.36	0.9	0.23	0.15	0.36
	4	54	0.19*	1.7	0.09	0.16	0.82
13	0	36	-0.13	0.8			
	1	69	0.18	1.2	-0.61	0.19	-0.91
	2	54	0.18	1.2	0.40	0.14	0.13
	3	56	0.28	1.0	0.21	0.16	0.81
14	0	32	-0.17	0.9			
	1	73	0.10	1.1	-0.84	0.20	-1.12
	2	68	0.23	0.7	0.17	0.14	0.06
	3	42	0.08*	1.2	0.67	0.18	1.05
15	0	31	0.16	0.9			
	1	28	0.38	1.3	0.33	0.20	-0.56
	2	69	0.43	1.3	-0.55	0.16	-0.13

	3	87	0.45	1.1	0.22	0.14	0.64
16	0	46	-0.10	0.8			
	1	41	0.20	1.2	0.19	0.17	-0.51
	2	49	0.33	0.6	0.01	0.15	0.00
	3	79	0.29*	1.0	-0.20	0.15	0.51
17	0	18	0.18	0.9			
	1	82	0.39	1.0	-1.20	0.25	-1.32
	2	51	0.50	0.8	0.90	0.14	0.36
	3	64	0.55	1.0	0.30	0.30	0.99
18	0	26	0.12	1.0			
	1	39	0.28	1.0	-0.20	0.22	-0.81
	2	83	0.37	1.1	-0.43	0.15	-0.15
	3	67	0.47	1.0	0.64	0.15	0.92
19	0	92	-1.06	1.0			
	1	108	-0.85	0.9	-1.12	0.14	-1.21
	2	15	-0.85	1.0	1.12	0.27	1.20
20	0	59	-0.39	1.0			

	1	70	-0.26	1.0	-0.50	0.16	-0.90
	2	56	-0.13	0.9	0.00	0.14	0.00
	3	30	-0.17*	1.2	0.50	0.20	0.90
21	0	54	0.18	0.9			
	1	53	0.28	0.8	0.31	0.16	-0.36
	2	107	0.50	0.9	-0.31	0.14	0.36
22	0	45	-0.42	0.9			
	1	92	-0.14	1.1	-0.98	0.17	-1.18
	2	49	-0.14*	1.2	0.48	0.15	0.19
	3	28	-0.02	1.0	0.50	0.21	1.00
23	0	57	0.16	0.9			
	1	45	0.29	0.7	0.52	0.16	-0.29
	2	112	0.49	0.9	-0.52	0.14	0.29
24	0	68	0.05	0.9			
	1	42	0.28	1.4	0.64	0.15	-0.26
	2	104	0.32	1.0	-0.64	0.14	0.25
25	0	24	0.47	0.9			
	1	103	0.62	0.9	-0.86	0.22	-1.00

	2	88	0.81	1.0	0.86	0.14	1.00
26	0	37	-0.07	0.9			
	1	35	0.24	1.8	0.05	0.19	-0.60
	2	37	0.19*	1.1	0.05	0.16	-0.13
	3	39	0.26	0.9	0.15	0.15	0.18
	4	67	0.21*	1.4	-0.26	0.15	0.56
27	0	124	-0.73	0.9			
	1	22	-0.45	1.2	1.08	0.15	-0.43
	2	32	-0.45	0.7	-0.93	0.16	-0.26
	3	21	-0.42	1.1	-0.04	0.19	0.05
	4	16	-0.43*	1.1	-0.11	0.27	0.53
28	0	123	-0.66	0.9			
	1	45	-0.37	0.7	0.47	0.14	-0.31
	2	47	-0.31	0.9	-0.47	0.17	0.30
29	0	130	-0.66	0.9			
	1	37	-0.36	0.7	0.71	0.14	-0.24
	2	48	-0.36	0.9	-0.71	0.17	0.24
30	0	119	-0.89	0.9			

	1	27	-0.68	0.6	0.67	0.14	-0.64
	2	22	-0.66	0.8	-0.51	0.16	-0.43
	3	39	-0.55	1.0	-1.20	0.17	-0.20
	4	8	-0.52	0.9	1.04	0.37	1.14

Table 5

Behavioural Descriptors for Item 3

Item	**Behavioural Descriptors**			
	1	***2***	***3***	***4***
Existing competencies of staff are compared with the needs of the business unit	Demonstrates knowledge of human resource plans, competency profiles and the needs of the business unit	Establishes links to business plans	Identified future skill requirements for the business unit	Promotes and provides opportunities for the development of skills and knowledge transferable to other Defence contexts (i.e. wider Defence network)

Table 6

Behavioural Descriptors for Item 12

Item	**Behavioural Descriptors**			
	1	***2***	***3***	***4***
Performance management processes that are applied to all staff are equitable and implemented in accordance with legislative requirements and organisational policies and processes	Applies performance management criteria and processes in accordance with legislative requirements and organisational policy and practices	Applies knowledge of different performance management schemes in accordance with legislative requirements and organisational policy and practices	Identifies and applies appropriate strategies to continuously improve performance management processes	Encourages staff self awareness and self appraisal to identify strengths and weaknesses

Table 7

Behavioural Descriptors for Item 26

Item	**Behavioural Descriptors**

	1	***2***	***3***	***4***
Referrals to appropriate support	Refers employees to appropriate	Refers employees to appropriate	Analyses, selects and plans	Maintains a supportive workplace

professionals and agencies are made to facilitate employee performance and well being	internal staff and agencies to facilitate performance and well being	internal and/or staff and agencies	approaches to counsel employees which may include referrals to external professionals and agencies	environment

Table 8

Rater Group Severity Estimates, Standard Errors and Infit

Rater Group	Number of Ratings	Severity Estimate*(logits)*	Error	Infit	Standardised Infit
Self	1380	-0.31	0.03	0.9	-2
Subordinate	1736	0.07	0.02	1.1	2
Peer	2040	0.11	0.02	1.1	2
Supervisor	1290	0.13	0.03	1.0	0
Mean	1611.5	0.00	0.03	1.0	0.7
SD	298.4	0.18	0.00	0.1	2.2

Table 9

Rater Group Severity Estimates when Not Observed is Missing Data

Rater Group	Number of Ratings	Severity Estimate*(logits)*	Error	Infit	Standardised Infit
Self	1184	-0.06	0.04	1.0	0
Peer	1450	-0.05	0.03	1.0	0
Subordinate	1186	-0.04	0.04	1.0	-1
Supervisor	866	0.14	0.04	1.1	2
Mean	1176.5	0.0	0.04	1.0	0.7
SD	199.6	0.8	0.00	0.0	1.2