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

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

Abstract

This paper presents preliminary findings from a study investigating the application of a multi-source measurement approach to the assessment of higher order competencies in the public service industry. The aim of the study 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. At present, 55 candidates have been assessed using a combination of assessment methods. This paper presents findings from a multi-faceted Rasch analysis of 360-degree assessment ratings. While candidate competence and item difficulty were well separated, there were only very small differences in the severity of different rater groups (self, supervisor, peer and subordinate). As such, rater group severity had no impact on candidate competence estimates.

Introduction

Competency based assessment

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

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

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

Assessing management and higher order competencies

CBA at higher levels refers to the assessment of covert, higher order competencies required for successful performance in professional and skilled work. While these higher order competencies are not confined to jobs at the higher levels of the Australian Qualifications Framework (AQF), their importance certainly increases at these levels (Hager & Gillis, 1995). The assessment of management competencies, such as decision-making, problem solving, leadership, conflict resolution, negotiation and strategic planning skills have traditionally involved the sole use of supervisor reports. These competencies are inherently difficult to assess as there is greater independence of action, less supervision, and the impact of decisions are often difficult to attribute to the person responsible due to the time lapse between the action and its consequences (Edmonds & Stuart, 1992). Thus, the importance of integrating evidence from multiple sources for the assessment of these competencies has been extensively documented in CBA literature (Griffin & Gillis, 2000).

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

Multi-source assessment

Within a CBA framework, multi-source assessment requires the use of a number of assessment methods or techniques. While candidates in the present study have been assessed using a combination of different assessment methods, this paper focuses on the analysis of 360-degree assessment ratings. The terms multi-rater appraisal and 360-degree feedback are used interchangeably in the performance appraisal 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 (Griffin & Gillis, 2000). 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 (Hurley, 1998).

360-degree 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 (Hurley, 1998). It is suggested that 360-degree assessment offers more objective assessments as multiple raters provide a fairer and less biased view of performance (Fletcher, Baldry, & Cunningham-Snell, 1998). 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 (Goudy, 1998).

Thus, 360-degree assessment appears well suited to the assessment of higher level management competencies, given their inherent complexity (Brutus, Fleenor, & London, 1998). 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 (Griffin & Gillis, 2000), 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 (Bozeman, 1997).

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 (Griffin & Gillis, 2000). The use of such techniques for the assessment of management competencies is invaluable given that they are particularly difficult to assess (Gregarus & Robie, 1998). Unfortunately, as is the case with CBA, research into multi-source assessment has not progressed at the same rate as its implementation in the workplace, with limited studies conducted in organisational settings using appropriate samples (Gregarus & Robie, 1998; Hurley, 1998).

Further, a method is needed for synthesising evidence from multiple sources to formulate an overall judgement of the competence 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 (Masters, 1993; Jaeger, 1989). This is because the purpose of assessment is to infer candidate competence beyond the sample of tasks used to estimate competence (Lunz & Wright, 1997).

Research aims

This study is investigating an innovative approach to the assessment of higher order competencies in the public service industry. 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 interest.

Method

Sample

At present, 55 candidates have been assessed using 360-degree assessment in combination with other assessment methods. Observer record forms (360-degree assessment instrument) were completed by 51 supervisors (M=0.93), 83 peers (M=1.51), 67 subordinates (M=1.27), and 3 clients. All candidates also completed a self-assessment. On average each candidate was rated by 3.71 (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 candidate's role (Thorndike, 1997).

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 workplaces, thus representing a variety of perspectives (Bennett, 1998). Rubrics are *"a set of scoring guidelines that describe the characteristics of the different levels of performance used in scoring or judging a performance"* (Gronlund, 1998, p. 225). Thus the central feature of rubrics are the ordered categories or levels of performance that comprise a description of the cognitive, affective and psychomotor skills embedded in competent performance (Griffin, 2000; Waltman, 1997). Underpinning the concept of rubrics is the criterion referenced interpretation in which an individual's achievement or competence is

described in terms of the tasks that they can perform (Glaser, 1981). The use of criterion referenced definitions for rating scales convey far greater information about the quality of performance, discriminates more accurately between individuals, and allows for candidates to be given more diagnostic feedback, feedback that they will likely perceive as more constructive and valid (Bondy, 1983).

Item writing for the 360-degree instrument involved a detailed analysis of all aspects of the Unit of Competency to ensure adequate coverage of all the components of competency as specified by the NTB (1992). The final 360-degree instrument consisted of 30 items based largely on the performance criteria of the unit. Performance criteria typically focus on task performance, however at the advanced diploma level there is a strong focus on the contingency management and transferability skill dimensions, thus it was necessary for items to expand on the performance criteria. Each item consisted of a prompt statement followed by a series of behavioural descriptors that 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 (as recommended by Goudy, 1998). 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 functions, and the amount of insightfulness, leadership and intuitiveness demonstrated. The descriptors were written, where possible, to be directly observable to members of the candidate's workplace as is the case with behaviourally anchored rating scales (BARS, Smith & Kendall, 1963). An example item is presented in Table 1. The behavioural descriptors for this item are coded 1, 2 and 3 in order of increasing competence (with 3 representing the highest level).

Rating process

Candidates, in conjunction with their assessor selected raters (or workplace observers) to complete the 360-degree instrument. According to Antonioni (2001), an objective criteria needs to be implemented for selecting raters, such as the extent of interdependency and the opportunity to observe work performance. Assessors considered whether raters were familiar with the skills and knowledge required to manage within the public service industry and were exposed to the candidate applying their skills and knowledge in the workplace. It was suggested that the most appropriate observers were those people who had known the candidate for a period of time (at least three months), have the most contact with the candidate, and understand the nature of the candidate's workplace role. It was recommended that ratings be obtained from at least a supervisor, two peers, and a subordinate or client. The candidate was also required to complete a self-assessment as part of the assessment process.

For each item raters were required to select the behavioural descriptor that best described the skills and knowledge demonstrated by the candidate based on their interactions with the candidate over time and across workplace contexts. Raters 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 workplace performance. For the self-assessment instrument this option was replaced with "have not yet had the opportunity to apply these skills and knowledge".

Rasch Analysis

Preliminary data analysis was conducted using a multi-faceted Rasch model, an extension of the simple logistic Rasch model (Rasch, 1960 and revised 1980). The simple logistic Rasch model states that the probability of a person answering an item correctly, or receiving a particular rating, is dependent only upon the ability (or competence) of the person (θ_n) and the difficulty of item (δ_i). This model can be extended to include additional facets of the assessment context, typically rater severity, by the addition of an additional (severity) parameter (ρ_r). In this multi-faceted model, it is not only the competence of the candidate and the difficulty of the item that governs the probability of a particular rating, but also the severity of the rater (ρ_r) making the judgement.

The Partial Credit model (Masters, 1982; Wright & Masters, 1983) allows for scoring one or more intermediate levels on an item (as opposed to simply correct/incorrect), and to award partial credit for reaching one of these levels. The Partial Credit model can be applied to situations where ordered response alternatives vary in number and structure across items (Linacre, 1994; Masters, 1982). Thresholds ($\tau_1, \tau_2, ..., \tau_m$) are estimated for each response alternative (Wright & Masters, 1983).

The multi-faceted Rasch model used in the present study is shown in equation 1.

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

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 rater groups for each item *i* separately.

Results

Figure 1 displays graphically the calibration of the candidate, item and rater group facets. The first column provides the linear measurement scale (*logit scale*) 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 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 competence estimate for the present sample was 0.17 logits (*SD*= 0.36). 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 Figure 1, candidate competence estimates ranged from –0.89 logits (candidate 18, the least competent) to 1.06 logits (candidate 50, the most competent), a range of 1.95 logits. The standard errors for the candidate estimates were all acceptable (*M*=0.09, *SD*=0.02). Despite this relatively narrow spread, the overall difference between the competence level of the candidates was significant, $\chi^2(54)=689.8$, $p<.01$. The separation ratio (*G*) for the candidates was 3.71. This index compares the true spread of candidate competence measures with their measurement error, and indicates the spread of this sample of candidates (Fisher, 1992). It is possible to determine the number of statistically distinct performance levels or discernible strata that are present in the data using the formula $(4G+1)/3$ (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 0.93. 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 ratio 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).

The infit mean square statistics are 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 (Englehard, 1994), or more conservatively between 0.7 and 1.3 (Adams & Khoo, 1995). 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. With the exception of 5 candidates, all infit values were within the range of 0.7 to 1.3, and all values were between 0.6 and 1.5 (M=1.0, SD=0.2).

Items

As can be seen in Figure 1, item difficulties ranged from –0.64 logits (item 7, the easiest item) to 1.03 logits (item 19, the most difficult item), a range of 1.67 logits. Looking at the variable map it appears as if there are at least four distinct levels of difficulty, based solely on statistical grounds. The standard errors for all items were acceptable (*M*=0.07, *SD*=0.02). Despite this somewhat restricted range, normally a range of –3.0 to 3.0 logits

may reasonably be expected (Myford & Wolfe, 2000), the overall difference between the items was significant, $\chi^2(29)=1263.9$, $p<.01$, with a separation reliability of 0.97.

Examining the infit mean square values, all items displayed acceptable fit, with all values between 0.9 and 1.3 (M=1.0, SD=0.1). 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.

To examine if the response categories are appropriately ordered and clearly distinguishable, the average candidate competence measure for each response category was examined (Linacre, 1999c). This value 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 rating categories for the instrument) display a degree of misfit. Most of these appear to be the result of a lack of ratings 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 one rating category displayed an oufit value approaching 2 (category 3, item 3, outfit=1.8). Instrument revision by SMEs based on this item analysis will be undertaken.

The range of item difficulties and thresholds indicates the range of competence levels that an 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 (Wright & Stone, 1979). This can be investigated using the item separation ratio and also through an examination of the placement of items on the variable map (Wright & Masters, 1983). While the range of item difficulties is only 1.67 logits, response thresholds ranged from -1.72 to 2.3 logits, a range of 4.02 logits. As thresholds and not item difficulties will be used for variable interpretation, this spread is considered sufficient.

Rater Groups

As can be seen in Figure 1, with the exception of self-ratings, which display a degree of leniency (-0.28 logits), all rater groups tend to cluster around the same severity (0.05 to 0.12 logits). An examination of the fair average for each group also reveals that self-ratings are typically higher (1.88) than the other rater groups (1.47 to 1.55). 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 (M=0.02, SD=0.00). The overall difference between the rater groups was significant, $\chi^2(3)=163.3$, $p<.01$, 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 four 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 (0.07 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=.993, $p<.01$), 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 (Gregarus, Robie, & Born, 2001). This finding is not however unexpected as research suggests that self-ratings of performance likely reflect a global evaluation of performance (Bozeman, 1997).

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. An analysis of the use of this category by rater group reveals that this category was used quite regularly by supervisors (M=8.05), peers (M=8) and subordinates (M=7.39), but much less frequently by candidates when conducting their self-assessments (M=3).

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. 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

In summary, a number of tentative conclusions are supported from this preliminary analysis. 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 (1993) 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 (Griffin, 2000).

Alternatively, this process can undertaken qualitatively by having SMEs 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 obtained, a meaningful interpretation of the variable under consideration can be undertaken.

References

Table 1
Example Item 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

```
-------------------------------------------------------------------------------
|   Measure  |  Candidates   |  - Items   |  - Raters
-------------------------------------------------------------------------------
|            |               |            |
|            |  *            |            |
+   1        +               +  *         +
|            |               |  *         |
|            |               |            |
|            |  *            |  ***       |
|            |  ****         |            |
|            |  ****         |            |
|            |  *******      |  **        |
|            |  **********   |  *         |
|            |  ********     |            |
|            |  ****         |  *         |  Subordinate  Peer  Supervisor
*   0        *  ****         *  ****      *
|            |  *            |  ***       |
|            |  ****         |  ***       |
|            |  **           |  ****      |  Self
|            |  **           |            |
|            |  *            |  **        |
|            |               |  ***       |
|            |  *            |            |
|            |               |            |
|            |  *            |            |
+   -1       +               +            +
-------------------------------------------------------------------------------
|   Measure  |  * = 1        |  * = 1     |  - Raters
-------------------------------------------------------------------------------
```

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

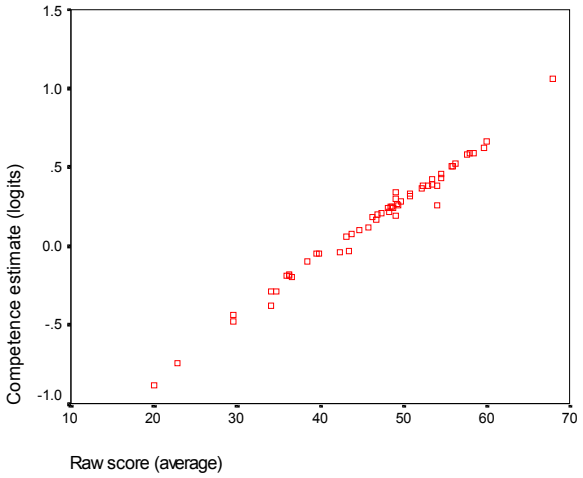

Figure 2: Plot of Candidate Raw Scores (average) to Competence Estimates (logits)