2011

Applying standards for leaders to the selection of secondary school principals

Helen Wildy
University of Western Australia

Coral Pepper
Edith Cowan University

Luo Guanzhong
Hong Kong Examinations and Assessment Authority

10.1108/09578231111129064 This article was originally published as: Wildy, H. R., Pepper, C. , & Guanzhong, L. (2011). Applying standards for leaders to the selection of secondary school principals. Journal of Educational Administration, 49(3), 276-291. Original article available here

Applying standards for leaders to the selection of secondary school principals

Abstract

Purpose

In this paper we describe innovative research to report on the process of selecting senior secondary principals by the public educational authority in Western Australia. Specifically we describe the development of performance-based tasks in the selection of senior secondary school principals over three years and describe the application of Rasch analysis to examine the construct validity and reliability of the tasks.

Methodology/Approach

Initially we describe previous research in the application of standards to selection in education, followed by a brief review of selection practices undertaken in two international, and the Western Australian educational settings. We then describe the innovative design of performance-based assessment tasks for the selection process including task and rubric development, rater training, and data validation. The Rasch measurement model is used to analyse the data sets gathered during three iterations of selection process for the Western Australian education authority.

Findings

The Rasch analysis of each data set provides evidence of construct validity and a robust measure of reliability. The Person and Item Location distributions indicate our tasks were better targeted for the highest performing candidates in the second and third iterations and that fine grained discrimination was evident across the candidate locations. The significance of the research lies in its applied nature and the potential offered by performance-based assessment tasks to make sound judgements about school leaders' ability to perform to a high standard in situations that are likely to confront them in large secondary schools.

Classification: Research paper

Keywords: performance-based assessment tasks, Rasch analysis

Applying standards for leaders to the selection of secondary school principals

Helen Wildy
School of Education, The University of Western Australia
Coral Pepper
Faculty of Regional Professional Studies, Edith Cowan University
Luo Ghanzhong
School of Education, The University of Western Australia

The first author of this paper is the chief investigator of the most recent large research project as well as the co-investigator of each of the earlier projects. The second named author was the research associate for the past three years of the project. The third named author of this paper provided the data analysis and advice throughout the decade of the research. We also acknowledge the input of: Professor Bill Louden who instigated the research and secured the first large grant; Dr Simon Clarke who worked on the project during its formative period and; Professor David Andrich who provided measurement expertise in the latter stages of the research.

Applying standards for leaders to the selection of secondary school principals

In this paper we describe innovative research to report on the process of selecting senior secondary principals by the public educational authority in Western Australia. The process derives from a decade of collaborative research funded by two large commonwealth grants as well as two small local grants. We investigated the extent to which standards for school leaders can be applied in the process of selecting principals for promotion to senior secondary schools using performance-based assessment tasks. Specifically, this paper describes the development of performance-based tasks in the selection of senior secondary school principals over three years. The paper also focuses on the application of Rasch analysis to examine the construct validity and reliability for each of the tasks.

There are three sections to the paper. The first provides an overview of background research conducted in the application of standards to selection in the field of education, reviewing briefly the practices adopted in two international educational settings. Included here is an outline of practices for selecting principals adopted in Australia. The second section focuses on the Western Australian context, giving a brief overview of the development of standards in this research project and their current application, particularly the development of performance-based tasks and their use in selection processes over three years (2004, 2005 and 2006). The third section introduces the Rasch measurement model and its use to analyse the research data from three rounds of selection.

Background

We know that 'standards' has been a dominant metaphor in educational reform for some decades. As well as standards for student performance, beginning teachers and experienced teachers in most jurisdictions internationally, there are also standards for school leaders' work. Perhaps most well known are the National Standards for Headteachers (NSHT, 2006) in England and Wales and the Interstate School Leaders Licensure Consortium (ISLLC) standards for school leaders developed by the Council of Chief State School Officers United States (1996). Characteristically, the developers of such standards acknowledge the complexity of the performance described by the standards. For some time now, however, we have been critical of both the nature of standards and the uses to which they are frequently put.

We have argued elsewhere that standards are frequently weakened by fragmentation into long list of duties that lend themselves to checklists, and by lack of attention to the essential qualities that characterise accomplished performances (Louden & Wildy, 1999a). Further, the standards are not always accompanied by psychometrically adequate assessments. In the United States, agencies responsible for professional standards have attended to the development of assessment against standards. For example, the National Board for Professional Teaching Standards (NBPTS, 1989) use sophisticated processes for assessing standards for accomplished teachers. Similarly, the Council of Chief State School Officers (CCSSO, 1996) has developed procedures for setting and assessing professional standards for school principals. The ISLLC assessment method is a pencil and paper test requiring problem-solving in response to scenarios and to data-based decision-making tasks. Such methods have strong scoring

rubrics based on established standards and have been subject to appropriate psychometric scrutiny (CCSSO, 1997).

Against this background of robust assessment practices in the United States, principals are prepared for leadership positions by participating in any of the numerous principal preparation programs on offer in the United States. Several of these are delivered through professional principal associations such as the National Association of Elementary School Principals (NAESP, 2006) and the National Association of Secondary School Principals (NASSP, 2002). Others such as the National Center on Education and the Economy deliver programs in partnership with state universities (NCEE, 2005). Principal selection may then be a combination of merit selection and negotiation between educational authorities and parent organisations. The NASSP has a twenty-year history of working with the assessment center process to strengthen principal preparation and selection. The process increases the 'best fit' factors in the placement and selection decisions for school leadership positions through performance-based activities (NASSP, 2002).

In the United Kingdom, it has been mandatory since 2004 for all first time Headteacher applicants to secure a place on the National Professional Qualification for Headship program (NPQH) prior to their first permanent substantive headship in the maintained sector. The NPQH is considered a benchmark qualification, underpinned by the National Standards for Headteachers (NSHT), and takes up to 15 months to complete (NAHT, 2006). However, a study undertaken into headteacher recruitment on behalf of the National College for School Leadership (NCSL) by a consortium of the Hay Group, Cambridge University, Eastern Leadership Centre and National Association of Head Teachers (NAHT) is critical of aspects of headteacher selection. For example, selection involves many stages; long-listing, which 50 percent of school governors consider unfair; interviews, which take between 1.5 and 2 days to conduct and require careful management to reduce bias; short listing activities, which may include panel interview, candidate presentation, psychometric testing or parent/community panels; and the use of external assessment centres (NAHT, 2006).

In both the US and the UK then it is customary for aspiring principals to undertake preparation programs and undergo complex selection processes both of which are grounded in standards frameworks. As late as the 1980s in all Australian public education systems, principals were selected for positions through a centrally administered bureaucratic and hierarchical system based on seniority (Blackmore and Barty, 2004). Since that time, selection based on merit has been introduced. The shift to merit selection corresponded with the decentralisation of authority to schools, regions or districts reflecting the general reform of public sector administration to emphasise devolution of responsibilities to local units, a reform undertaken by all jurisdictions across the country. However, what constitutes 'merit' continues to be a matter of concern because those who judge the candidates are frequently the same set of people as those who articulate standards. Selection processes rely on resumés, written applications against selection criteria, referee reports and panel interviews. On the basis of their investigation of the declining supply of principals in Australia, Blackmore and her colleague argue that the selection processes adopted in Australia have adversely affected the image and understanding of educational leadership in this country. Those involved in making judgements in such merit selection systems generally regard them to be 'superficial and prone to error' (p. 6) and to reward 'past

reputation and investment in the job' (p. 7). These authors characterise current selection practices as a 'reproduction' model, serving to normalize principal identity and exclude those who did not fit the model. In addition to these cultural issues, there is evidence that panel interviews, as means of making judgements about candidates, generate considerably less than robust data. For example, 25 years ago Hunter and Hunter (1984) found that, despite training, the agreement among panel members was found to be typically close to 0.2, that is, not much different from random.

In Australia aspiring principals learn to be principals 'on the job' without any mandated formal preparation programs. Furthermore, standards for both leaders and teachers typically have been characterised by weak assessments or no assessments at all (Louden, 2000). One of a few exceptions is the Level 3 Competency Standards developed for the Education Department of Western Australia (Western Australia, 1997). By adopting assessment procedures more closely related to performance than the traditional public service process of resumes, interviews and referees' reports, this process broke new ground in the Australian context. Despite the innovative characteristics of this process, reviews of assessment practices against standards in Australia show that the process fails to distinguish between different *levels of performance* (Chadbourne and Ingvarson, 1998; Wallace, Wildy and Louden, 1999; Jasman and Barrera, 1998). Our research over the past decade has attempted to address this omission, at least in relation to standards for school leaders in the Western Australian public education sector (Louden and Wildy, 1999b; Wildy, 2004; Wildy and Louden, 2000; Wildy and Pepper, 2005).

The Western Australian context

The research reported in this paper aimed to provide a more robust selection process using assessment tasks based explicitly on standards for leaders with the view to attract 'new blood' into leadership positions in senior public secondary schools in Western Australia. Our dissatisfaction with long lists as substitutes for standards led to the application of narratives as an alternative approach to the generation and use of standards for school leaders. Our research was conducted throughout a decade in collaboration with the public education authority, professional associations and many hundreds of school principals (Louden and Wildy, 1999a, 1999b; Wildy, 2004; Wildy and Louden, 2000; Wildy and Pepper, 2005). The research established that the quality of performance of school leaders is linked to personal attributes that shape the way leaders act, rather than to actions or duties. These personal attributes (*fair, decisive, supportive, collaborative, flexible, tactful, innovative and persistent*) are interrelated in complex ways often between conflicting demands such as decisiveness and collaboration, or tactfulness and persistence. Accomplished performance is characterised, not by displaying more of an attribute, but by balancing competing demands in particular contexts. Narrative accounts of leader performance can be rated and arrayed on continua showing variation in performance of the attributes. Previous publications describe examples of narratives, how they are developed, their ratings, and how they are used to illustrate this variation as levels - that is, standards - of performance.

This approach to standards for school leaders has been adopted and endorsed by the Western Australian public education system. Teachers and leaders access information about the standards framework through a Leadership Centre which offers professional

development programs linked explicitly to the framework (http://www.det.wa.edu.au/education/lc/standards.html). Central to the most recent large commonwealth grant has been the development of performance-based assessments for the selection of principals. In this paper we report on three applications although various project members conducted 13 different processes during earlier stages of the research. (We acknowledge the earlier input of Professor Bill Louden who instigated the research and secured the first large grant, Dr Simon Clarke who worked on the project during its formative period and Professor David Andrich who provided measurement expertise in the latter stages of the research).

Selection process design

Responsibility for the selection process is shared between the education authority and the researchers. The education authority sets up a Selection Panel of four, chaired by a senior Director with three experienced senior secondary principals. The Selection Panel is responsible for setting the timeframe for the process, inviting candidates to apply, conducting all correspondence with candidates, reading their written applications against selection criteria, reading referee reports and short-listing the candidates. Until our involvement through the current research program, the third part of the selection process was a Panel interview with each short-listed candidate, a process lasting many days. However, the Panel is responsible for recommending whether candidates are successful, based on data from written application, referee reports and the performance-based tasks.

Candidates apply to join a 'pool' of eligible principals who subsequently are offered opportunities to apply for positions in schools, as these positions become available through retirement or transfer of incumbent principals. The pool usually lasts for 12 months by which time most principals have been appointed and the pool is 'empty.' However, some principals decide to wait until a particular school position becomes vacant and will elect to remain in the pool. To remain in the pool requires reapplying for selection and being successful. A small number of candidates choose to do so. This is one way by which candidates repeat the selection process. Another reason for experiencing the selection process more than once is for candidates who were unsuccessful in being selected for the pool in one year to apply in subsequent years. A large number of candidates choose to do so. For these reasons we assume that the differences among candidates are similar from year to year. Furthermore, the candidates are likely to be similar in quality from year to year because they are drawn from a small set (80) of government secondary schools in a relatively homogenous and isolated educational jurisdiction. However, we might also assume that there is a practice effect for those candidates who apply more than once and increase in their knowledge of what is required by the process and in their confidence to perform in the process.

The Panel delegates responsibility to the researchers to design and develop a set of performance-based assessment tasks that will differentiate between candidates who are, and are not, suitable for appointment to the position of principal of a large secondary school. We also have responsibility for training the raters, organizing the selection day process, overseeing quality and subsequently validating data and analysing the data. Each of these is outlined in the following sections.

Task development

The aim of the project was to ensure fair decision-making. Tasks were designed to ensure both validity of the process and reliability of raters' judgements of candidates. Task development was a collaborative process involving the first and second authors and a reference group of senior and experienced secondary principals in half-day meetings over three months to ensure face validity of the Tasks. Tasks were developed inductively starting with a brainstorming process to identify challenges facing the secondary principals of the future. Through an iterative process of synthesis and elaboration of ideas, three key issues were identified and materials were developed to provide a context for investigating the issue. Examples of issues are: dealing with a poor performing department Head; handling a critical incident; implementing school-wide curriculum change. Tasks were designed to include a variety of modes of communication such as reporting to a superior, dealing with a subordinate, and addressing a large group. Table I below summarises the Tasks developed for each year.

Table I. Summary of Tasks 2004 - 2006

Year	Task A	Task B	Task C
2004	Address department Head on topic of poor performing department	Address whole staff on topic of vision and strategies for improving student performance	Address District Director on topic of handling a critical incident
2005	Address whole staff on topic of dealing with major curriculum reform (policy shift)	Address department Head on topic of vision and strategies for improving student performance	Address District Director on topic of handling a critical incident
2006	Address District Director on topic of dealing with increase in student leaving age (policy shift)	Address whole staff on topic of vision and strategies for improving student performance	Address department Head on topic of poor performing department

For each Task, descriptions of what constitutes high performance and what constitutes low performance were prepared, based on input from the reference group.
Tasks were the media through which candidates would demonstrate their knowledge, understanding and skill in relation to the Leadership Framework in general and the role of principal in particular. Specifically, the Tasks would provide opportunities for candidates not only to show they knew <u>what</u> was required of a principal of a large secondary school but also that they knew <u>how</u> to do the job. In relation to the Leadership Framework, candidates would need to show they were *fair, tactful, innovative, decisive, collaborative, persistent* and *flexible* in the right balance and in the right amount for the given context. Therefore Task development involved assembling contextual information about a large secondary school and also preparing models of high quality responses to each Task as well as low quality responses to each Task.

A distinctive feature of the WA Leadership framework is the articulation of the optimum amount of each attribute – neither 'too much' nor 'too little' – for a given set

of actions (competencies, capabilities, duties), in a given context (small school, rural school and so on). Table II provides an example of this articulation, in relation to one of the Tasks.

Table II. Attributes applied to Task B: Too little, Too much and Just right

Attribute	Too little	Just right	Too much
Decisive	Leaves decisions to others, shows no vision or views, dithers	Shares interpretation of research data, tackles issues, considers others' perception, asks hard questions, wants answers, proposes strategy	Dictates changes to be made, dictates who will do what and when, confrontational
Innovative	Puts forward a single solution or no suggestion for strategies, 'what is tried and true must be best'	Suggests a range of appropriate and realistic strategies with clear understanding for practice, builds on existing effective structures and processes	Puts forward many ideas not necessarily practical or relevant; fails to build on existing good practices

High performance is evident when candidates demonstrate 'just right' amounts of each attribute, in balance and appropriate to the context (the Task). Low performance is evident when candidates demonstrate either 'too little' or 'too much' of the attribute. Raters are trained to identify high performance and low performance on their respective Tasks.

Rubric development

Scoring rubrics were developed for each Task. Each rubric consists of five dimensions, with scales for each dimension from 1 (low) to high (5). The dimensions are particular to the Task, and collectively incorporate all eight attributes. To ensure consistency over time, the same dimensions are used each year. An excerpt from the scoring rubric for Task B appears in Table III below.

Table III. Excerpt from scoring rubric

Dimension	Low	1	2	3	4	5	High
Decisiveness	Imbalance in decisiveness and collaboration Delegates inappropriately Fails to identify tasks and responsibilities Shirks accountability Fails to ask 'hard' questions						Balances decisiveness and collaboration Delegates appropriately Identifies tasks and responsibilities Demonstrates accountability Asks 'hard' questions

Raters are trained to use the rubric in conjunction with the descriptions of high performance and low performance as well as the statements of performance on attributes.

Rater training

Each iteration of the Selection process involved 12 raters. Over time the number of experienced raters in the education system has increased. However, since the process is viewed as an educative one for raters as well as candidates, the practice has been to

invite novice raters to participate in the Selection process as well as those who have experience as raters. The raters in 2004 are coded Rater 1 to Rater 12. In 2005, half the raters were novice raters and they are coded Rater 13 to Rater 18; half the raters had been raters in 2004. In 2006, half the raters were novice raters and they are coded Rater 19 to Rater 24; half the raters had rated in either 2004 or 2006 or both. In the three years of this study, a total of 24 different raters were involved. Of these, two raters were involved in all three years, and eight raters were involved in two of the three years.

Raters participate in a half-day training program. There are two parts to the program: familiarization with rater teams, Tasks and scoring rubrics; and bias training.

Familiarisation with rater teams, Tasks and scoring rubrics

After the arrangements for Selection process are explained to raters, raters are allocated to teams of four and each team is allocated a Task. The teams are set up to ensure balance: male and female raters; experienced and novice raters; familiarity with the process (either as candidates or raters); status in the education system (principals, District Directors, Executive Directors) as well as ensuring that a representative of the formal Selection Panel belongs to each team.

Raters are familiarised with their Tasks by preparing for and making presentations within the same constraints imposed on the candidates. After one hour and 15 minutes of preparation, one rater from each team delivers the required presentation to the three raters in the team. The scoring rubric is given to the three raters to apply to the performance of their colleague. This process continues until all raters have made a presentation and each rater has applied the rubric three times. Interspersed between presentations, raters discuss issues that arise. First, the protocols for meeting and greeting candidates are agreed so there is no engagement with candidates or between raters about presentations. Raters are next provided with two sets of information to help them make informed and consistent judgements about standards of performance. One set of information is a description of both a high quality performance and a low quality performance of the particular Task. The other set of information is a description of the application to the particular Task of the attributes from the Leadership Framework. Raters are encouraged to examine their rating practices in light of both sets of information. Table II (above) contains the application of two attributes, *decisive* and *innovative*, to Task B in 2006, illustrating responses that show too little and too much of each attribute as well as the right amount of the attribute, for this Task, in the context to which the Task applies.

When raters are familiar with their role as raters, with the requirements of their Tasks, and when they have a thorough understanding of the standards implied by the scoring rubrics, they engage in a process designed to raise their awareness of the personal beliefs and value that they each bring to bear on the judgement of candidates' performance. The process is referred to as bias training.

Bias training

Bias training helps raters to focus their judgements on the performance, rather than on the person. Bias training is relevant to raters because candidates are known to raters as teachers, deputy principals or principals in a relatively small and isolated educational jurisdiction. The training aims to heighten raters' awareness of factors likely to

mediate their use of scoring rubrics to make judgements about candidates' performance on Tasks. In a half-day interactive facilitated session, raters discuss stereotypes about age, race, class and gender as well as the role of local knowledge of candidates in making judgements, focusing on what is, and what is not, included in scoring rubrics.

Selection process

The Selection process is conducted on one day and designed and monitored to ensure fairness to all candidates. Initially, following a formal welcome from the Chair of the Panel, the researchers explain to candidates the requirements of the Tasks and the schedule for the day. Candidates are allocated in random order to sets of six and, in 15-minute intervals, six candidates are taken to a preparation room where they spend 15 minutes reading a set of source documents relevant to all three Tasks. After 15 minutes, these candidates are given Task documents relevant to a Task: two candidates then prepare for Task A, two candidates prepare for Task B and two candidates prepare for Task C. One hour is available for Task preparation and candidates do this in a supervised setting. Ushers are employed to distribute materials to candidates and also to take the candidates at the appointed time to the appointed room for their presentations.

Candidates make an eight-minute presentation, as required by the Task, to two raters. The raters do not engage in conversation with candidates except to greet and farewell them. Raters then have seven minutes during which to complete the scoring rubric and enter a score for each of the five dimensions relevant to that Task. Raters rate independently and do not engage in discussion about the candidates or their performance. Throughout the day's process, raters change pairs so that by the end of the day each has rated with each other rater in the team of raters of one of the Tasks.

Candidates are required to leave with the raters any notes on which they based their presentation so that raters might subsequently be reminded of individual presentations. Candidates also hand their Task materials to the raters after their presentations. This procedure continues until each candidate has made three presentations, one on each Task. When candidates are not preparing for the next Task they are required to remain in the designated area of the building so that they might be given their next set of Task materials at the appointed time. When candidates complete the last of the three Tasks, they hand in the source documents and sign that all materials are returned. Raters hand their completed scoring rubrics to the researchers who enter the data into an Excel spread sheet ready first for validation and then for analysis.

Data validation

The half-day data validation session is an accountability process involving all raters in first, checking the data entry; second, reconciling discrepancies in ratings on each candidate between pairs of raters; and third, justifying judgements about performance on each of the three Tasks. After raters have signed off the accuracy of data entry on each of the Tasks, they are required to reduce, to one point or less, any difference of more than one point of more than one point between pairs of ratings on any dimension. Raters are also required to investigate reversals in ratings, that is, instances

of inconsistencies in patterns of harshness between pairs of raters. Such reversals indicates bias and raters are required to address the inconsistencies by negotiating in rating teams with reference to the scoring rubric, the information about standards of performance and also the notes candidates handed to raters at the conclusion of their presentation. Changes to the database are documented throughout these first two stages of validation.

The next stage of the validation process has two functions. The primary and explicit function is to ensure raters are accountable for the judgements they make about candidates' performances. The secondary function is implicit: when raters describe candidates' performances starting with the highest ranked and ending with the lowest ranked performance, they articulate standards of performance for this cohort of principals. This is an educative function, allowing experienced raters to induct novice raters into the culture of the rating process and also facilitating the articulation of quality for senior secondary principals in this educational jurisdiction.

To prepare for the third stage of the validation process, scores on five dimensions from two raters for each Task are calculated and candidates are ranked on each Task. The total Candidates' scores are then rank ordered by total scores for three Tasks, as well as rank ordered by each candidate's performance on each Task. Starting with the highest ranked candidate, raters of each candidate on each Task in turn describe the key features of the performance in relation to the scoring rubric. The descriptions of the highest ranked performances explicate what these raters identify as high quality. As the candidates further down the rank order are described, raters explicate moderate quality performance on each Task. The descriptions of the lowest ranked candidates make explicit the lowest quality performances. However, there are aberrant performances: a candidate might perform well on two Tasks and poorly on a third Task and in such instances possible explanations are canvassed. Perhaps this was the first Task undertaken and the candidate was anxious or misjudged the time. Perhaps the candidate misinterpreted the Task. Although scores are not changed, explanations for inconsistent performances are noted for inclusion in feedback to candidates. Similarly, if a candidate performed badly on an element of one Task but demonstrated on another Task that this element was well handled then a low overall score is considered sympathetically in terms of decision making at the margin.

Data analysis and findings

The Rasch measurement model was used to analyse the data in terms of construct validity and reliability for each of the Tasks as well as overall performance. Rasch analysis provides evidence of the construct validity of the measures, that is, the extent to which the three Tasks can be regarded as measuring a single construct. In this research, the construct is leadership for senior secondary principals. Using RUMM software (Andrich, Sheridan and Luo, 2000) the data sets for each year were analysed.

The Rasch analysis of each data set gives the measure of reliability (Person Separation Index). The Rasch analysis generates person (candidate) locations and item (score given by a rater to a dimension of a Task) locations. The candidate locations and item locations are shown in Figures 1A, 1B and 1C which indicate the relative spread of both candidates and items and allow for discussion of the extent to which the items are targeted to the candidates. Rasch analysis also provides rater profiles and a

discussion on this topic is the subject of another paper. A display generated of candidates' performance over time is shown in Figure II and this is followed by discussion of the clustering of candidates and the relative level of performance of each group of candidates over time.

Findings

In this section we present discussion of three topics: candidate performance; reliability; and targeting. Rater stability, rater profiles and rater harshness and item difficulty are addressed in a separate discussion paper.

Candidate performance

Each candidate has 30 scores, given by each of two raters on each of five dimensions, on a rating scale from 1 (low) to 5 (high), for each of three Tasks. Because of the large number of candidates to be assessed in one day, there are two rating pairs for each Task so that each rater assesses half the candidates. The data are arrayed as 60 items, with entries in 30 of the cells corresponding to the ratings on each dimension of each Task given by the pair of raters. The remaining 30 cells, corresponding to ratings on each dimension of each Task given by the second pair of raters, are treated as having missing data. The capacity to deal with missing data is a feature of the Rasch analysis that is used to generate Person and Item Locations relative to each other.

The total score for each candidate is the sum of ratings by two raters of each candidate's performance on five dimensions of three Tasks. Candidates can score a maximum score of 150 and a minimum score of 30. The total scores are then converted to relative locations on a common scale (logits) using the software program RUMM (Andrich, Sheridan and Luo, 2000).

Reliability

In the Rasch analysis, the measure of reliability of the data is provided by the Person Separation Index (analogous to Cronbach's alpha). The reliability measure for each year is shown in Table IV below.

Table IV. Person Separation Index for 2004 - 2006

Year	Person Separation Index
2004	0.95
2006	0.96
2006	0.98

Such high reliability encourages our confidence that the process is robust and has implications for policy and practice within the sector. As indicated earlier in this paper, there is evidence that despite training, panel interviews generate considerably less robust data of around 0.2 agreement among panel members (Hunter and Hunter, 1984).

Targeting

The Person and Item Location distributions indicate how well the items are targeted to the performance of the candidates.

Figure 1A. Person-Item Location Distribution 2004

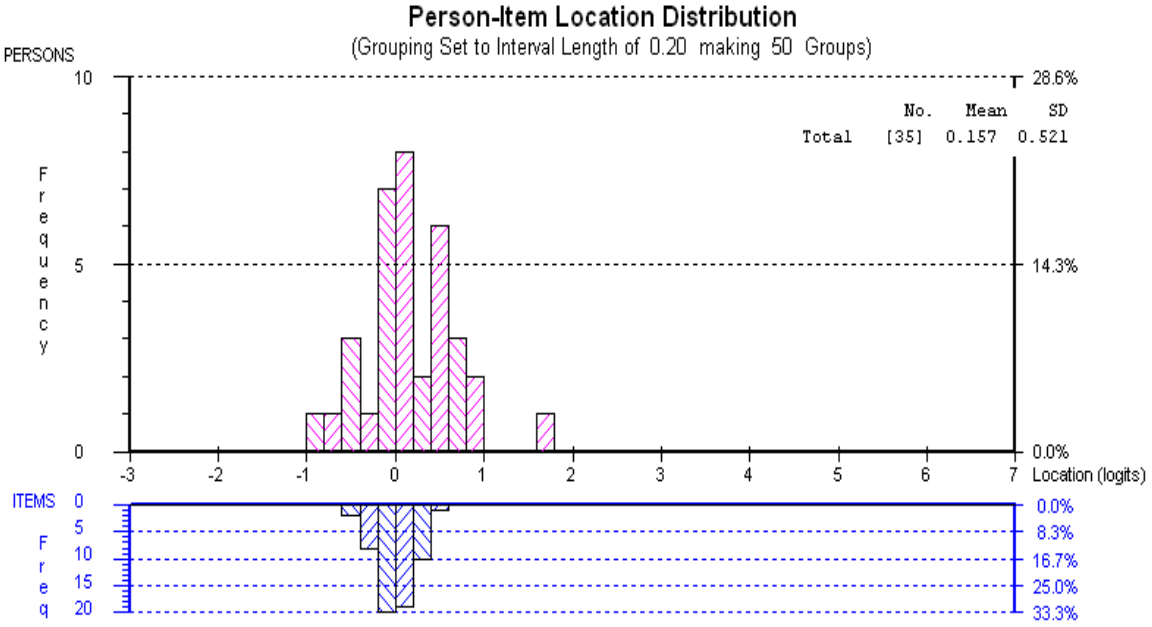
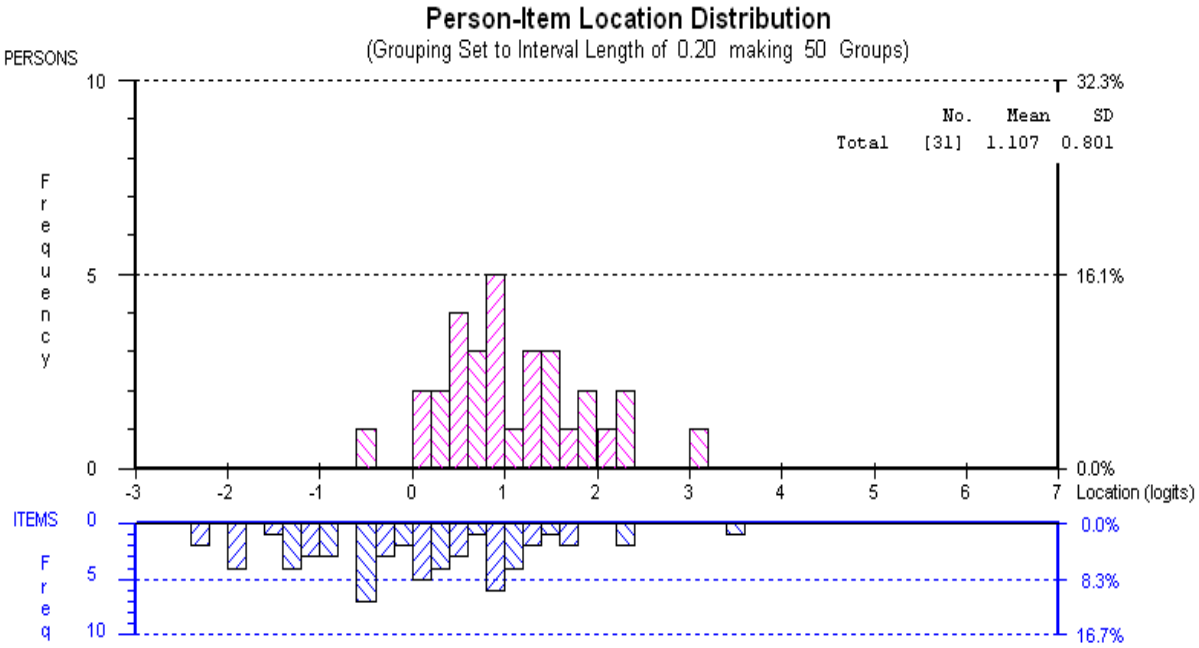


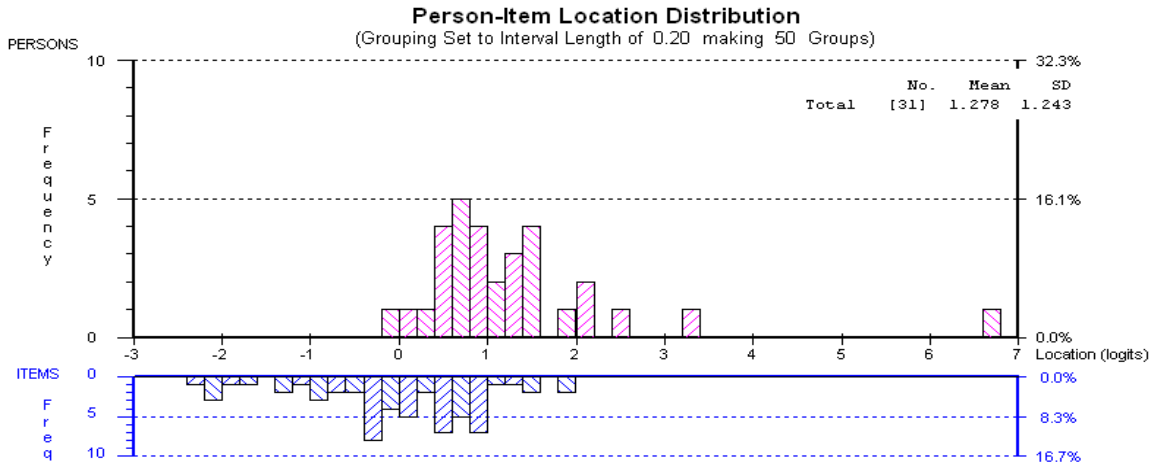
Figure 1A shows the distribution of the 2004 candidates in relation to the distribution of the items on the horizontal location scale in logits (the Rasch measurement units). The vertical axes show frequencies of candidates, and of items. The 35 candidates have a mean location of 0.157 logits and a standard deviation of 0.521 logits. However, the items are tightly bunched within the spread of the candidates, indicating that the items do not target the entire spread of candidates. Neither lower nor higher scoring candidates are targeted as well as they could be by the items. Now we examine the targeting of items to candidates in 2005 in Figure 1B.

Figure 1B. Person-Item Location Distribution 2005



The 31 candidates have a mean location of 1.107 logits and a standard deviation of 0.801 logits. In contrast to the targeting of items to candidates in 2004 where the items were tightly bunched within the spread of candidates, the items target the range of candidates better in 2005. Indeed the range of items exceeds the range of the candidates, especially at the lower end, with relatively few items being targeted to the higher scoring candidates. Here, there are some items that are very easy for most of the candidates. However, in practice, this is appropriate because the items are designed to identify candidates at the lower end who are not suitable for appointment. Fine-grained discrimination at the high end is not necessary because candidates at the top end will be considered suitable for appointment. Although the items might be too easy for some candidates, the items themselves are well spread in relation to most of the candidates. The single item location on the right (high) end of the item scale is likely due to rater harshness/leniency. In Figure 1C, we examine the targeting of items to candidates in 2006.

Figure 1C. Person-Item Location Distribution 2006



The 31 candidates have a mean location of 1.278 logits and a standard deviation of 1.243 logits. In contrast to the targeting of items to candidates in 2004 but similar to the targeting of items to candidates in 2005, the items target most of the candidates in 2006. However, unlike the targeting of 2005, some of the candidates are located above all the items and, like 2005, some of the items are located below all the candidates. Here some of the candidates perform at levels beyond the most difficult of the items. As we noted in relation to the 2005 data, this is appropriate because the items are designed to identify candidates at the lower end who are not suitable for appointment. Fine-grained discrimination at the high end is not necessary because candidates at the top end will be considered suitable for appointment. Although some of the items are too easy for this set of candidates, the items themselves are well spread in relation to most of the candidates. The most unusual feature of the 2006 candidate locations is the extremely high score of one candidate, three standard deviations above the next closest candidate and nearly four standards deviations above the mean. This candidate is indeed an outlier in this set of candidates.

In summary, the Figures presented show poor targeting for 2004, good targeting for most of the candidates in 2005, and slightly less focused targeting for the highest performing candidates in 2006. However, comparison of means shows that the candidates are on average located higher on successive years: the mean location of candidates in 2004 was 0.157 logits; the mean location of candidates in 2005 was 1.107 logits and in 2006 the mean candidate location was 1.278 logits. It is possible that the increasingly high location of candidates indicates that candidates are performing better in successive years. Alternatively, it could be argued that the Tasks are becoming easier, that what is required by the Tasks is better understood by candidates, or that there is a practice effect among candidates. Comparison of the spread of candidates shows increased spread on successive years: the standard deviation in 2004 was 0.521 logits; the standard deviation in 2005 was 0.801 logits and the standard deviation in 2006 was 1.243 logits. It is possible that the increasingly wide spread of candidates indicates that raters are performing better in successive years by being increasingly skilful in distinguishing the performance of candidates. Alternatively, it could be argued that the raters are becoming less consistent over time.

The relative position and spread of the candidates are shown in Figure 2 below.

Figure 2. Comparisons of Locations of Candidates 2004 - 2006

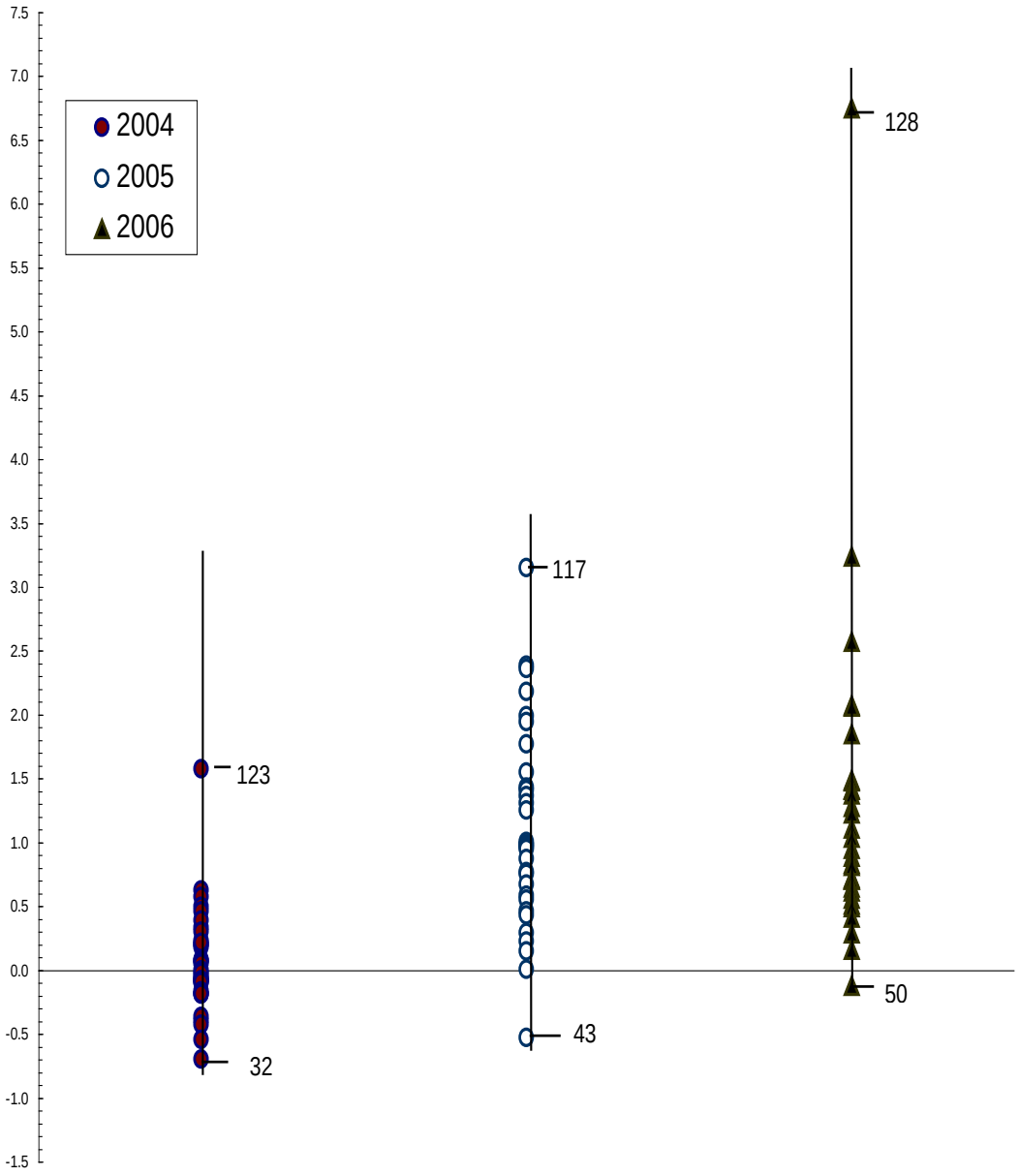


Figure 2 shows the locations of candidates for 2004, 2005 and 2006. The vertical axis is the locations scale (logits) showing the origin of 0 and a maximum of 7.5 logits and a minimum of -1.5 logits. The points on each of the three scales represent candidates for each year (2004 n = 35; 2005 n = 31; 2006 n = 31). In addition to showing the locations of candidates, each scale shows the maximum and minimum total scores obtained by candidates. The non-linear relationship between total scores and locations (logits) at the extreme ends of the scale is evidenced by the relationship between the maximum total scores (123, 117 and 128) and the locations for these candidates on each scale.

The important feature of the representation of data shown in Figure 2 is the differences in the spread of candidates on the three scales, both in terms of absolute range of locations and in terms of clustering of locations. First, each year shows one extreme location of a candidate: the highest location in each set is dramatically higher than all other locations on that scale. Indeed, the candidate who scored a total score of 123 in 2004 is the same candidate who scored a total score of 128 in 2006. The candidate who scored a total score of 117 in 2005 is the same candidate who scored the second highest total score in 2006. These two high scoring candidates demonstrated performances that are higher in quality than all other candidates and provide evidence of the validity of the measures. The reason for candidates applying for selection on more than one occasion was discussed in the earlier section on the Selection Practices.

The second aspect of the difference of spread of the three scales is the way the locations are clustered on each scale. For the 2004 data, the locations of candidates are tightly clustered, showing little spread, as was shown in Figure 1A in the earlier discussion of targeting. For the 2005 and 2006 data, the locations are more widely spread than for the 2004 data, as was shown in Figures 1B and 1C in the earlier discussion of targeting. The amount of spread indicates the capacity of the raters to discriminate between candidates' performances on the Tasks. The 2004 data indicates the raters were cautious in using the scoring rubrics to distinguish the quality of the candidates' performance on the Tasks and tended to rate all the candidates more or less the same as each other. Indeed, the raters in 2004 demonstrate that they may not have been skilled enough as raters to make distinctions between candidates. The greater spread in the data of 2005 and 2006 may indicate the raters were more confident in their use of the scoring rubric in 2005 and 2006 than were the raters in 2004. The raters in 2005 and 2006 may have demonstrated that they understood the descriptions of variations in performance and were skilled enough to make fine-grained distinctions between candidates' performance on the Tasks. In other words, the 2005 and 2006 data may indicate better discrimination of candidates' performance on the Tasks by raters. Alternately, the candidates in 2005 and 2006 may simply have been more different from each other.

Conclusion

The paper outlines the development and application of selection tasks that are grounded in practice, linked to the standards framework, and scored by trained raters using standards-based rubrics. We argue that the performance-based assessment tasks we have developed have high face validity because they are recognized by candidates to be every-day work for principals. Our performance-based assessment tasks also have high content validity because they are developed collaboratively with senior experienced principals and they represent what is considered important in the work of principals. Our analysis of the data gathered through the application of the Rasch measurement model gives us confidence that this selection process is robust. The performance-based tasks as described in this paper provide a sound basis on which to make judgements about the school leaders' ability to perform to a high standard in situations they are likely to confront in large secondary schools.

We believe that this education authority would be wise to continue this selection process, using the data that is generated to give feedback to both candidates and raters. Feedback based on performance on such tasks is focused, context specific, recent and likely to help candidates in monitoring and improving their performance. Most importantly, the feedback is linked to the standards framework from which

candidates can continue to monitor their own development, fostering professional responsibility rather than training candidates in dependency (Wallace and Wildy, 1995).

References

Andrich, D., Sheridan, B. and Luo, G. (2000), RUMM2020: A Windows interactive program for analysing data with Rasch unidimensional models for measurement, Perth, RUMM Laboratory: Western Australia.

Blackmore, J. and Barty, K. (2004), *Principal selection: Homosociability, the search for security and the production of normalized principal identities,* Paper presented to the Annual Conference of the Australian Association of Research in Education, Melbourne, Victoria.

Chadbourne, R. and Ingvarson, L. (1998), 'Self-managing schools and professional recognition: The Professional Recognition Program in Victoria's Schools of the Future', *Australian Educational Researcher,* Vol. 25, pp. 61-93.

Council of Chief State School Officers (1996), *Interstate School Leaders Licensure Consortium. Standards for school leaders,* Council of Chief State School Officers: Washington DC.

Council of Chief State School Officers (1997), 'Preliminary results from the content validation panel', *Work in Progress,* Vol. 2 No. 2.

Hunter J. E. and Hunter, R.E. (1984), 'The validity and utility of alternative predictors of job performance', *Psychological Bulletin*, Vol. 96, pp. 72-98.

Jasman, A. and Barrera, S. (1998), *Teacher career structure: Level 3 Classroom Teacher. Final report,* Perth: Murdoch University and Nexus Strategic Solutions.

Louden, W. (2000), 'Standards for standards: The development of Australian professional standards for teaching', *Australian Journal of Education,* Vol. 44 No. 2, pp. 118-134.

Louden, W. and Wildy, H. (1999a), ''Circumstance and proper timing': Context and the construction of a standards framework for school principals' performance', *Educational Administration Quarterly,* Vol. 35 No. 3, pp. 399-422.

Louden, W. and Wildy, H. (1999b), 'Short shrift to long lists: An alternative approach to the development of performance standards for school principals', *Journal of Educational Administration*, Vol. 37 No. 2, pp. 99-120.

National Association of Elementary School Principals (2006), Website http://www.naesp.org accessed 113 January, 2010.

National Association of Head Teachers (2006), Website http://www.naht.org.uk accessed 13 January, 2010.

National Association of Secondary School Principals (2002), Website http://www.principals.org/ accessed 13 January, 2010.

National Board for Professional Teaching Standards (1989), Website http://www.nbpts.org/ accessed 13 January, 2010.

National Center on Education and the Economy (2005), Website http://www.ncee.org/ncee accessed 17 August, 2006.

Wallace, J., Wildy, H. and Louden, W. (1999), 'Career progression for competent teachers: A case of portfolio selection', *Unicorn,* Vol. 25, No. 2, pp. 24-36.

Wallace, J. and Wildy, H. (1995), 'The changing world of school leadership: Working in a professional organisation today', *The Practising Administrator*, Vol. 17 No. 1, pp. 14-17.

Western Australia, Education Department (1997), 'Applying for Level 3 Classroom Teacher', *School Matters,* 18 June, pp. 13-14.
Wildy, H. and Louden, W. (2000), 'School restructuring and the dilemmas of principals' work', *Educational Management and Administration*, Vol. 28 No. 3, pp. 173-184.
Wildy, H. (2004), 'Using performance standards in the selection of district directors', *Measurement: Interdisciplinary Research and Perspectives,* Vol. 2 No. 2*,* pp. 119-124.
Wildy, H. and Pepper, C. (2005), 'Using narratives to develop standards for leaders: Applying an innovative approach in Western Australia', *Educational Research & Perspectives*, Vol. 32 No. 2, pp. 122-141.