Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments

Sandra Heldsinger
Stephen Humphry
The University of Western Australia

Abstract

Demands for accountability have seen the implementation of large scale testing programs in Australia and internationally. There is, however, a growing body of evidence to show that externally imposed testing programs do not have a sustained impact on student achievement. It has been argued that teacher assessment is more effective in raising student achievement levels. However, it is also often argued that teacher assessments are less reliable than the results of testing programs. This paper presents a study in which teachers judged writing scripts using the process of pairwise comparison to generate a scale. The analysis showed high internal consistency of the teacher judgements. The scale locations from pairwise comparisons were highly correlated with scale estimates for the same students from a large-scale testing program. The results demonstrate it is possible to efficiently obtain highly reliable and valid teacher judgements using the process of pairwise comparison. Reliability indices are also provided for a series of small-scale assessments that used the same methodology in a range of other domains. The results support the findings of the main study. The article discusses the benefits of using the method to supplement and validate results from large-scale testing programs.

Demands for accountability have given rise to the implementation of large scale testing programs in Australia and internationally, such as Australia's National Assessment Program – Literacy and Numeracy (NAPLAN), the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMMS) and Progress in International Reading Literacy Study (PIRLS) assessment programs. However, there is evidence that externally imposed testing programs do not have a sustained impact on student achievement and it has been argued that teacher assessment is more effective in raising student achievement. On the other hand, it is often argued that teacher assessments are less reliable than the results of testing programs.

If methods are found to obtain reliable and valid teacher judgements, they will solve many of the issues raised in the literature. The objective of the study reported in this paper is to show that it is possible to obtain highly reliable and valid information from teacher judgements using Thurstone's method of pairwise comparisons in the assessment of writing. The experimental approach taken in the study draws on the theoretical connection between Thurstone's model for pairwise comparison and item response models used widely in educational testing. The paper also discusses the use of ordered collections of scripts as exemplars of performance to characterise the development of writing ability.

The study was motivated by prior investigations in which a very high correlation was observed between scale estimates obtained from assessments of both written narratives and essays using (i) an elaborate marking rubric, and (ii) pairwise comparisons. To demonstrate validity of the pairwise method, the study also aims to replicate this previously observed correlation in the assessment of writing. The study was also motivated by a number of small studies in other domains that show promising results, which are reported in this paper after the main results.

Background

In Australia, substantial sums of money have been invested by state and federal governments in an attempt to obtain consistent teacher judgments, as evidenced for example by the scope of the evaluation by Louden, Chapman, Clarke, Cullity and House (2006). Even larger sums of money, however, have been invested in large scale testing programs, at a state level up until 2007, and at a national level from 2008 onwards. Australian governments invest this money in large part because of an expectation that educational systems should be accountable for student achievement and that parents and the community should have access to consistent and comparable information about the "performance of schooling" in Australia (Ministerial Council for Education, Employment, Training and Youth Affairs 2008, p. 12). Similar trends are evident at a national level in the US and England and at an international level, with many countries choosing to participate in the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMMS) and Progress in International Reading Literacy Study (PIRLS) assessment programs.

Despite the enormous investments, concerns have been expressed about the quality of information obtained from testing programs. Equally, though, concerns are implicitly expressed about teacher judgements as evidenced in rhetoric that tests "enable schools to develop a more objective view about the performance of their students compared to those in other schools and in relation to state-wide standards" (Performance Measurement Review Taskforce, n.d.).

Criticisms of Large-Scale Testing

The concerns expressed by critics of large-scale testing include that: the tests primarily provide normative information and rarely diagnostic information (for example, Sloane & Kelly, 2003); any accountability associated with the tests is unlikely to lead to deep, or long term, changes in teaching practices or student learning (Sloane & Kelly, 2003); the tests lead to a narrowing of the curriculum and reduce instructional time (Chudowsky & Pellegrino, 2003; Gregory & Clarke, 2003; Groves, 2002); high performance on the test does not generalise well to other tests for which students have not been specifically prepared (Shepard, 2003); the tests ignore higher-order concepts (Gunzenhauser, 2003); and that the high-stakes natures of the assessments cause undue stress on students and teachers (Gregory & Clarke, 2003). Apart from these considerations, given the objective of comparability across schools, the models of assessment in Australia and other countries are clearly premised upon consistent practices in the administration of tests by teachers. There is, however, a lack of evidence that the administration of the tests is sufficiently consistent in all necessary respects to ensure the results are actually consistent and comparable in accordance with the explicitly stated purposes of the tests.

In reviewing recent literature of the US assessment regime, Luke and Woods (2007, p. 9) conclude, "the combination of increased testing, standardised programs, increased accountability and incentives/sanctions for schools, districts and states who do not reach targets has not been a success…while states reported test score gains for at-risk students – these gains were not confirmed in reliable, national sample testing".

In reaction to the drive towards large scale testing programs, academics and teachers have called for a greater valuing of teacher judgement. It is argued that: teachers are the primary change agents (Wyatt-Smith, 2000); most of the information students, parents and teachers have about student learning comes from the classroom (Brookhart, 2003) and; teachers' assessments of their students have a greater impact on student performance than externally imposed tests (Black & Wiliam, 1998).

Criticisms of Teacher Judgement

On the other hand, without the use of an appropriate and rigorous methodology, it cannot be expected that teacher judgements will produce reliable assessments. There is some disquiet about teachers' ability to assess, which emanates not only from proponents for large scale testing, but also those proponents who value teacher judgement. Wyatt-Smith for example recommends, "there is an urgent need to invest in teacher judgement, training it up through professional development programs focusing sharply on assessment, and through system support mechanisms including those provided through internal and external moderation networks" (2000, p. 125). Stiggins (2001, p. 6) wrote, "We still cannot guarantee the accuracy of the assessments developed and used by teachers in the U.S. and, as it turns out, millions of teachers

around the world". In particular, there are concerns about teachers' ability to make judgements reliable enough to measure student growth in learning (Bond & Caust, 2005; McGaw, cited in Clarke & Gipps, 2000).

The Use of Pairwise Comparison in Education

The method of pairwise comparisons is uncommon in educational contexts and to date has been more commonly used in the measurement of judgement and choice in market research. It has also been used in sports (Bond & Fox, 2001, p. 152). Although seldom used in education, the method of pairwise comparisons was used by Bramley, Bell and Pollitt (1998, p. 14), who concluded that their studies showed "that applying the Thurstone paired comparison method to judgements of standard of academic performance can produce interpretable results". They provide background to the method, its use in Psychophysics, and the rationale for its application in Education.

Bramley et al. (1998) reported that "the most salient difficulty from a practical point of view is the monotony of the task and the time it takes to get a sufficient number of comparisons for reliable results" (pp. 14-15). In addition, Bramley et al. (1998) reported results from a second study in which they found "a lack of relationship between the original marks scale and the rated measure produced from the comparisons" (p. 10). However, in both of the studies they focused on cases in which each script was composed of responses to a number of questions in each of a number of sections of a paper.

A basic difference between the results reported here and those reported by Bramley et al. (1998) is that in the present study, the scripts involved a single task based on a single topic. In contrast with the results reported by Bramley et al. (1998), the study reported in this article shows a very strong relationship between scaled locations from a marking rubric and pairwise comparisons of scripts. The findings reported here are particularly relevant to those domains of learning in which extended performances are the most valid means of assessment.

Rationale: The Psychometric Theory Underpinning the Study

Item response theory, particularly the Rasch model, is used in the assessment of all data in the Australian National Assessment Program – Literacy and Numeracy (NAPLAN). Item-response formats in which students respond to items or questions, and the responses are scored either automatically or by markers, are the predominantly used formats. These include multiple choice and short response items. Item-response formats are used widely in programs such as PISA and NAPLAN in assessment domains such as reading and numeracy, which can be assessed readily using items that elicit short

responses. For extended performances, rubrics are often designed and the Rasch model for polytomous data applied.

Rubrics from standardised programs are in many instances made available to teachers. However, due to the training requirements and the complexities of the rubric, their use in schools is impractical and potentially unreliable. This study aims to provide evidence that the method of pairwise comparison is a viable alternative that is more efficiently implemented in school settings. The theoretical underpinnings of Thurstone's method of pairwise comparison is closely connected to the body of item response theory, particularly the Rasch model, as elaborated to follow. The study reported here exploits this connection by developing an approach based on the method of pairwise comparison and subsequently using well-established methods of analysis of the data based on Rasch models. Assessment using the method of pairwise comparison requires little training and appears to have broad practical application in school settings.

Context of the Study

This study was undertaken in a Western Australian primary school in which teachers from across the school were asked to compare pairs of narrative performances from students in each year level. This study is part of a broader project. Its objectives are to apply measurement theory to the assessment processes to facilitate the monitoring of student growth and, in turn, to provide data for the evaluation of curriculum programs. Specifically the project aims to draw on and support teacher's professional judgment; develop processes that are efficient and that are not unduly intrusive on teaching time; strengthen the nexus between assessment process and curriculum programs; and facilitate efficient school moderation processes.

Application of Measurement Theory: Pairwise Comparisons

The design and principle for the scale construction, in the method of pairwise comparison, is based on the work of Thurstone (1927, 1959). In arguing attitudes could be measured, Thurstone developed a process and model that can be used to scale a collection of stimuli based on simple comparisons between stimuli two at a time: that is, based on a series of pairwise comparisons. He referred to the formalisation of the process as the Law of Comparative Judgement. For example, suppose that someone wishes to measure the perceived weights of a series of five objects of varying masses. By having people compare the weights of the objects in pairs, data can be obtained and the law of comparative judgment applied to estimate scale values of the *perceived* weights. This is the perceptual counterpart to the physical weight of the objects. The scale locations are inferred from the proportions of judgements in favour of each stimulus versus each other.

The Rasch model is used in the analysis of data from national tests in Australia. The Rasch model contains both person and item parameters and is therefore appropriate for item-response formats such as multiple-choice questions. The method of pairwise comparison is used in this paper to place narrative samples on the same scale. The method draws on the teachers' understanding of what constitutes a better writer without requiring prior development of a rubric. Because the model used in this paper for the analysis of data obtained using the method of pairwise comparison contains only person parameters and no item parameters, it is appropriate for the direct comparison of student scripts.

The models

Thurstone's model for pairwise comparison is a predecessor to the Rasch model (Andrich, 1978a). In applications of Thurstone's response model for pairwise comparisons, the cumulative normal distribution is used. Substituting the numerically equivalent logistic function for the cumulative normal gives the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959). The BTL model is used in this study to analyse teachers' pairwise comparisons of writing performances.

To highlight the connection to the Rasch model within the framework of item response theory, note that the BTL model is also identical to the conditional form of the Rasch model that is obtained when the item parameter has been eliminated. See Andrich (1978a) and Bock (1997) for detailed elaboration of the theoretical connection between the models. In the same way that the item parameter can be eliminated in the Rasch model, the person parameter can also be eliminated (Rasch, 1960/1980). In practice, the person parameter is frequently eliminated to implement conditional maximum likelihood estimation of the item parameters (Andrich & Luo, 2003).

The BTL model is

$$p_{ij} = \Pr\{i > j\} = \frac{\exp(\theta_i - \theta_j)}{1+\exp(\theta_i - \theta_j)} \qquad (1)$$

where θ_i and θ_j are relative locations of persons i and j on the latent trait and p_{ij} is the probability performance i is judged better than performance j.

The Study

The study was conducted in an independent (private) school in the metropolitan area of Perth, Western Australia.

Collecting Samples of Narrative Writing

All students in the school from Year 1 to Year 7 were asked to write a narrative text. The students were given the same prompt, "What a discovery! This would change things for sure!". The administration of the assessment was standardised across the school and followed the guidelines used in the state testing program.

Each teacher was asked to select from their class's work, a performance that represented a weak performance, a performance that represented average ability in the class and a strong performance. From these performances, 30 narrative texts were selected for the pairwise exercise.

Participants

All teachers in the school were invited to participate in the rating exercise. Seventeen raters were classroom teachers and these teachers represented all class levels. Two raters were school administrators and one rater was the study coordinator. The study coordinator and one other teacher had previously participated in marker training for the state testing program, the Western Australian Literacy Assessment program, commonly referred to as WALNA. No other participants had received any specific training in the assessment of writing.

Training

Training the teachers for the exercise involved a brief explanation of the requirements of the task. Care was taken in the study to remind teachers that in each instance they had to select the better performance of each pair because experience with use of the method has shown that teachers occasionally select the weaker performance rather than stronger performance. The reason this occurs is likely to be that judging which performance is weaker is inherent in the process of deciding which performance is stronger. Training also included a brief discussion of what the teachers value in narrative writing and allowed teachers from lower primary and teachers from upper primary to discuss their understandings of what is meant by the phrase "a better performance". The training took approximately half an hour. In contrast, the use of the rubric against which the results are validated required a full day of training in the Western Australian assessment program.

Pairwise exercise

Each judge received the response sheet that listed the pairs of narrative scripts they were required to compare and on each sheet there was a reminder to the judge to circle the better performance in the pair.

The number of comparisons, when each of 30 narrative writing performances is compared with each other performance, is $I(I - 1)/2 = 435$. This is far too large a

number of comparisons for any one teacher. A design was constructed in which each performance was compared with between 64 and 74 other performances. The specific pairs to be compared by each judge were generated randomly from the list of all possible pairs of performances. The number of occasions on which each script (performance) was involved in comparisons with other scripts appears in the third column of Table 1. For example, script 4 was compared with other scripts on 64 occasions. Each possible pair of performances was judged twice. Each judge compared approximately 100 pairs of scripts in total.

Script ID	Preferred	Involved	Location	Std Error	Outfit
4	0	64	-12.244	1.484	0.009
22	3	67	-10.311	0.947	0.054
29	4	65	-9.71	0.916	0.082
28	5	65	-9.112	0.94	0.055
16	9	68	-6.94	1.176	0.014
2	10	68	-5.325	1.148	0.014
17	18	73	-2.72	0.689	0.378
20	18	65	-1.982	0.603	1.210
19	21	62	-1.09	0.531	0.289
25	20	63	-0.855	0.516	0.454
6	26	64	-0.531	0.479	0.312
14	27	71	-0.362	0.448	0.529
12	27	65	-0.186	0.48	0.403
18	29	65	0.39	0.443	1.101
30	31	64	1.039	0.434	0.751
13	39	72	1.438	0.413	0.532
21	36	63	1.916	0.44	0.897
26	40	67	2.158	0.423	0.532
9	41	71	2.191	0.405	2.616
24	42	66	2.644	0.412	0.306
5	52	74	3.215	0.401	0.319
23	49	66	3.864	0.437	1.068
15	49	65	3.914	0.431	0.769
1	53	68	4.018	0.443	0.570
27	53	66	4.475	0.449	0.236
11	51	63	4.591	0.47	0.533
10	54	64	4.961	0.478	0.540
8	59	67	5.377	0.509	0.183
7	64	68	6.494	0.693	1.265
3	69	69	8.682	1.452	0.008

Table 1: Pairwise Locations, Fit Indices and Statistics

Results

Analysis of the Pairwise Data

The teachers' ratings were analysed using custom-designed software PairWise (Holme & Humphry, 2008). The software implements maximum likelihood estimations, calculates a separation index and computes mean squared standardised residuals for the purpose of testing fit.

A location estimate for each writing performance was derived from the analysis of the teachers' ratings and is shown in Table 1. The mean location is constrained to 0 in the estimation of the person locations. Performance 4 was the weakest performance as no-one judged this performance to be better than any of the other performances. Performance 3 was judged the strongest performance because, in all comparisons it was judged to be the better performance; it was compared favorably against the 69 other scripts with which it was compared. In the BTL model, where two performances are very close in standard about half the teachers are expected to select one performance over the other and vice versa.

The outfit statistic was computed in the same way as it is computed in applications of the Rasch model (Wright & Stone, 1979; Wright & Masters, 1982), except that the observed and expected scores related to pairwise comparisons instead of person-item interactions. The outfit statistic indicates the internal consistency of judgements about each script relative to the others, referenced to expected outcomes of comparisons given the scale locations of the items.

Technically, the outfit index is defined as

$$U_i = \Sigma_{j \neq i} \frac{(x_{ij} - p_{ij})^2}{p_{ij}(1 - p_{ij})} / N_{ji} \qquad (2)$$

where $x_{ij} = 0$ denotes the comparison of script i as inferior to script j, $x_{ij} = 1$ denotes the comparison of script i as superior to script j, and p_{ij} is, as defined in Equation 1, the BTL model. A commonly used range of acceptable limits for the outfit index is 0.7 to 1.3 (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). It is stressed that, although they indicate misfit to the model, outfit values of less than 0.7 nevertheless indicate high levels of internal consistency.

The Person separation index (PSI) was .982. The separation index is defined as

$$PSI = \frac{\text{var}[\hat{\theta}]}{\text{var}[\hat{\theta}] - MSE}$$

where

$$MSE = \Sigma_n \sigma_n^2$$

is the mean square error of the estimates. The Person separation index is directly analogous to Cronbach's alpha (Andrich, 1988). It is an index of the internal reliability of the assessment as a whole, which is used in item response theory. The minimum value of the Person separation index is, in practice, effectively 0, and the maximum value is 1. A higher value indicates greater internal reliability.

The magnitude of the separation index observed in this study indicates very high internal consistency and, therefore, that teachers agreed about the relative differences of the performances. Although each teacher was very familiar with only a section of the ability range represented by the performances (the Year 1 teacher knew much about early writing development, and the Year 7 teacher knew much about later writing development for example), all teachers' judgements across all performances were highly consistent.

Validating the Teachers' Judgements

In addition to establishing internal reliability, to validate the results from the pairwise exercise, an experienced marker was asked to mark the performance using a marking rubric devised for the Western Australian Literacy and Numeracy Assessment. This marking was undertaken independently of the pairwise exercise. The rubric was extensively trialed during its development and the structure of the rubric is outlined in Andrich (2006). The data obtained from the rubric were analysed with the Rasch model for polytomous data (Andrich, 1978b; Masters, 1982; Rasch, 1960/1980; Rasch, 1961) in 2005 and 2006 for students in Years 3, 5 and 7 in Western Australian schools. Table 2 shows summary psychometric information for the WALNA writing rubric. The summary information is based on an analysis of a random sample of 500 students from the larger population of students.

Item-Trait Interaction	Reliability Indices
Total Item Chi Squ 88.7	
Total Deg of Freedom 81.0	Separation Index 0.923
Total Chi Squ Prob 0.2616	Cronbach Alpha 0.921

Table 2: Summary Psychometric Information for the Rubric

The total chi-square probability of 0.2616 shown in Table 2 is the probability the observed departures from the model occurred by chance alone. The chi-square probability for the pairwise results indicates relatively good fit to the model and therefore that the assessment provided a sound basis for measurement.

The chi-square test is based on comparisons of observed and expected means for each of 10 class intervals across 9 of the 10 items. The expected means are obtained from the polytomous Rasch model given the person and item parameter estimates. It has 9 degrees of freedom per item and a total of 81 degrees of freedom across the nine items. The tenth item, omitted from the analysis, is an on balance judgement. This item shows misfit that is most likely due to violations of local dependence because the description of the item summarises the descriptions of other items.

The reliability of the experienced marker's marking had been examined by the authors on four occasions during large scale marking operations and shows her to be a highly reliable marker. Marking centers were set up with facilities for online marking of scripts for primary (Years 3, 5 and 7) and secondary (Year 9). On each occasion, a single set of 25 scripts were marked by approximately 200 markers in the marking centre for primary students and by approximately 100 makers in the marking centre for secondary students. Each marker's scores were correlated with the set of averages for the scripts across the marking centre as a whole. The correlations of the marker who participated in this study were examined. On each of the four occasions, the correlation was approximately .99, indicating the marker is highly reliable.

Figure 1 shows the correlation between the marking of the performances, using the WALNA rubric, and the teachers' judgements of the relative differences of performance using the method of pairwise comparison. The regression line of best fit is included in the graph. The correlation was $r = .921$. This high correlation shows that in addition to being internally reliable, the pairwise results have high concurrent validity referenced to the rubric. The person separation for the WALNA test, shown in Table 1, is .923. Based on this and the separation index for the pairwise assessment, the disattentuated correlation between the estimates obtained from rubric and the pairwise method is approximately .966. This correlation with a well-developed rubric, used in large scale testing, indicates strong concurrent validity. Thus, the evidence indicates that the method of pairwise comparisons was a highly reliable and valid form of teacher assessment of narrative performances.

Summary of Results in other Domains

In addition, the pairwise methodology has been used in recent work in a number of other domains by the Curriculum Council of Western Australia carried out during 2007 and 2008. The method used in all of the exercises was virtually the same in all respects as the main study reported in this article. Table 3 summarises the studies and their results.

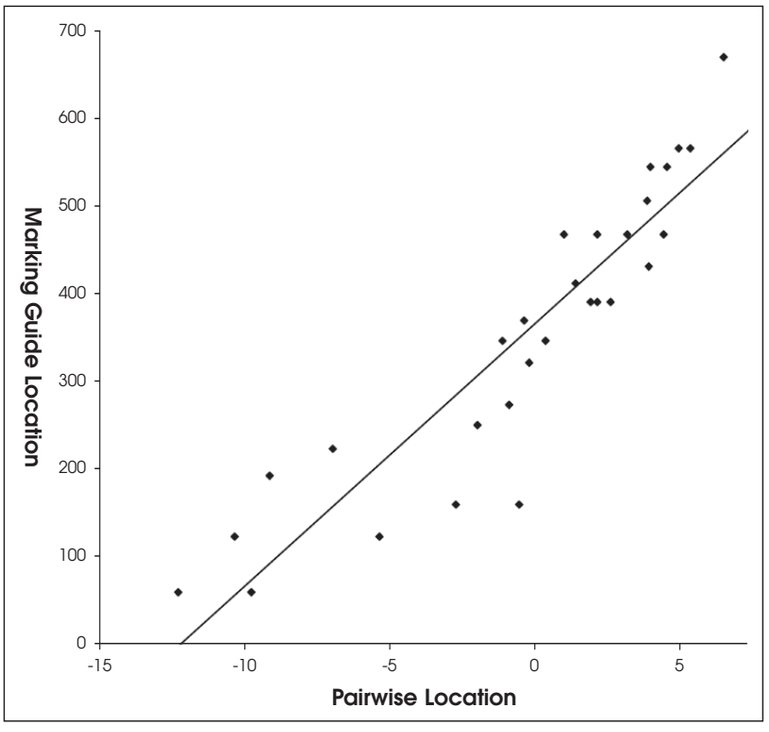

Figure 1: Correlation of Pairwise Locations and Locations from the Rubric

Discipline	Number of Judges	Number of Performances	Separation Index
English (Creative Writing)	20	31	0.95
English (Narrative & Essay)	15	56	0.97
Visual Arts	20	30	0.95
Philosophy	9	11	0.84
Accounting & Finance	11	19	0.95

Table 3: Summary Results from Pairwise Exercises in Other Domains

Table 3 shows that the separation indices were very high in all domains, indicating high internal consistency within and among judges. All of these exercises involved open-ended tasks and performances rather than tasks that required closed and short responses. In the case of Creative Writing, the performances included poetry and short stories. In the case of Visual Arts, the products of the performances were paintings and three-dimensional pieces of artwork (e.g., sculptures). The Philosophy exercise required students to clarify and evaluate a philosophical argument. For Accounting and Finance, students wrote an essay. Pairwise comparisons in each domain were based on global judgements rather than on judgements of aspects of the performances.

In addition to the exercises reported above, another was conducted for Chemistry at university level. The exercise involved comparing students' extended responses to questions about key concepts of interest in each of three laboratory experiments. In this exercise, there was a model answer to each of the questions. Pairwise comparisons were made among the responses of 18 students. The separation indices for the questions are shown in Table 4.

Question	Person Separation Index	Correlation With Lab Report Marks
1	0.952	0.967
2	0.945	0.968
3	0.952	0.984

Table 4: Person Separation Indices and Correlations for a Chemistry Laboratory Report

The scale estimates obtained from the pairwise comparison of responses were compared with marks awarded to the laboratory reports as a whole. The correlations between the scale estimates and the marks are shown in the last column of Table 4. All of the correlations are very high, again indicating concurrent validity.

The exercises for which summary results are provided above have not been described in the same level of detail as the main study. However, they are summarised to provide preliminary evidence of the potential viability of pairwise comparison as a form of valid and reliable assessment of extended performances in a range of disciplines.

Discussion

Using Teacher Judgements to Generate a Scale

Until 2007, the approach to obtaining consistent judgements in Western Australia was based on direct assessment against bands or levels of achievement described in curriculum documents. The State Government implemented a time consuming professional development program aimed at achieving consensus and comparability based on a Western Australian Outcomes and Standards Framework, which attempted to describe eight generic and broad levels of development across all learning areas. The approach did not, however, incorporate any checks of the reliability or validity of the assessments. In addition, the approach was complex and time consuming yet it produced very crude assessments and was considered unproductive by many teachers (Andrich, 2006; Louden et al., 2006).

Similar models of assessment have been used in various parts of Australia and more broadly. It is therefore clearly beneficial to develop rigorous, productive and efficient

methods of school-based assessment. In this study, virtually no training was required and the task took the staff approximately three hours. The findings indicate that the method of pairwise comparison is an efficient and productive method for drawing on teachers' professional knowledge to assess students' written work and to generate achievement scales. The summary results for other domains indicate the broader viability of the method.

As argued at the outset, the method of pairwise comparison can also be used by teachers to complement information from large-scale testing programs. If possible, it is advantageous to cross-reference the two sources of information to establish validity, to integrate them, and to identify anomalous information about student performance in the interests of individual students.

At the time of comparisons, in the study reported here, there was no attempt to ensure consensus through meetings of the kind involved in the approach to school-based assessment that was employed in Western Australia until 2007. However, because each script was compared with others by all of the judges, the location of each performance represents consensus of all the teachers' judgements provided the outfit statistic is within expected limits for all scripts. In this case, the outfit was below the upper limit of approximately 1.3 for all but one of the scripts. The outfit statistic for a given script provides an indication of the degree to which the comparisons were internally consistent given the location of the script and the locations of others with which it was compared. Thus, the outfit statistics indicate internal consistency at the level of each individual script.

Using Scaled Performances to Characterise the Developmental Continuum

The scale locations on their own are not a sufficient monitoring tool as there is no articulation of what the teachers value in writing and as such students, and to a certain extent teachers, are not provided with information about how to improve, or what needs to be taught next. However, the performances can be displayed graphically on the scale obtained from the method of pairwise comparison, in an analogous fashion to the way items are shown on scales in item maps. Displayed in this way, each performance provides an exemplar of a given level of development of writing ability and skills in the same way that items exemplify levels of performance along a continuum on an item map. That is, a collection of exemplars displayed in order on the scale exemplifies, and characterises the development of student writing. The Curriculum Council of Western Australia has used this approach to develop empirically based grade descriptors by determining grade boundaries on the scale.

In the second stage of the project, two teachers in the school will develop a resource to accompany the scale. This resource will articulate the ways in which writing

develops, as shown from the performances, but will also link to the many resources teachers currently use such as the "First Steps Writing Continuum" (Department of Education and Training, Western Australia, 1997). It is anticipated that this work will provide information for teachers' learning programs because it will provide fined-grained information about writing development not articulated in the curriculum documents which contain more generalised information. The links to the Curriculum documents are seen to be crucial, so providing those links will allow teachers to draw on a range of information to inform their teaching of writing.

Potential Limitations and Issues

Bramley et al. (1998) claimed that in practical terms, the most salient difficulty involved with the method is the monotony of the task and the time required to obtain reliable results. Although this may seem true relative to the time taken to mark external examinations, the method is considerably more efficient than many approaches to school-based assessments that use teacher judgements, such as those evaluated by Louden et al. (2006).

Bramley et al. (1998, p. 15) proposed that, in light of the perceived difficulty, "future research might concentrate on getting judges to put scripts in rank order". Relevant to this point, Thurstone (1928) proposed a practical application of his methodology, that the process of pairwise is only required in a calibration phase of a study. He suggested that once a set of statements had been scaled (using the pairwise process), it would be possible to measure other participant's position on that scale. He suggested presenting a final list of about 25 statements of opinion to participants and asking them to indicate with which statements they agreed or disagreed. The score for each person, and hence their position on the scale, is the average scale location of all the statements endorsed.

Analogously, student performances may first be calibrated and listed in order. Future performances may then be measured against the calibrated sample of performances by deciding which of the performances it is most alike in the ordered list. This has the effect of giving the same score to the new performance as the calibrated script that has been chosen. The same total score implies the same scale estimate because the total score is the sufficient statistic for the scale location. Provided that teachers in different contexts use common assessment tasks, calibrated samples of performances may in principle be used to efficiently assess new performances by this approach.

The psychometric theory involved with the second stage of this process warrants further investigation and elaboration. The reliability and validity using the second stage must be empirically tested to determine whether it matches that obtained in the calibration stage. These tasks are beyond the scope of the current article. This is an important area

for future research because it potentially avoids unnecessary duplication of the most time consuming phase of calibrating performances.

At the time of undertaking the comparisons, the teachers expressed concerns about their capacity to judge the performances of students in the ability range that was unfamiliar to them. The data show however that this was not an issue and the teachers shared a common understanding of development in writing. On completion of the exercise, the teachers commented on how useful it was to see the full ability range in the school and to see through the examination of the performances how writing develops. Much of the subsequent discussion focused on: (i) the ordering of the performances, (ii) the fact that performances from different year groups had been judged to be of a similar standard, and (iii) the fact that some performances from lower year levels had been judged to be better than performances from higher year levels.

The teachers who made the judgements generally found the information obtained from the method of pairwise comparison to be strong evidence of the face validity of the assessment results. A common remark is that the process of pairwise comparison forces consideration of the qualitative characteristics that distinguish one performance from another. This lends validity to the practice of using ordered collections of performances to characterise the development of ability and skills. Some of the teachers involved in the study perceived the ordering the scripts from lowest to highest scale estimates to provide valuable information for future teaching programs by characterising the zone of proximal development.

Conclusion

The requirement for accountability has given rise to large scale testing programs in Australia and internationally. There is, however, a growing body of evidence that externally imposed testing programs do not have the sustained impact on student achievement that is intended by funding these programs. Whilst it is argued that teacher assessment is more effective in raising student achievement levels, it is also often argued that the results of teacher assessments are less reliable than the results of large scale external testing programs.

This paper described a study in which teachers' judged writing scripts using the process of pairwise comparison. Scale locations were generated for the students' writing performances and data were analysed using custom software for the method of pairwise comparison using approaches that are used in the application of the Rasch model. The analysis of the data showed high internal consistency of the teacher judgements. The study also showed that the scale locations from pairwise

comparisons were highly correlated with scale estimates for the same students from a large-scale testing program, indicating strong concurrent validity.

It was argued that teacher judgements can supplement results from large-scale testing programs when appropriate. A two-staged method based on Thurstone's original work was proposed as an area of future research aimed at efficient use of the method of pairwise comparison to assess student performances.

The results demonstrate it is possible to efficiently obtain highly reliable assessments of narratives from teacher judgements using the process of pairwise comparison. Reliability statistics were also provided for a series of small-scale assessments that used the same methodology in a range of other learning domains. These statistics support the findings of the present study and indicate the viability of pairwise comparison for assessment of extended performances more broadly, in a range of domains. The findings potentially carry broad implications for monitoring student achievement and improvement given the ongoing debate about the relative advantages and disadvantages of large-scale external assessments and teacher judgements in the school setting.

Acknowledgements

The authors wish to acknowledge the valuable contributions of Andrea McCracken, Greg Stowe, and the staff at Scotch College Junior School; as well as the contributions of Jan Brandreth. We also wish to acknowledge Wai Mun Loke and the staff of the Curriculum Council of Western Australia, Raymond Driehuis, Christine Woods, Gerard Morris, and Lynda Kuntyj for the their work on the additional assessments whose results are summarised in the paper. The work was supported in part by an Australian Research Council Linkage grant with the Australian National Ministerial Council on Employment, Education, Training and Youth Affairs (MCEETYA) Performance Measurement and Reporting Task Force, UNESCO's International Institute for Educational Planning (IIEP), and Pearson Assessment as Industry Partners; David Andrich and Stephen Humphry Chief Investigators.

References

Andrich, D. (1978a). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 449 – 460.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications.

Andrich, D. (2006). *A report to the Curriculum Council regarding assessment for tertiary selection*. Perth: Curriculum Council of Western Australia. [Available from: www.curriculum.wa.edu.au/internet/_Documents/Publications/Andrich+Report.pdf].

Andrich, D., & Luo, G. (2003). Conditional Pairwise estimation in the Rasch model for ordered response categories using principle components. *Journal of Applied Measurement*, *4*, 205-221.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*: *Principles, Policy & Practice, 5*(1), 7-74.

Bock, D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21-33.

Bond, T., & Caust, M. (2005, November). *Silk purses from sows' ears? Making measures from teacher judgements*. Paper presented at the Australian Association for Research in Education Conference, Sydney [published January 2006].

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika*, *39*, 324-345.

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone's paired comparisons. *Education Research and Perspectives*, *2*, 1-23.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5-12

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: what will it take? *Theory into Practice*, *42*(1), 75-83.

Clarke, S., & Gipps, C. (2000). The role of teachers in teacher assessment in England, 1996-1998, *Evaluation and Research in Education*, *14*(1), 38-52.

Department of Education and Training, Western Australia (1997). *First Steps Writing Developmental Continuum*. Richmond, Australia: Heinemann.

Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into Practice*, *42*(1), 66-74.

Groves, P. (2002). "Doesn't it feel morbid here?" High stakes testing and the widening of the equity gap. *Educational Foundations*, *16*(2), 15-31.

Gunzenhauser, M. (2003). High-stakes testing and the default philosophy of education. *Theory into Practice*, *42*(1), 51-58.

Holme, B,. & Humphry, S.M. (2008). *PairWise software*. Perth: University of Western Australia.

Louden, B., Chapman, E., Clarke, S., Cullity, M., & House, H. (2006). *Evaluation of the Curriculum Improvement Program Phase 2*. Report for the Department of Education and Training prepared in the Graduate School of Education, University of Western Australia. Accessed January 10, 2009, from http://www.det.wa.edu.au/education/accountability/docs/curriculumreport.pdf

Luce, R. D. (1959). *Individual Choice Behaviours: A theoretical analysis*. New York: J. Wiley.

Luke, A., & Woods, A. (2007). Learning lessons: What No Child Left Behind can teach us about literacy, testing and accountability. *QTU Professional Magazine*, *November*, 5-9.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Ministerial Council for Education, Employment, Training and Youth Affairs (2008). *National declaration on educational goals for young Australians*. Retrieved December 12, 2008, from http://www.mceetya.edu.au/mceetya/natgoals,24767.html

Performance Measurement Review Taskforce. *A paper about the benefits of participating in national assessments*. Retrieved December 12, 2008, from http://www.curriculum.edu.au/verve/_resources/Benefits_of_participation_in_national_assessments1.pdf

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press. [Copenhagen, Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B. D. Wright].

Rasch, G. (1961/1980). On General Laws and the Meaning of Measurement in Psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine,* pp. 321-333. Berkeley: University of Chicago Press. [Available from: http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.bsmsp/1200512872]

Shepard, L. A. (2003). The hazards of high stakes testing. *Issues in Science and Technology*, *19*(2), 53-58

Sloane, F. C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory into Practice*, *42*(1), 12.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *Medical Research Methodology*, *8*, 33.

Stiggins, R. J. (2001).The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*(3) 5-15.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 278-286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529-54.

Thurstone, L. L (1959). *The measurement of values*. Chicago, USA: The University of Chicago Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wyatt-Smith, C. (2000). Exploring the relationship between large-scale literacy testing programs and classroom-based assessment: A focus on teachers' accounts. *Australian Journal of Language and Literacy*, *23*(2), 109-127.