CAV05748

MEASURING STUDENT PERCEPTIONS OF CLASSROOM ASSESSMENT

Robert Cavanagh
Curtin University of Technology

Bruce Waldrip
University of Southern Queensland

Joseph Romanoski
Curtin University of Technology

Jeffery Dorman
Australian Catholic University

Darrel Fisher
Curtin University of Technology

Paper presented to the Assessment and Measurement Special Interest Group at the 2005 Annual Conference of the Australian Association for Research in Education: Sydney.

Abstract

The investigation developed and applied an instrument to measure student perceptions of the assessment procedures applied to gauge their classroom learning. The rationale for the research centred on the paucity of research into students' involvement in decisions about assessment in light of the importance often assigned to teacher initiated and executed assessment of students' learning. The study aimed to develop an interval-level scale to measure five aspects of student perceptions of classroom assessment: congruence with planned learning; authenticity; student consultation; transparency; and accommodation of student diversity. Following item writing and piloting, data were obtained from 320 students responding to 30 items on a four point response scale (*almost always*, *often*, *sometimes*, and *almost never*). The Rasch rating scale model was then applied to examine the fit of the data to the measurement model for the items. This revealed good data to model fit for the majority of the items when data were assigned to a three point response scale (collapsing of *almost always* and *often* categories). The report describes the analytic techniques and results, how the instrument could be improved, and identifies common and uncommon student perceptions based on a *post-hoc* analysis.

Address correspondence to Dr Rob Cavanagh
Curtin University of Technology
Department of Education
GPO Box U1987
Western Australia 6845
Email: R.Cavanagh@curtin.edu.au

MEASURING STUDENT PERCEPTIONS OF CLASSROOM ASSESSMENT

Background

Epistemological considerations

The role and function of student assessment in the classroom can be viewed from two interrelated perspectives: first as part of the teaching process; and second as part of the learning process.

Assessment has traditionally been identified as an essential component of teaching. For example, Barry and King (1998) proposed a three phase model of teaching in which teaching is explained as a cyclical process of planning, teaching and evaluation. They also identified the purposes of assessment in relation to the persons and organisations that have use for the results of assessment. These include students, teachers, parents, schools, educational system, government, the community, employers, and tertiary institutions (Barry and King, 1998, p. 330). However, the forms of assessment and specific assessment tasks employed in schools are overwhelmingly decided by teachers and administrators. A separate yet related matter concerns the resulting choice of assessment tasks. Even though reports like *The Status and Quality of Teaching and Learning in Australia* (Goodrum, Hackling, & Rennie, 2001) have asserted that assessment is a key component of the teaching and learning process, teachers tend to utilise a very narrow range of assessment strategies and in practice, there is little evidence that teachers actually use diagnostic or formative assessment strategies to inform planning and teaching (Radnor, 1996). The likely cause of this problem is a bifurcation between the assessment practices of teachers and the reasons for assessment - the pedagogical basis for assessment is neglected.

With regard to the pedagogy underlying assessment, Barry and King (1998, p. 330) considered that in an ideal world, assessment enhances learning, provides feedback about progress, stimulates motivation, builds self-confidence and self-esteem, and develops skills in evaluation. Similarly, Reynolds, Doran, Allers, and Agruso (1995) argued that for effective learning to occur, congruence must exist between instruction, assessment and outcomes. So, while assessment is a core component of teaching, it also has a key role in learning. Consequently, the rationale for this study was a view that students need to be more involved in decisions about classroom assessment and that this involvement requires they understand assessment processes and the implications for themselves as learners. Notwithstanding the strength of the argument for students being involved in decision-making about their assessment tasks, there is little contemporary evidence of such involvement (Fisher, Waldrip, & Dorman, 2005), and hence a problem is presented to researchers embarking on investigation of student involvement in classroom assessment.

Given the paucity of research into student involvement in classroom assessment, there is an absence of relevant theory upon which to ground an empirical investigation of this phenomenon. One way to overcome this problem is to examine the research on assessment from the teaching perspective, and then to reframe the findings of this research from the student perspective.

In terms of teacher practice, Harlen (1998) advises teachers that both oral and written questions should be used in assessing student's learning. Similarly, the inclusion of alternative assessment strategies, such as teacher observation, personal communication, and student performances, demonstrations, and portfolios, have been offered by experts as having greater usefulness for evaluating students and informing classroom instruction (Brookhart, 1999; Stiggins, 1994). With a similar intent, Barksdale-Ladd and Thomas (2000) identified five best practices in assessment: providing feedback to help students improve their learning; conceptualising assessment as part of a student's work which can go into a working portfolio; providing flexibility so that assessment does not dominate the curriculum; ensuring that assessment informs instruction to help teachers improve their teaching thereby ensuring student learning; and using more than one measuring stick to assess students' learning. Further, McMillan (2000) identified authenticity, feedback opportunities, validity, fairness, ethics, efficiency, feasiblity and utilising multiple methods as important characteristics of assessment.

The North West Regional Educational Laboratory (1995) took a more global view of assessment and identified the characteristics of quality assessment in schools. These included: reviewing assessment instruments and methods for cultural and other bias, aligning assessments of student performance with the written curriculum and actual instruction, and teaching students to evaluate their own work through peer and self-assessment. In another more general study, Stern and Algren (2002) employed three assessment criteria in their review of assessment in science curriculum materials: the extent to which assessment tasks align with the goals of the materials, the extent to which the items focus on student understanding, and the extent to which assessment informs instruction.

Of particular significance for the present study is that when Dietel, Herman and Knuth (1991) noted several important characteristics of good assessment, they also drew attention to the need for student involvement in the design and implementation of assessment. They argued that good assessment should involve students in setting the goals and the criteria for assessment and also performing tasks that measure meaningful instructional activities - activities that should be contextualised in real-world situations.

Methodological considerations

Several decades of research into classroom learning environments has shown the merit of using rating scale survey instruments to elicit data on multiple dimensions of the teaching and learning in schools. Significantly, this method has been successfully applied, albeit on a small scale, to profile students' perceptions of assessment. This study was an American sample of 174 students in Years 4 to 12 responding to a specially-designed questionnaire (Schaffner, Bury, Stock, Cho, Boney, & Hamilton, 2000). The *Perceptions of Assessment* questionnaire developed by Schaffner et al. (2000) asked students to respond to 55 questions on "how you feel about the way your teacher finds out how much you have learned".

After reviewing the aforementioned literature on student assessment and in cognisance of the proven effectiveness of the survey research method in studies of similar phenomena, a five-element theoretical framework was developed to inform construction of an instrument to measure student views of classroom assessment.

Theoretical framework

In anticipation of the analytic techniques that were to be applied in the study, the theoretical framework underpinning the empirical investigation defines a student trait. From a behavioural research perspective, a trait is a relatively enduring characteristic of the individual that is evidenced by a certain manner of response or behaviour(s) in all situations (Kerlinger, 1986). The trait of student view of classroom assessment was conceptualised to comprise the following elements:

1. Congruence with planned learning - Students affirm that assessment tasks align with the goals, objectives and activities of the learning program;
2. Authenticity - Students affirm that assessment tasks feature real life situations that are relevant to themselves as learners;
3. Student consultation - Students affirm that they are consulted and informed about the forms of assessment tasks being employed;
4. Transparency - The purposes and forms of assessment tasks are affirmed by the students as well-defined and made clear; and
5. Accommodation of student diversity - Students affirm they all have an equal chance of completing assessment tasks.

Research objectives

The study aimed to construct a measure of how students view the assessment procedures applied in the science classroom based upon a five-element conception of learning assessment. Attainment of this objective was contingent on confirmation of the following hypotheses concerning the data elicited by the scale:

1. Measures of student ability to affirm the presence of the elements in their classroom and measures of item difficulty can be plotted on one interval-level scale.
2. The items in the measure elicit data on a dominant and possibly uni-dimensional trait.

Methodology

The Rasch rating scale model was developed to test whether data obtained from rating scale instruments conforms to the requirements of measurement - testing that data from the items of the instrument constitute measurements of the trait under question. Wright and Masters (1982) described the postulates or requirements for measurement. These requirements can be rephrased to describe those features which a number must manifest in order to be considered a measurement:

- Uni-dimensionality - the reduction of experience to a one dimensional abstraction (height, weight, intelligence);
- Qualification - more or less comparisons among persons, items, etc. (taller or smaller, heavier or lighter, brighter or duller);
- Quantification - a unit determined by a process which can be repeated without modification over the range of the variable (feet, inches, pounds, logits); and
- Linearity - the idea of linear magnitude inherent in positioning objects along a line by some device or instrument (tape measure, scale).

It should be noted that data conforming to these four requirements is interval-level in contrast to data which, whilst displaying ordinality (qualification and possibly uni-dimensionality), do not manifest intervality (quantification and linearity). This is an important distinction when the data are required for subsequent analyses. Fraenkel and Wallen (2004, p. 241) drew attention to this issue in the use of the data applied for parametric analyses:

> "It turns out that in most cases parametric techniques are most appropriate for interval data, while nonparametric techniques are most appropriate for ordinal and nominal data. Researchers rarely know for certain whether their data justify the assumption that interval scales have actually been used."

The distinction between the properties of data is not an esoteric matter concerning how data are classified. Rather, it concerns the accuracy of the data in providing a valid representation of what is being investigated. Bond and Fox (2001, p. 2) expressed particular concern about this issue in research concerning humans; "… psychometricians, behavioural statisticians, and their like conduct research as if the mere assignment of numerical values to objects suffices as scientific measurement". They further asserted that "Quantitative researchers in the human sciences need to stop analysing raw data or counts, and instead analyse measures" (p. 2). In cognisance of these concerns and their pertinence to this research, data were analysed using the Rasch Unidimensional Measurement Model (RUMM) computer program (Andrich, Sheridan, Lyne & Luo, 2000).

The empirical research proceeded through three phases. First, the instrument was developed - items were written for each of the five elements in the theoretical framework and then a pilot study was conducted utilising qualitative methods (see Fisher, Waldrip & Dorman, 2005). Students were asked to explain their understanding of the meaning of items and of the constituent terminology in each item. This process resulted in the rewording of some items.

Second, a 30-item instrument utilising a four point response scale (*almost always*, *often*, *sometimes* and *almost never*) was administered to a sample of 320 students. The sample comprised Year Eight to Ten students in 16 classes from Queensland metropolitan and rural schools. Data were analysed using RUMM. Responses were scored: 0 - *almost never*; 1 - *sometimes*; 2 - *often*; and 3 - *almost always* (missing responses were entered as 9). RUMM estimated the thresholds between the response categories for each item. A threshold is the student ability location level (logit) at

which the probabilities of students choosing two adjacent response categories (e.g. *almost always* and *often*) are equal.

Third, the results of a series of RUMM analyses were applied to refine the instrument to improve its capacity as a measure. The measurement properties of the refined instrument were then examined by a final Rasch model analysis. This procedure was *post hoc* because the original data was modified prior to the analyses. Finally, the difficulty students displayed in affirming the items within the instrument was gauged by calculating the individual item's logit location. A "logit" is a logarithmic unit, defined as the item's log odds that it will present difficulty to the students in their attempts to affirm it. The logits for items provided a calibrated estimate of overall student difficulty in affirming the presence of the elements of classroom assessment under investigation.

Results

The first RUMM analyses of data from the 30-item instrument with the four-point response scale revealed problems with the ordering of the three thresholds between the four response categories - none of the thresholds were sequentially ordered. This finding can be illustrated by examination of the category probability curve for Item 1 generated by RUMM (see Figure 1 below). Student locations (logits) are plotted on the horizontal axis ranging left to right from students who found it difficult to affirm the item to those who found this easy. In addition, the probability of a particular response category being chosen is plotted on the vertical axis and the four curves are labelled according to the respective response categories - 0 for *almost never*, 1 for *sometimes*, 2 for *often*, and 3 for *almost always*.

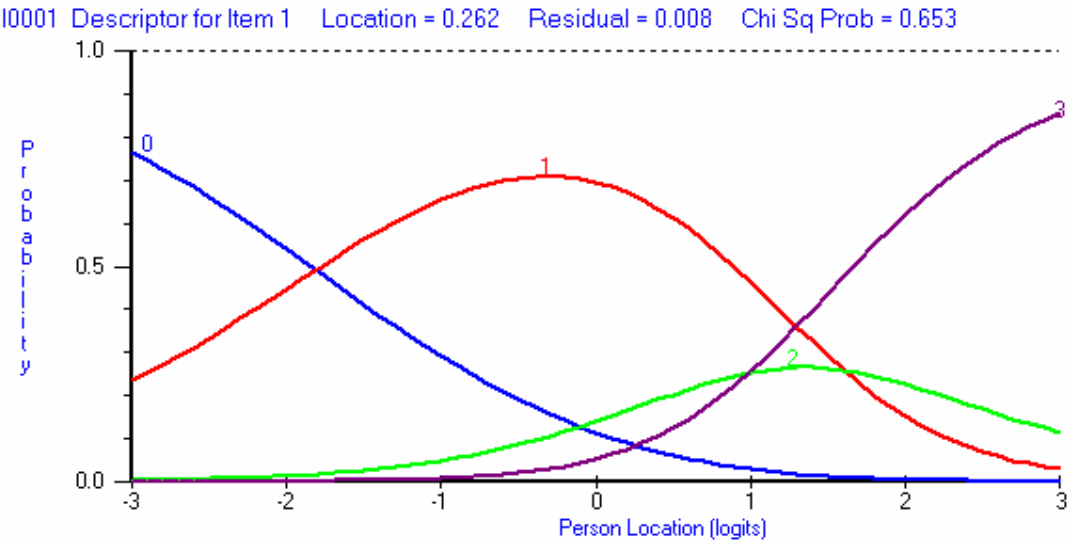


Figure 1: Four-point scale with disordered thresholds

In Figure 1, the curve for the *often* category (Curve 2) shows that the students who had a relatively high affirmative view of their assessment had a relatively low probability of choosing the *often* category in this particular item, although such a choice would be the most logical. With regard to thresholds, the person location (logit) value on the horizontal axis corresponding to the intersection of two response category curves is the person location (logit) value for the threshold between the two response categories. Two of the three thresholds are disordered since the Curves 1 and 2 intercept has a higher person location value (logit 1.7) than the Curves 2 and 3 intercept (logit 1.0).

A similar pattern of thresholds was shown for the other 29 items with the disordering being evident for the thresholds between the *often* and *almost always* categories. This finding suggested that students were confounded in their selection of these two categories in comparison to their selection of the other two categories. To test this assumption, the data from the *often* and *almost always* categories were combined into one category (*often*) and a new category probability curve was generated for Item 1 (see Figure 2).

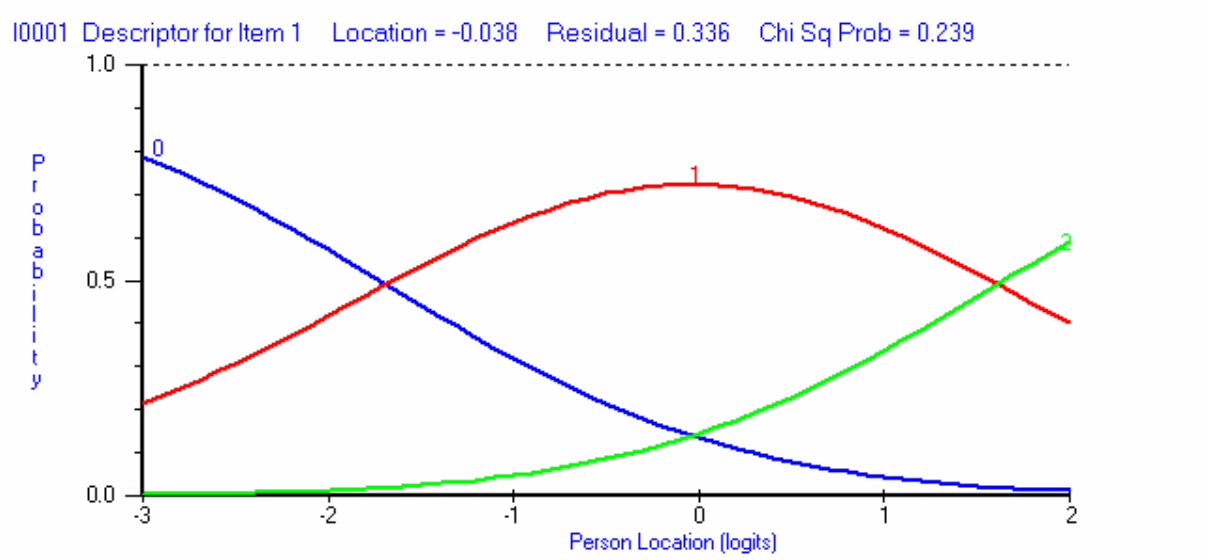


Figure 2: Three-point scale with ordered thresholds

This shows an ordering in the points of intersection between the curves in relation to student ability to affirm the items postulated to comprise the scale. The *almost never* and *sometimes* curves intersected at the student location logit of -1.8 and the *sometimes* and *often* curves intersected at the student location logit of +1.6. At this stage of the inquiry, there was evidence of some problems with how the respondents responded to the response categories offered in the original instrument and of the need to ascertain whether a three-point response rating scale might produce a better measure. Data from the *often* and *almost always* categories were combined for the other 29 items and uncentralised thresholds were estimated in a second RUMM analysis (see Table 1 below).

Table 1
Uncentralised thresholds for three-point response scale

Item	Threshold 1 locat'n	Threshold 2 locat'n	Items cont'd	Threshold 1 locat'n	Threshold 2 locat'n
1	-1.69	1.61	16	-1.45	0.36
2	-1.63	0.96	17	-0.95	1.26
3	-1.14	0.99	18	-0.63	0.59
4	-1.75	0.64	19	-0.45	1.18
5	-0.88	1.55	20	-0.58	1.58
6	-1.63	-0.04	21	-0.44	1.23
7	-1.65	-0.02	22	-0.44	1.75
8	-1.35	0.96	23	-0.78	1.14
9	-1.43	0.52	24	-0.92	1.31
10	-2.01	0.69	25	-0.05	0.91
11	-1.74	0.76	26	0.25	0.83
12	-1.00	0.59	27	-0.12	1.03
13	-0.92	1.13	28	-0.09	1.07
14	-0.54	0.86	29	-0.52	0.56
15	-0.39	0.99	30	-0.61	0.54

The thresholds for all 30 items were ordered suggesting that the three-point response scale was more appropriate for the data.

When individual item fit statistics were calculated, six of the items had high residuals (>3.0) and low Chi-square probabilities (<0.01). When data for an item fit the model well, the residual, the difference between the actual score and the score predicted by the model, should ideally be less than 2.0 and greater than -2.0 with a Chi-square probability greater than 0.05. The Rasch rating scale model is a measurement model, and when the data conforms to the requirements of the model, the data meets the aforementioned four criteria of a measure. With this in mind and given the number of items written for each element, data from the six items with high residuals and low Chi-square probabilities were deleted from further analyses.

A third RUMM analysis was conducted using the three-point data for the remaining 24 items. The summary test-of-fit statistics (see Table 2 below) showed good overall data to model fit. In an ideal fit, the mean locations of persons and items should be zero and the standard deviations should be 1.0. For these data, the mean person location was -0.89 suggesting the students experienced difficulty in affirming many of the items. Also, the standard deviation of the student locations was 1.35 due to the large variance in student transformed scores. The standard deviations of the residuals for both the persons and the items were greater than 2.0 indicating a large range in the distribution of the residuals which is probably evidence of some noise in the data. While caution should be exercised in interpreting the Chi-square probability value for large samples, the total Chi-square probability value of 0.000 suggests a dominant rather than uni-dimensional trait was measured. Cronbach alpha scale internal reliability could not be estimated due to missing data. The Rasch model tests for separation of person ability and item difficulty parameters. The separation index of 0.91 shows the data met this requirement and that the scale is an objective measure of the trait investigated.

Table 2
Summary test-of-fit statistics (N=320)

Item-person interaction

	Items		Persons	
	Location	Fit Residual	Location	Fit Residual
Mean	0.00	0.39	-0.89	-0.52
SD	0.41	1.50	1.35	2.40

Item-trait interaction		Reliability indices	
Total Item Chi Squ	193.5	Separation Index	0.91
Total Deg of Freedom	96.0	Cronbach Alpha	N/A
Total Chi Squ Prob	0.000		

Power of test-of-fit

Power is excellent
[Based on SepIndex of 0.91]

Individual item fit statistics were then calculated (see Table 3 next page). The item difficulties were located within a range from -0.71 to +0.63 logits showing that overall, the students experienced varying levels of difficulty in affirming that respective items described what was happening in their classroom.

The residuals are generally within a range from -2.0 to +2.0 with the exception of data for items 4, 7, 15, 22 and 30. Hence, as was noted for the summary test-of-fit statistics there was noise in

some of the data. Also, the majority of the Chi-square probability values were greater than 0.05 with the exception of data from items 4, 7, 20, 21, 22, 24 and 30 - some of the items did not elicit data with an ideal fit to the model. So overall, the items were measuring reasonably well but the scale could be further improved by deleting items with high residuals (>±2) and low Chi square probability vales (<0.05).

Table 3
Individual item fit statistics

Item	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob	degF
1	-0.09	0.11	0.31	295.04	310	2.85	0.58	4
2	-0.38	0.10	1.08	294.09	309	5.12	0.28	4
3	-0.11	0.10	1.06	294.09	309	18.90	0.00	4
4	-0.61	0.10	2.10	294.09	309	15.02	0.00	4
5	+0.30	0.10	-0.97	295.04	310	2.81	0.59	4
7	-0.89	0.09	2.77	295.04	310	20.45	0.00	4
8	-0.24	0.10	-1.07	295.04	310	3.49	0.48	4
10	-0.71	0.10	1.23	294.09	309	5.40	0.25	4
11	-0.54	0.10	1.44	294.09	309	3.48	0.48	4
12	-0.25	0.09	-0.22	294.09	309	3.96	0.41	4
13	0.07	0.10	0.76	294.09	309	5.44	0.24	4
14	0.12	0.10	-1.82	294.09	309	6.46	0.17	4
15	0.25	0.10	2.44	294.09	309	4.59	0.33	4
17	0.11	0.10	1.68	294.09	309	3.26	0.52	4
18	-0.06	0.09	0.18	294.09	309	4.38	0.36	4
20	0.46	0.10	-1.72	294.09	309	22.28	0.00	4
21	0.36	0.10	0.19	294.09	309	14.42	0.01	4
22	0.63	0.11	-2.72	294.09	309	13.29	0.01	4
23	0.14	0.10	-1.19	294.09	309	6.67	0.15	4
24	0.16	0.10	-1.51	294.09	309	11.19	0.02	4
25	0.40	0.10	0.95	294.09	309	3.03	0.55	4
26	0.50	0.10	1.35	294.09	309	0.98	0.91	4
28	0.45	0.10	0.60	294.09	309	2.54	0.64	4
30	-0.07	0.09	2.35	294.09	309	13.49	0.01	4

Note: Item labels are according to the original 30-item scale.

RUMM also generated an item map (see Appendix 1). The students' ability to affirm the items is plotted on the left against the difficulty of items on the right. The item difficulty plot includes the uncentralised thresholds for each item. For example, Item 22 had a difficulty of 0.63 logits (see Table 3) and this is reflected in the position of the threshold between *sometimes* and *often* for the item in the item map of +1.6 logits. The range of item difficulty logits was low (-2.2 to +1.6) compared to the range of student ability logits (-5.6 to +2.4) as was revealed by the summary test-of-fit statistics in Table 2. Also, as was previously noted, many of the items were difficult for the students to affirm and this is shown in the item map by the locations of student ability extending well below the item difficulty locations.

Finally, the difficulty of each item (logits) was presented alongside the wording of each item with the items organised according to the original five-element theoretical framework (see Appendix 2). The respective item location logits were then examined to identify common and comparatively uncommon student views of their classroom assessment.

The meaning of the scale

At least four of the items for each of the five elements hypothesised to comprise student views of classroom assessment elicited interval-level data on student ability and item difficulty. This finding provides strong evidence that the refined scale of 24 items was an accurate measure of the student trait investigated. The good fit of data to the Rasch measurement model complied with the measurement requirement for uni-dimensionality, although the trait might better be considered as dominant and comprised of elements. These findings provide confirmation of the two hypotheses tested in the study and thus show the objectives of the research were attained.

The Rasch analysis results reveal that the sample of students investigated differed markedly in their ability to affirm the elements of classroom assessment measured. Student ability to affirm the items ranged over eight logits indicating a large variation in this ability and this is likely due to differences between classrooms and within classrooms.

The items (and elements of assessment) presented students with varying degrees of difficulty in affirming the item content - the range of item difficulties was slightly less than four logits As was previously indicated, the item difficulty logits can be interpreted in terms of common and comparatively uncommon student views. For example, the logits for the items concerning *congruence and planned learning* ranged from -0.61 to+0.30 with a mean value of -0.18 logits. In comparison, the range of logits for *authenticity* ranged from -0.89 to -0.24 with a mean value of -0.53. This shows it was easier for the students to affirm the *authenticity* items compared to the *congruence and planned learning* items. Similar comparisons can be made for the other elements and these show students had more difficulty affirming the *student consultation*, *transparency* and *diversity* items (means respectively +0.10, +0.35 and +0.32). This is because the difference between the respective levels of affirmation for the five elements of assessment was consistent for the whole sample - for example *authenticity* of assessment (mean logit -0.53) was more prevalent than *transparency* of assessment (mean logit +0.35) across all of the classrooms investigated.

Conclusion

Application of the Rasch rating scale model enabled refinement and validation of a scale to measure student views of their classroom assessment. In calibrating item difficulty against student ability to affirm the presence of classroom assessment practices, the differences within both parameters were measured. That is, both student ability and item difficulty were measured reasonably accurately.

While the refined scale elicited data that complied with the criteria for measurement, it could be improved by inclusion of items for each element that would be easier for the students to affirm. If this were done, there could be a better match between the student ability transformed scores and the item difficulty measures. Also, this could allow the students who consistently expressed negative views of their classroom assessment to provide more affirmative responses and perhaps to use the full range of response categories in a more logical manner. It is possible that the problems with the original four point response scale might be redressed if extra items that were easier to affirm were included in the instrument. The refined scale might also be further improved by deleting the seven items identified from Table 3 that elicited data with less than ideal fit to the model.

In summation, the refined scale shows promise as a measure of the trait of student view of classroom assessment and the results of the Rasch analyses will have application in further development of the scale.

References

Andrich, D., Sheridan, B., Lyne, A., & Luo, G. (2000). *RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models*. Perth: Murdoch University.

Barry, K., & King, L. (1998). *Beginning teaching and beyond (3rd ed.).* Katoomba: Social Science Press.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: teachers and parents speak out. *Journal of Teacher Education*, 51, 384-397.

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

Brookhart, S.M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. ASHE-ERIC Higher Education Report. 27 (1).

Dietel, R. J., Herman, J. L. & Knuth, R. A. (1991). *What does research say about assessment?* Retrieved June 5, 2004, from http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm

Fisher, D.L., Waldrip, B.G., & Dorman, J.P. (2005, April). *Student perceptions of assessment: Development and validation of a questionnaire*. A paper presented at the Annual Meeting of the American Educational Research association: Montreal.

Fraenkel, J.R., & Wallen, N.E. (2004). How to design and evaluate research in education. New York: McGraw Hill.

Goodrum, D., Hackling, M., & Rennie, L. (2001). *The status and quality of teaching and learning in Australian schools*. Department of Education, Training and Youth Affairs: Canberra.

Harlen, W. (1998). Teaching for understanding in pre-service science. In B.J. Fraser and K.G. Tobin (Eds.) (1998). *International handbook of science education*. (pp. 183-198) Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kerlinger, F.N. (1986). *Foundations of behavioural research (3rd ed.).* New York: Holt, Rinehart & Winston).

McMillan, J. A. (2000). *Basic assessment concepts for teachers and school administrators*. ERIC/AE Digest. (ERIC Document Reproduction Service No. ED447201)

North West Regional Educational Laboratory. (1995). *Classroom characteristics and practices*. Retrieved June 5, 2004, from http://www.nwrel.org/scpd/esp/esp95_1.html

Radnor, H. (1996). *Evaluation of key stage3 assessment in 1995 and 1996 (research report).* Exeter: University of Exeter.

Reynolds, D.S., Doran, R.L., Allers, R.H., & Agruso, S.A. (1995). *Alternative assessment in science: A teacher's guide*. Buffalo, NY: University of Buffalo.

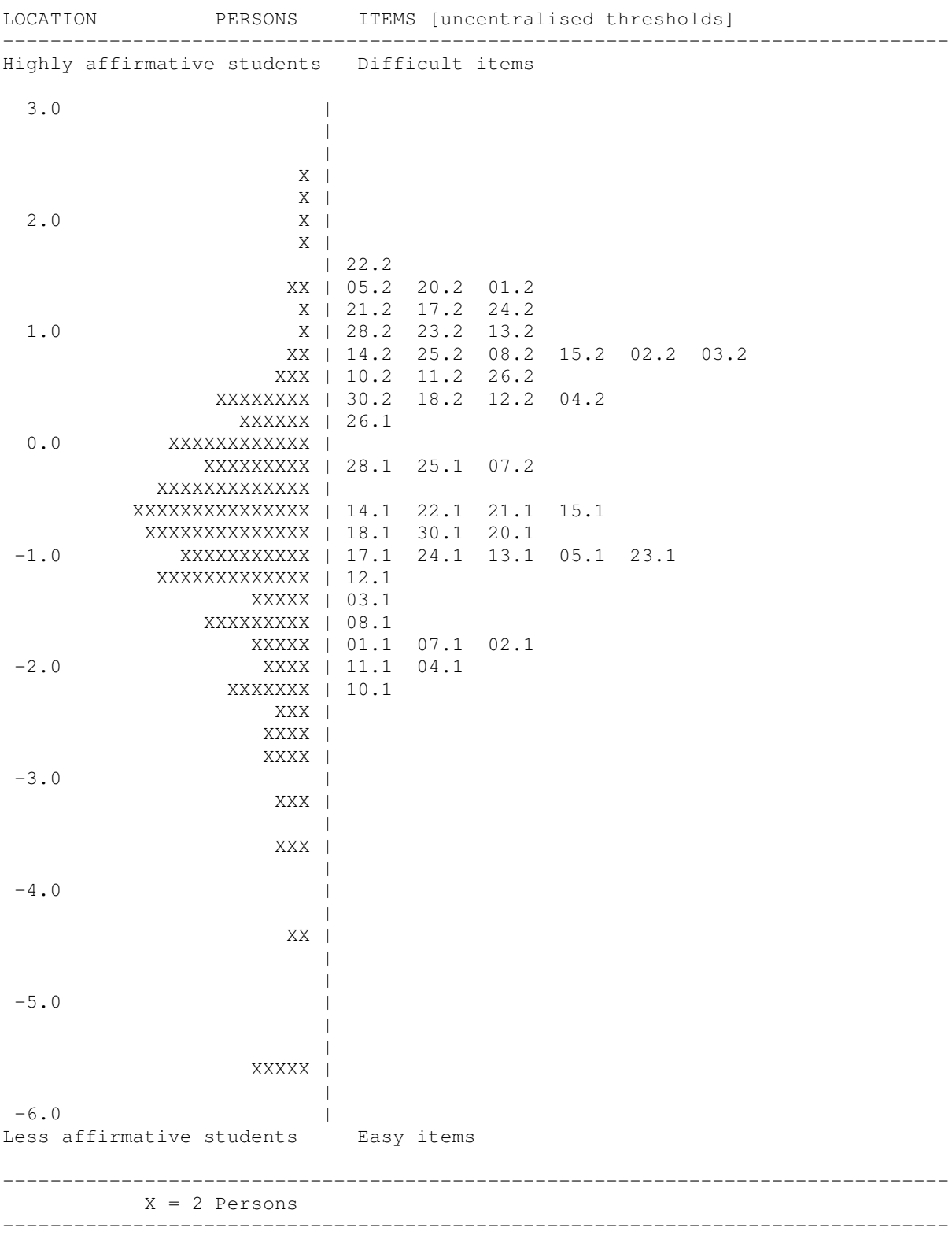
Schaffner, M., Burry-Stock, J.A., Cho, G., Boney, T., & Hamilton, G. (2000, April). *What do kids think when their teachers grade?* Paper presented at the Annual Meeting of the American Educational Research Association: New Orleans.

Stern, L., & Ahlgren, A. (2002) Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*, 39, 889-910.

Stiggins, R. (1994). *Student-centered classroom assessment*. Ontario: Macmillan College Publishing Co.

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

Appendix 1: RUMM item map



Appendix 2: Students' Perceptions of Assessment Questionnaire

Items	Logits
Congruence with planned learning	
1. My assessment in science tests what I know.	-0.09
2. My science assignments/tests examines what I do in class.	-0.38
3. My assignments/tests are about what I have done in class.	-0.11
4. How I am assessed is like what I do in class.	-0.61
5. How I am assessed is similar to what I do in class.	+0.30
6. I am assessed on what the teacher has taught me.	
Mean	-0.18
Authenticity	
7. I am asked to apply my learning to real life situations.	-0.89
8. My science assessment tasks are useful in everyday things.	-0.24
9. I find science assessment tasks are relevant to what I do outside of school.	
10. Assessment in science tests my ability to apply what I know to real-life problems.	-0.71
11. Assessment in science examines my ability to answer every day questions	-0.54
12. I can show others that my learning has helped me do things.	-0.25
Mean	-0.53
Student Consultation	
13. In science I am clear about the types of assessment being used.	+0.07
14. I am aware how my assessment will be marked.	+0.12
15. I can select how I will be assessed in science.	+0.25
16. I have helped the class develop rules for assessment in science.	
17. My teacher has explained to me how each type of assessment is to be used.	+0.11
18. I can have a say in how I will be assessed in science.	-0.06
Mean	+0.10
Transparency	
19. I understand what is needed in all science assessment tasks.	
20. I know what is needed to successfully accomplish a science assessment task.	+0.46
21. I am told in advance when I am being assessed.	+0.36
22. I am told in advance on what I am being assessed.	+0.63
23. I am clear about what my teacher wants in my assessment tasks.	+0.14
24. I know how a particular assessment tasks will be marked.	+0.16
Mean	+0.35
Diversity	
25. I have as much chance as any other student at completing assessment tasks	+0.40
26. I complete assessment tasks at my own speed.	+0.50
27. I am given a choice of assessment tasks.	
28. I am given assessment tasks that suit my ability.	+0.45
29. When I am confused about an assessment task, I am given another way to answer it.	
30. When there are different ways I can complete the assessment.	-0.07
Mean	+0.32

Note: All the 30 items in the original instrument are included and the six items that were deleted from the refined scale have been shown by italicised type.