Mathematics Education Research Journal 2006, Vol. 18, No. 2, 27-39

Surveying Primary Teachers about Compulsory Numeracy Testing: Combining Factor Analysis with Rasch Analysis

Peter Grimbeek and Steven Nisbet
Griffith University

This paper reports the use of several quantitative analytic methods, including Rasch analysis, to re-examine teacher responses to questionnaire items probing opinions related to the compulsory numeracy tests conducted in Years 3, 5, and 7 in Queensland, Australia. Nisbet and Grimbeek (2004) previously reported an interpretable and statistically acceptable 6-factor exploratory factor solution. The present paper improved on this outcome by utilising Rasch analysis to identify items with orderly sequences of scores across response categories, and to subject these to fresh exploratory and confirmatory factor analysis. The resulting 3-factor scale proved acceptable in terms of exploratory and confirmatory factor analysis as well as in terms of Rasch item analysis. The paper briefly discusses the implications of these outcomes in relation to the refined instrument's capacity to gather information about how teachers view the Queensland numeracy reporting system.

In the last 10 years, numeracy skills have been subjected to much debate and scrutiny, and increased pressure has been placed on primary schools to improve outcomes and report on progress. A review of the school curriculum (Wiltshire, McMeniman, & Tolhurst, 1994) led to the introduction of the *Year 2 Diagnostic Net* and *Year 6 Test* in Queensland schools in the mid 1990s (Queensland Schools Curriculum Council, 1996). Although the Year 6 Test was discontinued in 1997 (making way for the federally-initiated Year 3, 5, & 7 Tests), the Year 2 Net continues to be used. It has been received well by primary teachers and has had a positive impact on their teaching of mathematics (Nisbet & Warren, 1999).

Furthermore, at a national level, performance-based assessment and reporting was promulgated in the mid 1990s (Australian Education Council, 1994a), and all states were given individual responsibility for the implementation of these procedures. Consequently in Queensland, Student Performance Standards (Australian Education Council, 1994b) were introduced system-wide, but unsuccessfully, in the face of teacher opposition in spite of the provision of substantial funds for professional development of teachers (Nisbet, Dole, & Warren, 1997). Teachers rejected the increased workload imposed by the system and used the union to pressure the government to abandon the scheme.

In 1997, a National Literacy and Numeracy Plan was adopted in all states to (a) identify students at risk, (b) conduct intervention programs, (c) assess all students against national benchmarks, and (d) introduce a national numeracy reporting system (Department of Education, Training, & Youth Affairs, 2000). Consequently, annual compulsory state-wide testing was

introduced for students in Years 3, 5, and 7 in 1998. As a result, in August each year, all students in Years 3, 5, and 7 in Queensland government schools sit numeracy tests.

In Queensland tests, a broad interpretation of *numeracy* has been assumed, embracing the perspectives offered by Willis (1998) that numeracy (a) includes concepts, skills and processes in mathematics, (b) is described in terms of everyday situations in which mathematics is embedded; and (c) implies that students can choose and use mathematical skills as part of their strategic repertoire. Hence the Queensland tests cover number, measurement, geometry, chance, and data. They also test skills of calculation (written, mental and calculator methods), and real-world problem solving.

A review of the Year 3, 5, and 7 testing program (Queensland School Curriculum Council, 1999) identified potential benefits and concerns related to such state-wide testing. The suggested benefits for teachers included the identification of student strengths and weaknesses, data to inform planning and teaching, the provision of results for various groups (boys, girls, students of non-English speaking backgrounds [NESB], indigenous students), and identification of teacher professional development needs. Issues of concern included the potential of the test to narrow the curriculum, the possibility of teachers teaching to the test, the potential that assessment items would not be based on the classroom program, and the misuse of results (e.g., the publication of 'league tables' of 'good' and 'bad' schools).

The reports sent to schools after the annual tests are intended to provide administrators and teachers with information that will allow them to identify strengths and weaknesses of the school's program, compare their results with those of other schools, and to take what they might consider to be appropriate action. Information provided to the school includes results for each test item and each section (number, space, measurement and data) for each year-level, for each subgroup (boys, girls, NESB, and indigenous students), and for each student, and comparisons of these measures with the state averages. Furthermore, the incorrect answers were recorded for each item for each student, and items for which the school scored 15% above and 15% below the state average was listed.

However, it is not known whether the schools' intentions to use the information provided in the results reflect the views of class teachers (and not just those of the principal) and whether schools and teachers actually put the test results to such uses. Evidence gathered in a pilot study suggested that, although schools might have good intentions, they do not necessarily carry them out and use the results for the benefit of their school programs or the students' performance levels. The previous study by Nisbet and Grimbeek (2004) was designed to determine the extent to which schools analyse and use the test data and teachers' views of the validity or otherwise of the Year 3, 5, and 7 tests.

Traditional models of implementing innovation assume that teacher change is a simple linear process: Staff development activities lead to changes in teacher knowledge, beliefs and attitudes, which, in turn, lead to changes in

classroom teaching practices, the outcome of which is improved student learning outcomes (Clarke & Peter, 1993). Later models of teacher change regard teacher change as a long-term process (Fullan, 1982) with the most significant changes in teacher attitudes and beliefs occurring *after* teachers begin implementing a new practice successfully and observe changes in learning (Guskey, 1985). The professional development (PD) models of Clarke (1988) and Clarke and Peter (1993) are refinements of the Guskey model that recognise the ongoing and cyclical nature of PD (focussing on knowledge, attitudes, and beliefs), and teacher change.

Such models can help explain why some educational innovations are successful and others are not. The introduction of the Year 2 Diagnostic Net was successful because teachers saw positive outcomes for pupils and because they valued the Net's overall effect (Nisbet & Warren, 1999). However the introduction of *Student Performance Standards* in mathematics was a failure because teachers did not believe that the extra work entailed in performance-based assessment and reporting was worthwhile. Furthermore, they received little support from their employers for the initiative (Nisbet, Dole, & Warren, 1997).

After five years of administration of the Years 3, 5, and 7 tests we considered it appropriate to investigate the impact of the tests on schools. Hence, the current dataset was collected. The aim of the study was to investigate teacher attitudes to, and beliefs about, the Year 3, 5, and 7 tests (agreement with tests, their validity, and purposes), to determine how school administrators (i.e., principals and deputy principals) and teachers use the test results (identifying students with difficulties, and gaps in the curriculum), to explore the impact of the tests on teachers' practices (preparation for the test, influence on content and method), and to investigate the responses of teachers and pupils to the tests. A further aim was to determine the effect of school location, school size, experience, and extent of PD on attitudes, beliefs and practices (Nisbet & Grimbeek, 2004).

Responses to items revealed the following about teacher beliefs and attitudes:

- Feedback: a minority of teachers give students feedback on strengths and weaknesses, or use the results to encourage pupils.
- Diagnosis: A majority of teachers report that their school uses the results to identify topics causing difficulties, identify gaps in content taught, and identify pupils experiencing difficulties.
- Teacher change: A minority of teachers report that test results have influenced what they teach in mathematics lessons, and how they teach and assess it.
- Comparison: A minority of teachers agree that the tests are a good way of comparing their school with other schools or with the state.
- Validity: Most teachers think that the tests have little validity, in that these tests do not give an indication of numeracy ability, the quality of the school's numeracy program, or the teacher's ability to teach mathematics.
- Preparation for tests: The vast majority of teachers report showing pupils how to fill in answers before the day of the test and giving

pupils a practice test before the day of the actual test.

Nisbet and Grimbeek (2004) used exploratory factor analysis (EFA) procedures to examine the dataset. They reported that the 29 items were factorable (KMO > 0.800), although the initial Principal Axis Factoring[1] (PAF) and Varimax (orthogonal) rotation produced a six-factor solution that was neither simple (some items loaded > 0.30 on more than one factor) nor interpretable (items did not group sensibly).

After removal of items with loadings exceeding 0.30 on two or more factors or without significant loadings, a refined analysis with 15 of the 29 items resulted in a factorable (KMO = 0.766) six-factor solution that was both simple and highly interpretable. As shown in Table 1, the six factors could be labelled and described as follows (in order of factors):

- Feedback (three items): Teachers using the test results to encourage students, and to give them feedback on their strengths and weaknesses.
- Diagnosis (three items): School using results for diagnostic purposes —to identify pupils with difficulties, identify gaps in content, and identify topics causing difficulties.
- Teacher Change (three items): Tests influencing teacher practice in mathematics — what and how they teach it, and how they assess it.
- Comparison (two items): Tests as a good way of comparing the school with other schools and the state system.
- Validity (two items): The tests seen as valid indicators of the teachers' ability and the school's numeracy program.
- Preparation for Tests (two items): Teachers showing pupils how to fill in answers, and giving practice tests.

Factor scores based on these clusters of items were used to examine the relationships between the six factors and specific background variables (geographical location, school size, teacher experience, and amount of PD). For instance, it was revealed that the factor, Teacher Change, was affected by school size (teachers in smaller schools were influenced more in their teaching by the results of tests), and also the factor, Diagnosis, was affected by amount of PD (those with exposure to mathematics PD were more likely to use the tests to identify difficult topics, gaps in the curriculum and students experiencing difficulties).

The present paper set out to improve on this outcome by utilising Rasch analysis to identify items with orderly sequences of scores across response categories, and subject these to fresh exploratory and confirmatory factor analysis. The following statistical software packages were used to conduct the analyses: Statistical Package for the Social Sciences (SPSS), Analysis of Moment Structures [AMOS] (Arbuckle, 1999), and WINSTEPS (Linacre, 2004).

---

[1] The Principal Axis Factoring method extracts factors from the original correlation matrix, with squared multiple correlation coefficients placed in the diagonal as initial estimates of the communalities (See SPSS help files for more detail).

Table 1
Six-factor Solution for 15 items (PAF extraction, Varimax rotation, ≥ 0.25 loadings)

Items/factors	1	2	3	4	5	6
Use numeracy tests to give feedback on strengths (Q19)	0.92					
Use numeracy tests to inform students about weaknesses (Q20)	0.82					
I use the results of the numeracy tests to encourage students (Q21)	0.71					
School analyses numeracy tests to identify gaps in content (Q7)		0.85				
Numeracy tests identify topics causing difficulty (Q8)		0.84				
Numeracy tests identify pupils with difficulties (Q6)		0.76				
Numeracy tests influenced how I teach mathematics (Q28)			0.88			
Numeracy tests influenced how I assess pupils in mathematics (Q29)			0.81			
Numeracy tests influenced what I teach in mathematics (Q27)			0.72			
Numeracy tests a good way to compare schools (Q14)				0.88		
Numeracy tests a good way to compare schools with State (Q15)				0.82		
Numeracy tests indicate ability to teach mathematics (Q2)					0.80	
Numeracy tests indicate quality of school's numeracy program (Q3)				0.28	0.73	
School shows pupils how to fill in the answers before day of test (Q4)						0.79
School gives pupils a practice test before day of test (Q5)						0.78

Note. PAF = Principal axis factoring.

Methodology

As reported by Nisbet and Grimbeek (2004), the dataset was collected by survey method.[2] A questionnaire was constructed containing items about teachers' attitudes, beliefs and practices relating to the state-wide Year 3, 5, and 7 Tests, plus items relating to the teachers' grade level, teaching experience, school location and school size, and an item for any other comments. The results of a pilot study of 34 teachers in city and rural schools conducted in the

[2] Staff from Australian Council for Educational Research (ACER) provided assistance with the sample design and selected the sample of schools. The ACER sampling frame is compiled annually from data provided by the Commonwealth and each State and Territory education system.

previous months (Nisbet, 2003) were used to revise and expand the questionnaire items. A five-point Likert scale (*disagree strongly*, *disagree*, *undecided*, *agree*, *agree strongly*) was provided for responses, and teachers invited to comment on selected items. A sample of 56 primary schools representative of size, disadvantaged-schools index and geographical location across Queensland was selected and a total of 500 questionnaires sent to schools (from an estimate of the number of teachers in each school from the data on pupil numbers). Although the response rate (24%) was small ($N = 121$), a range of responses was received in terms of year level and position (i.e., Years 1-7; principal, deputy, and mathematics coordinator), and in terms of teaching experience (1-40 years), geographical location (capital city, provincial city, rural and remote), and school size (<20 pupils to >400 pupils).

SPSS was used as a first approximation of the frequencies per response category for Likert Scale items. It was also used to collapse response categories from five into four after using WINSTEPS (Linacre, 2004) to judge the ordering of response categories, and to conduct exploratory factor analyses. Table 2 shows the ordered response categories for one item.

Table 2
Example of Disordered Response Categories Identified Using WINSTEPS for Q4

Code (Response category)	Average score
1	-1.85
2	-0.22
3	-0.64
4	-0.24
5	-0.06

WINSTEPS (Version 3.53) (Linacre, 2004) was used to examine item statistics related to the ordering of Likert scale response categories across the 29 items (Bond & Fox, 2001). Table 2 illustrates an item in which the average score for the *disagree* (code 2) was more positive than for the *undecided* (Code 3) or *agree* (Code 4) response categories. It appears that participants found it as easy to tick *disagree* as *strongly agree* for this item, and found it more difficult to tick *undecided* or *agree*, whereas normally one would expect the *agree* response to be more difficult to tick than *undecided* or *disagree*. Based on such an examination, the two upmost-response categories of the 5-point scale were collapsed to form a 4-point scale (see guidelines for collapsing Likert scale categories in Bond & Fox, 2001, pp. 166-170). A subset of five items that continued to display disordered response categories was excluded at this point, and the remaining 24 items re-entered in an iterative sequence of exploratory (Principal Axis, Varimax rotation) and confirmatory factor analyses (AMOS CFA, Arbuckle, 1999). Finally, WINSTEPS (Linacre, 2004) was utilised to re-examine both the difficulty level and degree of fit (via Infit and Outfit statistics) for items forming part of the refined factor structure.

A fundamental issue with the use of Likert scale items is the problematic measurement properties of multi-choice response categories per item.

Differing assumptions about measurement properties determine conflicting rules of thumb for analysing such data (Grimbeek, Bryer, Beamish, & D'Netto, 2005). A reason for resorting to Rasch analysis in this paper is that it explicitly takes into account the categorical and ordinal nature of such data (Bond & Fox, 2001; Byrne, 2001; Michell, 1999). In contrast, the initial exploratory and confirmatory techniques also reported in this paper implicitly assume that the data to be analysed are parametric (equal interval, ratio) in nature. The implausibility of this assumption in relation to Likert scale items (Bond & Fox, 2001) does not deter researchers from using parametric tools. A rule of thumb for such work is to assume that scales with 4 or more points approximate the properties of interval measures (Byrne, 2001, pp. 91-92).

SPSS factor analysis was used to conduct fresh exploratory factor analyses (PAF extraction, Varimax rotation) on the 24 items remaining after using WINSTEPS (Linacre, 2004) to identify and exclude five items with persistent disorderly response categories. Subsequently, AMOS (Version 4.01) (Arbuckle, 1999) was used to undertake confirmatory factor analysis (CFA) on the items and factors identified via the EFA procedure outlined previously. A reason for using the CFA as well as EFA (i.e., PAF, Varimax rotation) procedure is that whereas EFA makes no assumption about item-scale associations, CFA explicitly tests the proposition that items cluster in specific subscales. Its rigorous testing procedures include a suite of fit estimates, ranging from statistics through various types of model fit that permit a more rigorous scrutiny of outcomes than is afforded via EFA. Table 3 shows the estimates of goodness of fit for the 15- and 7-item models.

Table 3
Estimates of Goodness of Fit for the 15-item and 7-item CFA models

Measure	Ideal estimates	15-item model	7-item model	χ^2_{diff}
X^2		87.079	12.165	74.914
df		75	11	64
Probability	$p > 0.05$	0.161	0.351	0.165
Chi/Df	0.00-3.00	1.161	1.106	
RMR	0.00-0.05	0.048	0.025	
RMSEA	0.00-0.05	0.037	0.03	
NFI	0.90-1.00	0.927	0.979	
RFI	0.90-1.00	0.897	0.960	
TLI	0.90-1.00	0.984	0.996	
CFI	0.90-1.00	0.989	0.998	
GFI	0.90-1.00	0.917	0.972	
AGFI	0.90-1.00	0.868	0.928	

Note. CFA = Confirmatory factor analysis; RMR = Root mean square residual; RMSEA = Root mean square error of approximation; NFI = Normed fit index; RFI = Relative fit index; TLI = Tucker-Lewis Index (Non-normed fit index); CFI = Comparative fit index; GFI = Goodness of fit; AGFI = Adjusted goodness of fit.

Results

WINSTEPS-based examination of the 29 items indicated that the average response was out of sequence across response categories for 13 of the 29 items.

Of these items with disordered categories, nine involved the top-most category, usually because of sparse selection of this response option. Accordingly, the two topmost categories (representing *agree* and *strongly agree*) were collapsed to form a 4-point scale. On examination, it was found that the average response remained out of sequence for just five items using this 4-point response scale. While more radical collapsing into trichotomous or dichotomous response categories could have minimised the number of out of order sequenced response categories, doing so would have infringed conventional rules of thumb regarding the use of such survey items in exploratory or confirmatory factor analyses: that is, as stated above, 4-point response scales are regarded as at the lower limits of acceptability (Byrne, 2001) for factor analysis.

The initial 6-factor solution was further examined by confirmatory factor analysis (CFA). Given the very low level of missing data (1 case in each of 9 items, 2 cases in the 10th), SPSS Replace Missing Values (SPSS, 2004) was used to replace these with the average for that item. As indicated in Table 3, estimates of goodness of fit for the initial 15-item, 6-factor model either exceeded or approximated ideal values, indicating the model to be highly acceptable in statistical as well as conceptual terms.

The 6-factor model dataset was subjected to an iterative sequence of confirmatory factor analyses (including the exclusion of six cases classified as extreme examples of multivariate kurtosis – Mahalanobis estimates) that resulted in the 7-item, 3-factor model illustrated in Table 4 in terms of EFA style output. This three-factor solution included Diagnosis (3 items), Feedback (2 items), and Validity (2 items) factors present in the initial analysis but excluded seven items related to Teacher Change (3 items), Comparison (2 items), and Preparation for Testing (2 items). From a confirmatory factor analytic perspective, the revised 7-item, 3-factor model achieved a high standard. As shown in Table 3, all 10 listed estimates of goodness of fit achieved highly acceptable levels.

Table 4

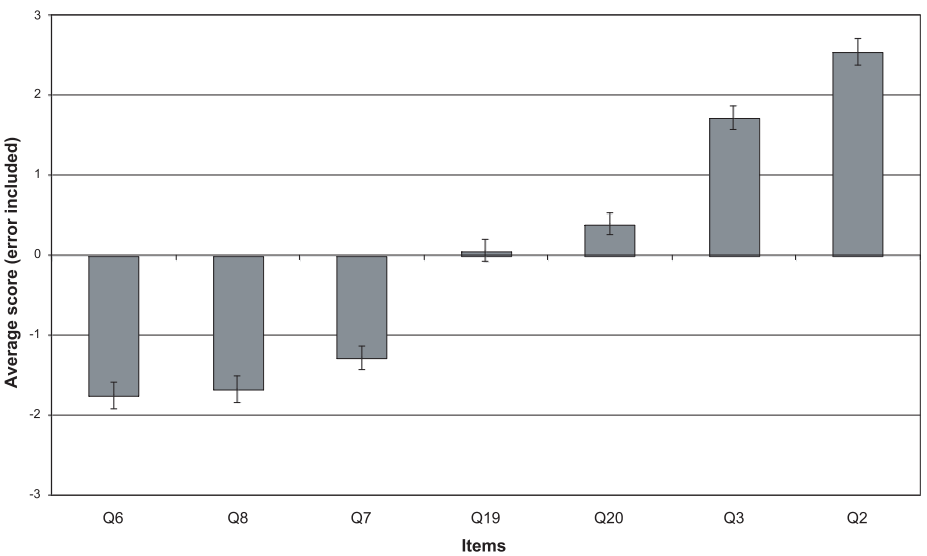
Three-factor Solution for 7 items (PAF extraction, Varimax rotation, ≥ 0.25 loadings)

Factors/Items	1	2	3
School analyses numeracy tests to identify topics causing difficulties (Q8)	0.92		
School analyses numeracy tests to identify gaps in content taught (Q7)	0.89		
School analyses numeracy tests to identify pupils with difficulties (Q6)	0.78		
Use numeracy tests to give students feedback on strengths (Q19)		0.91	
Use numeracy tests to inform students about weaknesses (Q20)		0.90	
Numeracy tests indicate quality of school's numeracy program (Q3)			0.80
Numeracy tests indicate teacher's ability to teach mathematics (Q2)			0.80

Note. PAF = Principal axis factoring.

The two models were compared by using chi-square values to compute the chi-square difference test. This test examines the significance of the chi-value obtained by taking into account the difference in chi values and the difference in degrees of freedom. As indicated in Table 3, the two models appear statistically equivalent in terms of this test.

Finally, WINSTEPS (Linacre. 2004) was used to examine the average scores per item and response category. Consistent with the prior selection process (Figure 1), the average response was in sequence across response categories for all seven items. This means the average estimates of level of agreement per item clustered consistently with the three subscale



factor structure.

Figure 1. Illustration of average score per measure.

In addition (see Figure 2), the mean square fit values computed for measures of Infit and Outfit per item all occupied the 0.5-1.7 bandwidth of values considered consistent with these items being neither too easy nor too difficult (Smith, Schumacker, & Bush, 1998; Wright & Linacre, 1994). In other words, in the context of response categories based on level of agreement, none of these items attracted 100% of responses for *strongly agree* (i.e., "too easy" would be equivalent to all participants strongly agreeing) or *strongly disagree* (i.e., "too difficult" would be equivalent to all participants strongly disagreeing).

Discussion and Conclusions

The rationale for the present study was that Rasch item analysis, when

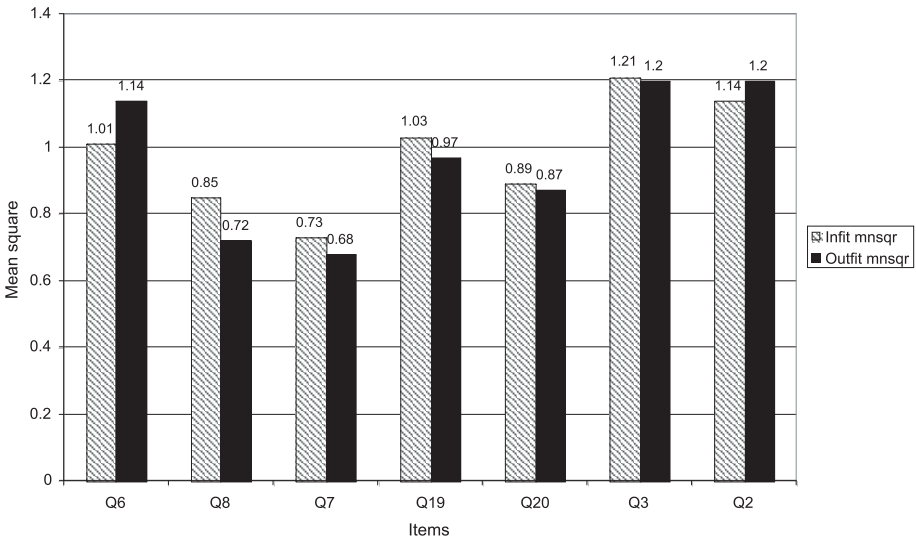


Figure 2. Illustration of Infit and Outfit measures per item.

combined with exploratory and confirmatory factor analysis, would optimise the resulting factor solution. As it turns out, the initial 15-item solution could be replaced by a 7-item model with highly acceptable statistical criteria, including those based on Rasch item analysis.

From the point of view of sampling teacher opinions related to mathematics testing, the optimised instrument offers increased ease of use in that the diminished number of items is more readily administered via the web-based surveys in vogue at present. The statistical equivalence of these two models, as per the chi-square difference test, places the two models on a level footing, but the 3-factor model is certainly more parsimonious as well as also being easier to analyse. More generally, the outcomes reported here suggest the importance and pertinence of items related to Diagnosis (3 items), Feedback (2 items), and Validity (2 items) when canvassing teacher views related to numeracy education.

As a bonus, the three subscales associated with the seven items range neatly along a single scale in terms of item difficulty (Figure 1). This means, from a Rasch item analysis perspective, the seven items could be viewed as forming a single scale that collects information across a set of seven items where the responses vary in terms of level of agreement. This interpretation of the three subscales is supported by the outcomes of a single factor PAF (details not reported here) indicating that all seven items load significantly on a single scale and could be administered and scored as such. The benefit to the test administrator is that responses to these seven items can either be scored in terms of three more specific factors (Diagnosis, Feedback, and Validity) or as a single 7-item measure indicative of the positivity of teacher attitudes concerning the Year 3, 5, 7 tests of numerical literacy.

This is not to say that the instrument has reached its academic destination

in terms of refining items to minimise ambiguity, enhancing the reliability of outcomes by generating additional items, and enhancing the validity of outcomes by ensuring that the set of items is systematically extended in terms of the measurement of factors of interest. In all of this, we would expect the combination of Rasch item analysis and confirmatory factor analysis to continue to provide highly useful information.

In these terms, a criticism of the revised factor structure is that it retains approximately one-quarter of the initial set of 29 items. As stated above, this trimmed set of items is highly acceptable from a variety of statistical perspectives (exploratory factor analysis, confirmatory factor analysis, Rasch item analysis), but the loss of three-quarters of the items could be seen to compromise the statistical reliability of judgments based on the total number of items. It reduces the precision of the Rasch person measures (larger error terms) and also diminishes validity insomuch as these rigorously applied statistical criteria enhance the purity of the measure but diminish the breadth of the field from which items are drawn. Anecdotally, as a research methodologist in contact with Australian academics, the first author can report that several seasoned researchers have rejected Rasch item analysis as a legitimate procedure precisely because of the tension between the development of such pure but sparse measures and the pressing need to develop measures with ecologically reliable and valid qualities.

It follows that potential test users in the field (e.g., mathematic educators) might consider such sparse measures to lack practical application but the authors are of the opinion that practitioners could use this test to identify responses relevant to the three remaining factors – diagnosis, feedback and validity. While the revised factor structure omits three of the six initial factors (i.e., Teacher Change, Comparison, and Preparation were omitted), the three remaining factors (Diagnosis, Feedback, and Validity) are crucial to the process of monitoring teacher beliefs about the numeracy tests.

As stated above, a major advantage of the trimmed factor structure is that it not only clarifies the measures but also reduces the length of the questionnaire, and in doing so makes the exercise of gathering teacher opinions about the compulsory Year 3, 5, and 7 numeracy tests easier and thus more likely to result in higher response rates. In short, insomuch as this instrument was designed to determine the extent to which schools analyse and use the test data, and to probe teacher views of Year 3, 5, and 7 tests, this revision meets that need very well.

Finally, it is clear that the application to the dataset of Rasch item analysis together with confirmatory factor analysis has produced an instrument with a factor structure that is statistically and possibly conceptually elegant by comparison with the model initially reported, and could be used to survey teacher views related to numeracy education testing as either a three-scale or single-scale instrument.

References

Arbuckle, J.L. (1999). Amos (Version 4.01) [Computer Software]. Chicago:

Smallwaters.
Australian Education Council. (1994a). *Mathematics work samples*. Melbourne: Curriculum Corporation.
Australian Education Council. (1994b). *Student performance standards in mathematics for Queensland schools*. Melbourne: Curriculum Corporation.
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.
Byrne, B. M. (2001). *Structural equation modeling with AMOS.* London: Lawrence Erlbaum.
Clarke, D. J. (1988). Realistic assessment. In D. Firth (Ed.), *Maths counts – who cares?* (pp. 187-192). Melbourne: Mathematical Association of Victoria.
Clarke, D. J., & Peter, A. (1993). Modelling teacher change. In B. Atweh, C. Kanes, M. Carss, & G. Booker (Eds.), *Contexts in Mathematics Education*. (Proceedings of the Sixteenth Annual Conference of the Mathematics Education Research Group of Australasia, pp. 167-175.). Brisbane: MERGA.
Department of Education, Training and Youth Affairs. (2000). *Numeracy, a priority for all*. Canberra: Commonwealth of Australia.
Fullan, M. (1982). *The meaning of educational change*. New York: Teachers College Press.
Grimbeek, P., Bryer, F., Beamish, W., & D'Netto, M. (2005). Use of data collapsing strategies to identify latent variables in questionnaire data: Strategic management of junior and middle school data on the CHP questionnaire. *Proceedings of the 3rd. Annual International Conference on Cognition, Language and Special Education,* Gold Coast, QLD.: Griffith University.
Guskey, T. (1985). Staff development and teacher change. *Educational Leadership*, *42*, 57-60.
Linacre, J.M. (2004). WINSTEPS Rasch measurement computer program (Version 3.53) [Computer software]. Chicago: Winsteps.com.
Michell, J. (1999). *Measurement in Psychology: A critical history of a methodological concept*. United Kingdom: Cambridge University Press.
Nisbet, S., & Grimbeek, P. (2004). Primary teachers' beliefs and practices with respect to compulsory numeracy testing. In I. Putt, R. Faragher, & M. McLean (Eds.), *Mathematics education for the third millennium: Towards 2010.* (Proceedings of the 27th Annual Conference of the Mathematics Education Research Group of Australasia, Townsville, Vol. 2, pp. 406-413). Sydney: MERGA.
Nisbet, S., & Warren, E. (1999). The effects of a diagnostic assessment system on the teaching of mathematics in the primary school. In O. Zaslavsky (Ed.), *Proceedings of the 23rd Annual Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 337-344). Haifa, Israel: PME.
Nisbet, S., Dole, S. & Warren, E. (1997). Cooperative professional development for mathematics teachers: A case study. In F. Biddulph & K. Carr (Eds.), *People in mathematics education*. (Proceedings of the 20th Annual Conference of Mathematics Education Research Group of Australasia, Vol. 2, pp. 369-375). Rotorua: MERGA.
Queensland School Curriculum Council. (1999). *Review of Queensland literacy and numeracy testing programs, 1995-1999: Issues paper.* Brisbane: Queensland School Curriculum Council.
Queensland Schools Curriculum Council. (1996). *The Year 2 Diagnostic Net handbook for schools – Shaping the future.* Brisbane: The Queensland Government Press.
Smith, R.M., Schumacker, R.E., & Bush, M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.
SPSS (2004). SPSS for Windows [Computer Software]. Chicago IL: SPSS.

Willis, S. (1998). Which numeracy? *Unicorn, 24*, 32-42.
Wiltshire, K., McMeniman, M., & Tolhurst, T. (1994). *Shaping the future: Review of the Queensland School Curriculum.* Brisbane: Queensland Legislative Assembly.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 6*, 205.

Authors

Peter Grimbeek, c/o: CLS, Mount Gravatt campus, Griffith University, QLD 4111. Email: <p.grimbeek@griffith.edu.au>

Steven Nisbet, c/o: CLS, Mount Gravatt campus, Griffith University, QLD 4111. Email: <s.nisbet@griffith.edu.au>