Monitoring Standards in Education: Mathematics 2002 Assessment

Andrew Stephanou
Australian Council for Educational Research
<stephanou@acer.edu.au>

Barry McCrae
Australian Council for Educational Research
<mccrae@acer.edu.au>

Rhonda Farkota
Australian Council for Educational Research
<farkota@acer.edu.au>

John Lindsey
Australian Council for Educational Research
<lindsey@acer.edu.au>

Elena Stoyanova
Department of Education and Training, Western Australia
<Elena.Stoyanova@eddept.wa.edu.au>

This paper describes the 2002 Western Australia Monitoring Standards in Education system-wide random sample assessment of student performance at Years 3, 7 and 10 in mathematics. It presents the design of the sample, outlines the methodology used to analyse the data collected, and summarises the results. Students found the Working Mathematically tests harder than the Content strand tests, and there was variation in their performance across the Content strands. Strong evidence was found that the Content and Working Mathematically items fitted on a single measurement scale.

For more than a decade, *Monitoring Standards in Education* (MSE) has provided information about educational standards in the Western Australian Government school system. MSE has two major objectives:

- To monitor and report on system-level performance in key areas; and
- To assist schools to monitor the performance of students in key areas.

MSE focuses on the reporting of students' performance in relation to the WA Outcomes and Standards Framework. This is achieved by using Rasch measurement (Andrich, 1988) to map student achievement onto scales that link the range of skills observed during the assessment programs to the achievement levels described in the Framework, and enables comparison of the levels of system performance over time.

MSE assessment programs draw random samples of approximately 10% of the state Years 3, 7, and 10 populations in Government schools. The sample draws students from the full range of metropolitan and country schools, including Remote Community Schools, Education Support Centres, and the School of Isolated and Distance Education. Random samples of this size draw sufficient numbers of students from most subgroups (males, females and students with Language Backgrounds Other Than English) to ensure reliable reporting. To help ensure reliable and valid reporting of the performance of Aboriginal and Torres Strait Island students, oversampling of these students is employed.

An important principle which guides task development for MSE assessment programs is that the assessment materials reflect good assessment practices. Extensive consultation with curriculum consultants, practicing teachers, school administrators and representatives from tertiary institutions is an integral part of the process of developing assessment materials. Particular attention is given to designing innovative assessment tasks that permit students to demonstrate a wide variety of skills and problem-solving strategies.

The 2002 MSE Mathematics Assessment

The Australian Council for Educational Research (ACER) was responsible for the development of assessment instruments, drawing of the random sample, and analysis of student performance data for the MSE Mathematics 2002 assessment program. The program had two distinct components:

- It assessed Year 10 student performance in the Space, Measurement, Chance and Data, Number, and Algebra strands of Mathematics (referred to subsequently as the Content strands).
- It assessed Year 3, 7 and 10 student performance in the Working Mathematically strand.

Student performance in the Content strands had previously been assessed in 1992, 1996, 1998 and 2000, and the items were calibrated onto a single Historical MSE Mathematics scale. Working Mathematically was assessed for the first time in 2000. It was analysed separately, with the items being assumed to belong to the same scale as the Content items for equating to the Historical scale.

Two Working Mathematically tests were developed for each of the three year levels, and two Content strand tests were developed for Year 10. The production of two tests each time means that a representative selection of items can be released for teacher reference, while sufficient items can be kept secure for equating purposes in the future. Development of assessment tasks was guided by the Outcomes and Standards Framework for Mathematics (Department of Education, 1998). About twice the number of items needed were developed and trialled, and the selection of final forms was based on analysis of the trial data.

Each Content strands test consisted of about 50% multiple-choice items. Most of the remaining items only required a short response that could be easily judged correct or incorrect. By contrast, over half of the Working Mathematically items required a more open response and many allowed for partial credit. It was kept in mind that tasks developed to assess the Working Mathematically strand should be guided by good classroom practice in this regard (see Perso, 2001), and that they should reflect the fact that the mathematics that students use spontaneously is generally that studied in earlier years (Clarke, 1988). Further, there needed to be sufficient tasks of a sufficiently "open" nature to extend students, so that "growth" across year levels could be properly assessed without the need for separate equating studies at intermediate year levels.

Random Sample Design

The design of the random sample consisted of a combination of common-case equating (i.e., same students doing pairs of tests) and common-item equating (i.e., items common to pairs of tests). This allowed for the possibility of constructing either two separate scales, one for Content (C) and one for Working Mathematically (WM), or a single MSE Mathematics scale to be equated to the Historical scale. It required three of the WM tests used in 2000 (one for each year level—3EQ, 7EQ and 10PQ), and one of the Year 10 C tests (10CQ), to be administered in addition to the six "2002" WM tests (3P1, 3P2, 7P1, 7P2, 10P1 and10P2) and the two "2002" C tests (10C1 and 10C2).

Figure 1 shows the design for the data collection. Each box represents student responses to the items in a test. At each year level, about 300 of the students who sat a P1 test, and about 300 of the students who sat a C2 test, also sat a "2000" test for equating purposes. All students in Year 10 were expected to sit two tests, a C and a WM test, or a "2002" test and a "2000" test. In Years 3 and 7 most students sat only one WM test.

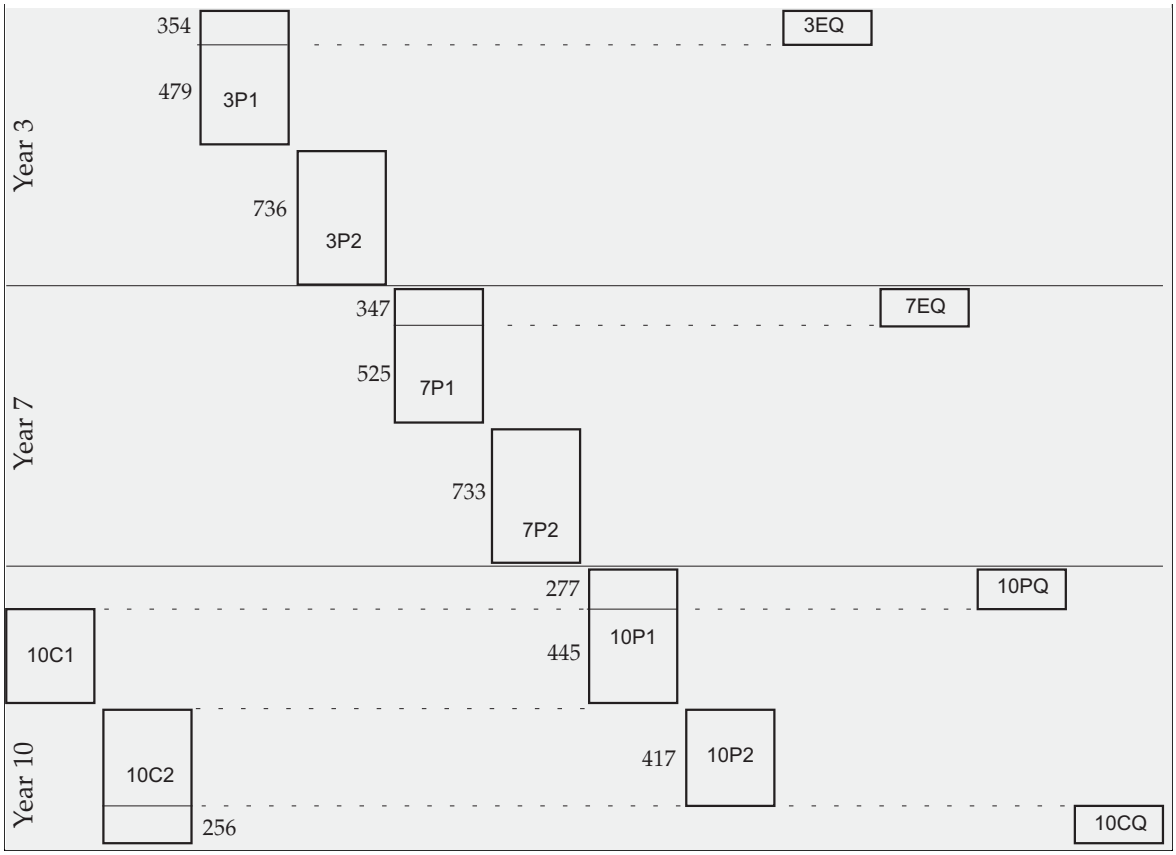

Figure 1. MSE Mathematics 2002 design showing common-case equating.

Figure 2 shows how WM or C tests for the same year level, and WM tests at different year levels, were linked through common items. For example: the P1 and P2 tests at each year level had items in common, as did 10C1 and 10C2; test 7P1 had items in common with 10P1 and 10P2 as well as (different) items in common with 7P2; etc.

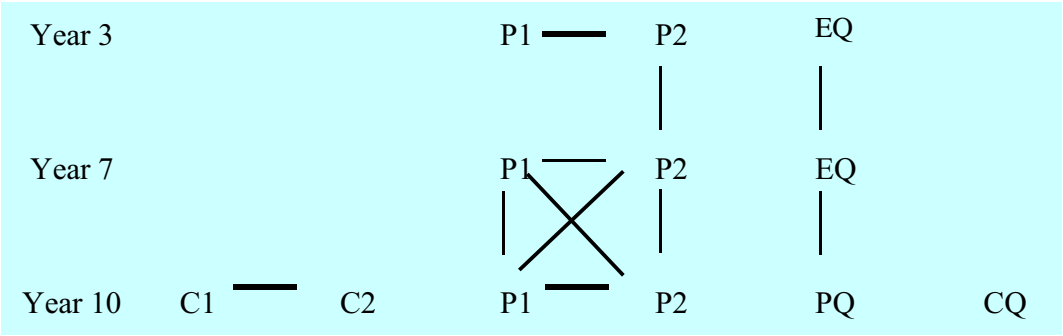

Figure 2. MSE Mathematics 2002 design showing common-item equating.

Construction of MSE Mathematics 2002 Scale

The construction of the MSE Mathematics 2002 scale started with a separate analysis, using Quest (Adams & Khoo, 1999), of the data collected with each of the 12 tests. The aim of this analysis was to check for any remaining noise in the data, after data entry and cleaning, and that each item worked satisfactorily in the test in which it was used. The analysis of fit at this initial stage led to the collapsing of the score categories of a few partial credit items, and to the exclusion of a few items from a joint analysis with the other items.

Analysis of fit was based on a number of fit indicators, and on a substantive examination of the items flagged to be problematic. All final decisions were based on substantive arguments using fit indicators as a guide. The item fit indicators used were:

- Infit statistic of Quest. This indicates how well the observed proportions of students at various ability levels who are successful on an item compare with the expected proportions (model probabilities).
- Point Biserial Correlation (PBC). The PBC of an item is the correlation between the scores on that item (0 or 1 for dichotomous items) and the score on the whole test.
- Mean and standard deviation of the abilities of the students in each score category. A necessary, but not sufficient, condition for category fit is that the mean ability of the cases assigned to each category increases with the hierarchy of the categories.
- Distance between category boundaries. When difficulty boundaries between adjacent categories are too close, there is a case for collapsing the categories.

Following the test-by-test analysis, two scales were constructed, one for Content and one for Working Mathematically, through two joint partial credit analyses (Wright & Masters, 1982) with Quest. Examination of item fit led to the recoding of a few more items, and the exclusion of a few other items whose fit did not justify their use in the construction of the scales.

The possibility of constructing a single scale which could be equated onto the Historical MSE Mathematics scale was then investigated and it was found that item fit did not change significantly from the item fit with two separate scales. Examination of the correlations between student abilities measured with two tests was also found to justify the construction of a single scale. Figure 3 shows the true and observed correlations between abilities estimated for each of two tests. The variance of ability estimates is partly

due to measurement error and consequently the observed correlation is partly due to measurement error. True correlations have been calculated from observed correlations using test reliabilities. It can be seen in Figure 3 that the correlation between 10P1 and C1, and that between 10P2 and C2, are not less than the correlation between C tests or those between WM tests. Strong evidence is thus provided to support the calibration of WM and C items on the same scale.

Consequently, a single MSE Mathematics 2002 scale was constructed by a joint analysis of all the remaining "2000" and "2002", Working Mathematically and Content, items. Once again, a few items had to be excluded because of poor fit, but altogether only 4 (out of 156) WM items, and 5 (out of 100) C items, were excluded at the various construction stages. Finally, it was determined that 1.26 logits had to be added to "2002" logits to transform them to Historical scale logits (Stephanou, 2003, pp. 21–22), and a new Quest analysis was then performed, with all items "anchored" to their Historical difficulties, to produce the final 2002 scale.

More evidence for the calibration of the Content items together with the Working Mathematically items was provided at this stage by observing that items classified as C tested similar ideas to items located in the same region of the scale and classified as WM. Further, the estimation of subgroup (all, girls, boys, LBOTE, ESB, ATSI, non-ATSI) achievement by strand (including Working Mathematically) showed that estimated abilities were generally the same whether based on all administered items or only on those of a single strand (Stephanou, 2003, p 37).

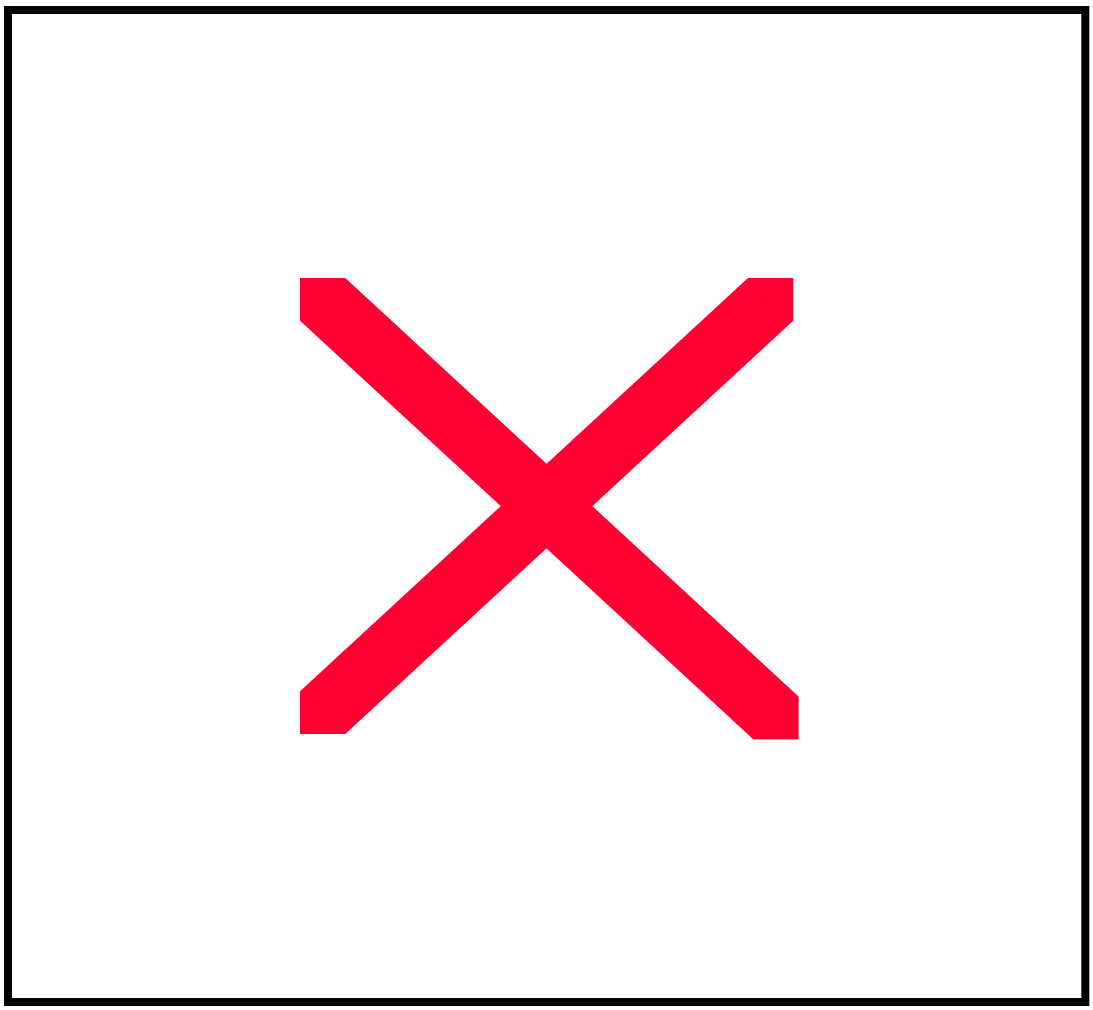

Figure 3. MSE Mathematics 2002 design showing correlations between performance on tests.

Difficulties of Year 10 Strands

A strand, from the point of view of a Rasch analysis, is a subset of the items calibrated onto the same scale, similar to the subset of items making up a test. The difficulty of a strand may be defined as the mean difficulty of the items belonging to that strand. Figure 4 shows the difficulties of the six strands—WM, Algebra (A), Measurement (M), Number (N), Space (S), Chance and Data (CD)—in logits on the Historical scale.

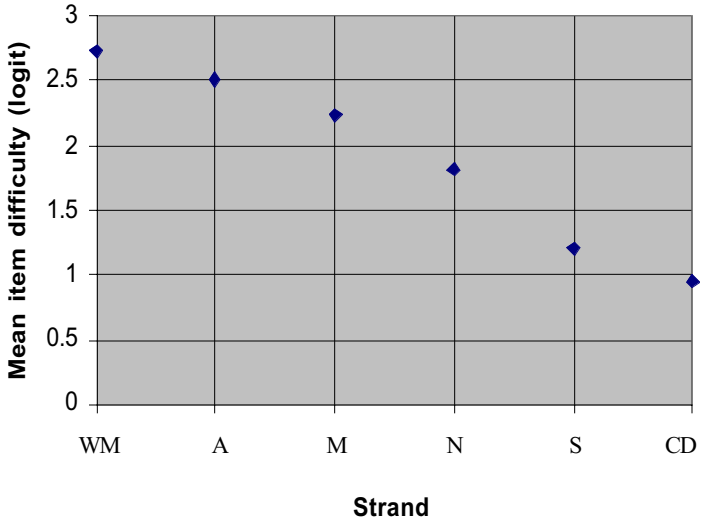

Figure 4. MSE Mathematics 2002 Year 10 strand difficulty.

The Working Mathematically items administered to the students were found to be on average the most difficult. The Algebra items were found to be on average the most difficult of the Content strand items and were about 1.5 logits more difficult than the least difficult group, the Chance and Data items. All groups of students performed better on the CD items than on the A items according to the mean percentage scores on CD and on A items. Figure 5 shows the Year 10 items by strand on the Historical scale and the distribution of overall abilities of the Year 10 students. Items are identified by number (Quest id). Each score category for a partial credit item is separately mapped (e.g., 145.1, 145.2).

It is important to note that the difficulties of items from different strands overlapped. For example, Chance and Data item 148 was more difficult than most Algebra items. Further, care must be taken not to draw conclusions such as "Students find Algebra more difficult than Chance and Data" solely on the basis of the average difficulty of the Algebra items used in the tests being greater than that of the Chance and Data items. An alternative explanation is that the Algebra items included on the 2002 tests were a more difficult selection of items according to the Outcomes and Standards Framework than the Chance and Data selection. However, a preliminary analysis does not, in general, support this alternative explanation, and in fact suggests that the Outcomes and Standards Framework over-estimates the difficulty of the Chance and Data strand (in particular, the Understand Chance substrand).

Subgroup Achievement

A multilevel modelling of the data, taking into account the hierarchical structure of the data (students clustered within schools and schools within districts), made it possible to answer questions on the statistical significance of differences in subgroup achievement through the effect of explanatory variables. MLWin software (Rasbash et al, 2002) was

used for this analysis. The explanatory variables employed were gender, ATSI, LBOTE, school location (country or metropolitan), socio-economic status, and school population.

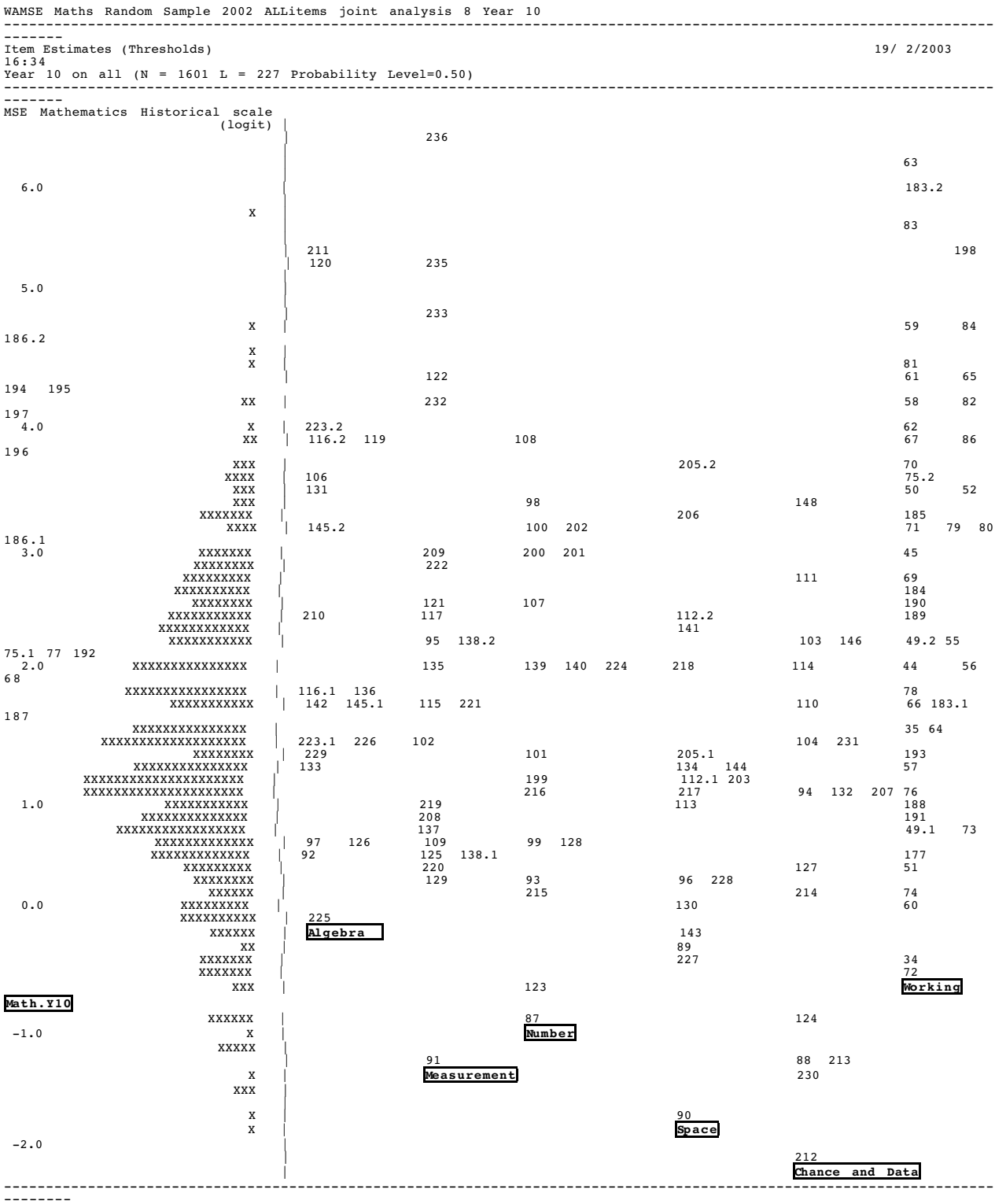

Figure 5. MSE Mathematics 2002 Year 10 items by strand on the Historical scale.

The multilevel modelling of all data in the random sample was first completed to check the statistical significance of year level differences. Data were then modelled separately for each of the three year levels to account for the variation within each year level, recognising the different characteristics of schools and students.

Full details of the analysis can be found in Stephanou (2003). The main findings were as follows:

- The difference between the mean achievements in adjacent year levels was statistically significant in favour of higher year levels.
- The differences in mean achievement between ATSI and non-ATSI students were statistically significant within each year level in favour of non-ATSI students.
- Gender differences within each year level were statistically significant just beyond the 0.05 probability level in favour of girls.
- Socio-economic status had a significant effect at each year level, but the effect of LBOTE was only significant at Year 3.
- Year 3 students were likely to perform better in smaller schools, but Year 10 students were likely to perform better in larger schools.

Conclusion

Careful design of the random sample, using common-case and common-item equating, together with innovative item design, enabled the MSE Mathematics 2002 assessment program to show that the Working Mathematically and Content strand items fitted on a single Mathematics scale. Not surprisingly, in view of their additional interpretive demands, students found the Working Mathematically items harder on average than items from the Content strands.

Students found the Algebra items harder on average than items from the other Content strands, and the Chance and Data items easier. A possible explanation for the relative difficulty of the Algebra items is that students are less exposed to Algebra than to the other Content strands. On the other hand, it is possible that Year 10 students could cope with more difficult Chance and Data work than is currently expected of them.

More gender effect (in favour of girls), and less LBOTE effect, was found than in 2000 (see Department of Education, 2002).

Acknowledgement

This paper is largely derived from the confidential technical report on MSE Mathematics 2002 prepared by Andrew Stephanou for the Western Australian Department of Education and Training (Stephanou, 2003).

References

Adams, R., & Khoo, S. (1999). *ACER QUEST*: *The interactive test analysis system*. Melbourne: Australian Council for Educational Research.

Andrich, D. (1988). *Rasch models for measurement*. CA: SAGE Publications.

Clarke, D. (1988). *Assessment alternatives in mathematics.* Melbourne: Curriculum Corporation.

Department of Education. (2002). *Student achievement in mathematics: Western Australian government schools 2000.* Perth: Author.

Department of Education. (1998). *Outcomes and standards framework. Mathematics student outcome statements.* Perth: Author.

Perso, T. (2001). Working mathematically: What does it look like in the classroom? In *Mathematics: Shaping Australia* (Proceedings of the 18th Biennial Conference of the Australian Association of Mathematics Teachers, pp 352–356). Adelaide: AAMT (CD-ROM).

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2002). *MLwiN (Version 1.10.007): Interactive software for multilevel analysis*. London: Centre for Multilevel Modelling, Institute of Education, University of London.

Stephanou, A. (2003). *Technical report (MSE Mathematics 2002) for the WA Department of Education and Training*. Melbourne: Australian Council for Educational Research.

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