A Comparison Among Three Different Approaches to Mathematics Assessment

Rosemary Callingham
University of New England
<rcalling@pobox.une.edu.au>

The same group of Year 10 students undertook three different mathematics assessments. Two of these were based on developmental continua and addressed higher order thinking, but in different ways. The third was a multiple-choice test of mathematical skills and knowledge, appropriate to the Year 10 cohort. Results suggested that although all the assessments were reliable, and no test bias was detected, there were some differences between the assessment outcomes across sub-groups and by mathematics course being followed. The implications of these findings for teachers and test developers are discussed.

Over the years there have been many calls to improve mathematics assessment (e.g. Grouws & Meier, 1992). These have ranged from appeals by Newmann and associates (1996) for "authentic achievement" and the development of "rich tasks" (Clarke & Clarke, 1999), to the use of clinical interview techniques, especially in the early years of schooling (Doig & Hunting, 1995) and developmental assessment (Pegg, 2002). Alongside this has come increased accountability, with all Australian states now implementing whole cohort numeracy testing in Years 3, 5, and 7 for benchmarking purposes, most of which is now measured by multiple-choice machine-scored tests. Underlying much of this activity is an implicit assumption that the same construct is being measured regardless of the assessment approach, and that different assessment methods essentially provide the same information.

Different approaches to assessment, however, are underpinned by diverse philosophies, and this affects the assessment development and construction process. Authentic assessment is claimed to be closer to teaching and learning strategies and thus to deliver improved quality of information because the students are not placed in an unfamiliar situation. Teachers who know the students generally mark the work and students are familiar with the classroom teacher's expectations. In contrast, proponents of objective testing point to the need for reliability, and the potential bias of teacher marked assessment (Shepard, 2000).

There is considerable evidence that students perform differently on diverse types of tests (Caygill & Eley, 2001). Where students are allowed to verbalise their answers and talk to an interviewer, for example, achievement tends to be higher than when they are asked to write a response to the same question. Similarly, providing a set of answers, as in a multiple choice format, typically leads to higher responses than a format where students have to construct their own answer. This raises questions about the inferences drawn from assessment information. Is it possible, for example, to infer that a student who performs well on a multiple choice mathematics test will also do well on a performance assessment demanding higher order thinking skills? Conversely, does a student who demonstrates higher level thinking necessarily perform well on a test of mathematical skills?

This raises issues about the nature of the construct being assessed, and the choice of assessment method according to the information that is wanted, and its potential use. If higher order thinking is the target then the assessment should be designed to address this, and similarly, if mathematical skills are the target then this should

become the focus (Shepard, 2000). This reinforces the notion that the target construct needs to be clarified before the assessment is developed. In turn, the nature of the target construct, together with the purpose of the assessment, may influence the choice of assessment method.

This paper reports the outcomes from three different mathematics assessments undertaken by a group of Year 10 students in Tasmanian government high schools. Each assessment addressed mathematical ideas, but was underpinned by a different perspective, and each assessment took a different form. There were two questions of interest:

1. How closely related were the outcomes from the three assessments?
2. Did the assessments provide the same information across all groups of students?

Details of the instruments used and the underlying perspectives are described below.

Assessment Instruments

Mathematics skills were addressed in an original 34 item multiple-choice test that covered the range of content strands in the mathematics curriculum. The test items came from trial versions of a Survey of Mathematics Skills (Assessment Research Centre (ARC), 2001), the Third International Mathematics and Science Study (TIMSS) publicly released items for the final year at school (The Third International Mathematics and Science Study, 1995), and the Tasmanian Year 7 Numeracy Monitoring Program 1998 (Department of Education & Catholic Education Office, 1998). All items had been used in other contexts and were considered appropriate for the cohort by experienced mathematics teachers. Students filled in their choice of answers on a scannable recording sheet. The variable was labelled MAT.

The Collis-Romberg Mathematical Problem Solving Profiles Form B (Collis & Romberg, 1992) was used to gain an objective measure of students' higher-order thinking skills in mathematical problem solving. This instrument was based on the SOLO Taxonomy (Biggs & Collis, 1982) and has been extensively validated. The test follows a super-item format (Cureton, 1965) in which there is a common stimulus and a set of questions targeting increasingly higher levels of thinking. A correct response to a particular question is considered to be evidence of a particular level of thinking – in SOLO terms, unistructural, multistructural, relational, or extended abstract and the method of response is not considered when the judgment is made. Teachers marked each question correct or incorrect on the scannable sheet. Following computer scanning, the final codings were transformed into a 0-4 score on each question following the instructions in the manual. The variable was labelled CRPC.

The final test was a performance assessment task, "In a Spin", using teacher judgment, that aimed to bring assessment closer to teaching. This was set in the context of a game involving spinners and targeted students' understanding of statistical variability and probability. Students made and tested spinners, and predicted the outcomes from spinning pairs of spinners. The eleven activities that comprised the task were treated as learning and teaching tasks in their own right, and were completed by students in their classrooms supported by any of the normal teaching strategies employed by the teacher concerned. The only proviso was that teachers could not tell students the answers. Each activity had an analytical scoring rubric associated with it based on the anticipated quality of students' responses, and was linked to a developmental continuum that addressed higher order thinking

responses to the activities using the rubrics provided and their professional judgment. Filling in the appropriate bubble on the scannable sheet indicated the marks. The variable was called IAS.

The three assessment processes thus had similarities and differences. All assessments required mathematical skills and understanding. Two, the performance task (IAS) and the problem-solving test (CRPC), were developmental in nature, while the mathematical skills test (MAT) addressed increasingly difficult mathematical content. The performance task asked students to provide written explanations of their responses, whereas the problem-solving test demanded minimal writing skills and the mathematical skills test required no writing. Teachers marked both the performance task, using a set of scoring rubrics, to inform their professional judgement, and the problem-solving test, using a set of right/wrong answers that required minimal teacher judgment. The mathematical skills test was machine scored. A desire to link assessment more closely to teaching and learning underpinned the performance task, the other two assessments were driven by the need for objectivity.

Methodology

The Sample

Students in Year 10 came from 13 different government high schools from all parts of Tasmania. The total number of students completing all three assessments was 685, of whom 350 (51.1 percent) were male. The students were undertaking one of three mathematics courses, MT420 (lowest), MT421 (middle) and MT422 (highest). The numbers of students in each course is shown in Table 1. The course being taken by some students was not indicated, and thus the overall number of students shown in this table is lower than the sample total.

Table 1
Number of Students by Course

Number	MT420	MT421	MT422
Males	49	177	74
Females	35	167	101
Total	84	344	175

Administration

The tests were all undertaken within a three week period in October/November 2001, approximately three weeks before these Year 10 students left high school to move to senior secondary colleges. All responses and codings from the performance task were filled in on scannable computer mark sheets. Following scanning of the mark sheets, the data were Rasch scaled using the computer program Quest (Adams & Khoo, 1996) and all results were placed on the same scale so that they could be directly compared. Ability estimates, in logits, were obtained for each student on each assessment from the Rasch scaling process (Bond & Fox, 2001).

Results

Reliability and Bias

Reliability was determined using the Cronbach alpha measure of internal consistency computed by Quest. Results are shown in Table 2. The ideal value is

close to 1. For all assessments the value was acceptably high, and each of the assessments provided reliable data.

Table 2
Reliability of Each Assessment

Reliability measure	IAS	CRPC	MAT
Cronbach alpha	0.82	0.74	0.79

The infit mean square (IMSQ) measure provides a measure of fit to the Rasch model. Acceptable values lie between 0.77 and 1.3, and this can be used to indicate the fit of different groups to the model. Where there is misfit detected, there may be test bias present. The IMSQ measures for each variable are shown in Table 3. The values indicated that there was no detectable bias across gender or course.

Table 3
Infit Mean Square of Each Assessment

Group	IAS	CRPC	MAT
Male	0.95	0.87	0.99
Female	0.93	0.77	0.97
MT420	1.10	0.88	1.14
MT421	0.92	0.77	0.98
MT422	0.91	0.80	0.90
Overall	0.94	0.81	0.97

To determine the extent to which the assessment data were related, correlations were undertaken among the three variables. These are shown in Table 4, corrected for attenuation, so that measurement error was minimised. They indicate that correlations among the three variables obtained from the different assessments are only moderate. The highest value, R = 0.61, is for the association between the test of mathematics skills, MAT, and the test of mathematical problem solving, CRPC. These low correlations were unexpected, especially the low association of the performance task, IAS, with each of the other variables.

Table 4
Correlations Among Variables

Variable	IAS	CPRC	MAT
IAS	1.00	0.45	0.37
CPRC		1.00	0.61
MAT			1.00

Comparisons were also undertaken among the outcomes on all three assessments across boys and girls. Figure 1 shows the box plots of students' performances on the three variables by gender. The scale is in logits, produced by the Rasch analysis.

Examination of the box plots suggested that girls achieved slightly better than boys on the performance task (IAS), but the reverse was found for the mathematics skills test (MAT). There appeared to be little difference on the mathematical problem-solving test.

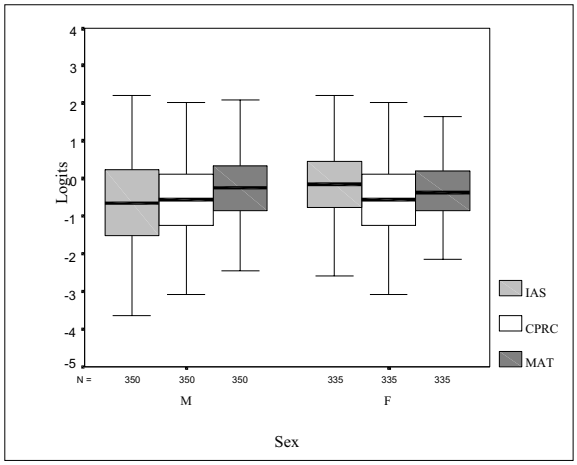

Figure 1. Student ability by sex across three variables.

The differences were confirmed by comparing the outcomes for each group using an independent samples t-test. The results are shown in Table 5. Findings for both IAS and MAT indicated that the null hypothesis that the means for males and females were equal was rejected in these two instances.

Table 5
Independent Samples T-Test Males and Females on All Variables

Test	Category	Mean	df	t	p
IAS	Male	-0.78	683	-4.75	0.00**
	Female	-0.29			
CRPC	Male	-0.57	683	-1.23	0.22
	Female	-0.46			
MAT	Male	-0.18	683	2.02	0.04*
	Female	-0.32			

* $p < 0.05$. ** $p < 0.001$.

Similar analyses were undertaken for each variable by mathematics course. The box plots are shown in Figure 2. There were some noticeable differences across the three variables. As expected, there was increasing achievement on all three variables as the mathematics course went from lowest to highest. The overall spread of ability was more contracted in the mathematical skills test, MAT, than in either of the tests based on cognitive developmental perspectives (CRPC and IAS). One noticeable feature was the wide spread of ability shown by students in the lowest mathematics course, MT420, on the IAS variable. The top students in this course were achieving almost as well as the top students in the highest course, MT422, on this test. The performance task, IAS, also allowed students in the two lower level courses to demonstrate higher levels of achievement than did either of the two other assessments. On both CRPC and MAT, only students in the top-level course, MT422, achieved at the highest levels.

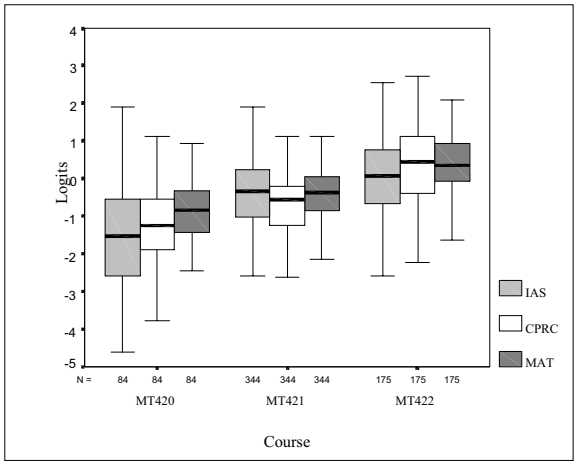

Figure 2. Student ability by mathematics course across three variables.

ANOVA analyses indicated that the null hypotheses that the mean abilities between each of the mathematics courses were equal could be rejected for all variables ($p<0/001$). This was not surprising since the courses were organised according to ability level. The implications of these findings are discussed below.

Discussion

The three different assessments all provided good reliability and showed no bias across gender or mathematics course. The variables produced by the various assessments were, however, less highly related than might have been expected of assessments all measuring the same construct. There were significant differences between males and females on the performance task and the test of mathematics skills. There were also significant differences on each assessment among students undertaking different mathematics courses.

The performance task asked students to undertake practical investigations and to explain these in writing. Females achieved better on this task than males, but the converse was true of the mathematics skills test. This finding is in keeping with other research that suggested that girls performed better in classroom based assessment whereas boys achieved more highly in examinations (e.g., Brew, Leder & Rowley, 1999).

The finding that achievement on all assessments was related to the mathematics course studied was not surprising. However, the different spread of achievement across the three variables was unexpected. In particular, the performance task, IAS, appeared to provide opportunities for students in the lower level courses to demonstrate their ability. Two inter-related aspects may be involved here, namely nature of the construct being assessed and the nature of the assessment process.

The performance task, In a Spin, was specifically designed to address higher order thinking within a numeracy context. It included practical tasks, making and testing spinners, as well as theoretical ones, such as determining the outcomes from spinning two spinners, but did not expect a particular format or approach to the activities. Students were free to draw on any skills and knowledge that they had at their disposal to answer the questions. In contrast, the test of mathematics skills, MAT, specifically targeted particular mathematical knowledge appropriate for Year 10 students, including symbol manipulation, trigonometry, and use of complex measurement formulae. Unless students had had opportunities to learn the appropriate content, they

would be unlikely to be able to achieve at the highest levels. The mathematical problem-solving test, CRPC, appeared to fall somewhere in between these two. It also had thinking skills as an underlying construct but, for achievement at the highest levels, required the application of specific aspects of mathematics. The correlation between the MAT and CRPC variables also reinforced this view. However, the performance task, IAS, and the mathematical problem-solving test, CRPC, allowed students to demonstrate higher achievement than did the mathematics skills test, MAT.

The second aspect was the nature of the assessment process. The performance task, IAS, took place in normal classrooms, where students could talk, discuss, and work together, although each student did have to produce an individual product. The other two assessments were undertaken under test conditions. More activity was required by the task –using concrete materials and writing explanations. It has been shown in earlier studies (Callingham & Griffin, 1998) that lower ability students preferred an open-ended assessment format, whereas higher ability students liked multiple-choice formats. This too may have affected the outcomes.

There are a number of implications arising from these findings. Firstly, the relatively low correlations among the variables, suggests that inferences about students' understanding, knowledge, and skills were less transferable across assessments than might have been expected. A student with high levels of mathematical skills might not apply these well in an open-ended practical situation, and it would be unwise to surmise this. Secondly, although no test bias was detected, the dissimilar formats appeared to allow different groups to achieve. Some students with lower levels of mathematical skills appeared to perform better on an open-ended, practical application task. In itself, this may not be surprising, but the level of thinking that lower ability students could demonstrate indicates that current methods of course choice may not be based on cognitive processes.

The Cronbach alpha figures indicated that all the assessments, including the teacher-judged performance task, In a Spin, were reliable. The lack of bias in the assessments indicates that no group of students was systematically disadvantaged or advantaged by the assessment, including the teacher-judged task. The multiple-choice test of mathematics skills did not allow students to demonstrate the highest levels of ability, and the problem-solving test only allowed students in the top mathematics course to reach the highest levels. This suggests that teacher-judged tasks may be useful in providing reliable information about students' higher order thinking, especially if these students have poor mathematics skills.

Finally, that differences between males and females on different test forms continues to be observed is worth noting. Clearly there is still a need for gender issues to be considered when assessment is planned.

These findings about the same students' performances on different assessments at one point in time reinforce the need for both the underlying construct and the nature of the assessment to be considered at the assessment development stage. In addition, they indicate that multiple approaches to assessment are likely to provide a more rounded picture of a student's mathematical ability.

References

Adams, R.J., & Khoo, S. (1996). *Quest: The interactive test analysis system, Version 2.1*. Melbourne: Australian Council for Educational Research.

Assessment Research Centre. (2001). *Survey of mathematics skills. Trial 1 and trial 2*. Melbourne: Author.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

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

Brew, C., Leder G. & Rowley, G. (1999). Mathematics teachers and the VCE: Broadening the educational landscape. In J. M. Truran & K. M. Truran (Eds.), *Making the difference* (Proceedings of the 22nd annual conference of the Mathematics Education Research Group of Australasia, Adelaide, pp. 98-104). Sydney: MERGA.

Callingham, R., & Griffin, P. (1998, August) *Issues encountered in using an investigative task as part of a large-scale assessment program.* Paper presented at the British Education Research Association Annual Conference, Belfast.

Callingham, R., & Griffin, P. (2001). Beyond the basics: Improving indigenous students' numeracy. In J. Bobis, B. Perry & M. Mitchelmore (Eds.), *Numeracy and beyond*. (Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia, Sydney, pp. 122-129). Sydney: MERGA.

Caygill, R., & Eley, L. (2001, September). Evidence about the effects of assessment task format on student achievement. Paper presented at the annual conference of the British Educational Research Association, Leeds.

Clarke, D., & Clarke, B. (1999). Developing and using rich assessment tasks: Some models, some lessons. In K. Baldwin & J. Roberts (Eds.) *Mathematics, the next millennium*. (Proceedings of the 17th biennial conference of the Australian Association of Mathematics Teachers, Adelaide, pp. 266 – 273). Adelaide: AAMT.

Collis, K.F., & Romberg, T.A. (1992). *Collis-Romberg mathematical problem solving profiles.* Melbourne, VIC: Australian Council for Educational Research.

Cureton, E.E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25*, 326-346.

Department of Education & Catholic Education Office. (1998). *Tasmanian numeracy and literacy assessment and monitoring program.* Hobart: Author.

Doig, B., & Hunting, R. (1995, November). *Interpreting student response using clinically-based mathematics assessment.* Paper presented at the 25th annual conference of the Australian Association for Research in Education, Hobart.

Grouws, D.A., & Meier, S.L. (1992). Teaching and assessment relationships in mathematics instruction. In G. Leder (Ed.) *Assessment and learning of mathematics*. (Pp. 83-107). Melbourne: Australian Council for Educational Research.

Newmann F. M. & Associates, (1996). *Authentic achievement: restructuring schools for intellectual quality*. San Francisco: Jossey-Bass.

Pegg, J. (2002). Assessment in mathematics: A developmental approach. In J. Royer (Ed.), *Mathematical cognition*. (Pp. 227 – 259). Greenwich, CT.:Information Age Publishing.

Shepard, L.A. (2000). *The role of classroom assessment in teaching and learning. CSE Technical Report 517.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Third International Mathematics and Science Study (TIMSS). (1995). *Released item sets*. Retrieved September 14 2001 from http://timss.bc.edu/TIMSS1/Items.html