Developing Task Specific Criteria: A Preliminary Report

Dawn Bartlett
NSW Department of Education and Training
<dawn.bartlett@det.nsw.edu.au>

A case study methodology was utilised to document the processes involved in the development and marking of extended response tasks within a systemwide numeracy assessment. One of the tasks included in the final test is used to exemplify how numeracy and measurement requirements can be satisfied within this process. This paper is a preliminary report of a research project being undertaken as part of the development of the pilot of the NSW Department of Education and Training Secondary Numeracy Assessment Program.

Secondary Numeracy Assessment Program (SNAP)

The *Secondary Numeracy Assessment Program* (SNAP) is a systemwide assessment of students' numeracy skills at the beginning of Year 7. It consists of three sections: short responses, connected responses, and extended responses. The focus of this research project is the development of scoring rubrics for the two extended response tasks that would be professionally marked.

Students are given twenty minutes to work on each of the extended response tasks. Each student is provided with a plastic instrument sheet, which features a ruler, a protractor, a 2 mm grid and a 1 cm grid. Triangles, as shown in Figure 1, is one of the tasks in the 2000 pilot.

Extended Response Task 2: Triangles

Part A	Part B
Using your plastic instrument sheet, draw three different types of triangles. Name the types of triangles you have drawn. Explain how your triangles are different from each other. Use appropriate mathematical terms in your work when naming and explaining.	If one triangle has a large perimeter than another triangle, will it always have a larger area? Answer Yes or No and give reasons for your answer. To do this task, you need to: 1. Draw some triangles 2. Show their areas and perimeters 3. Use the correct units of measurement for the areas and perimeters. 4. Use the plastic instrument sheet to help you.

Figure 1. Final version of Triangles task.

Teachers, parents and students are provided with detailed reports of each student's achievement on the numeracy aspects assessed in the SNAP paper. One of the reports provided for teachers indicates how students achieved on each item and task in the test. Teachers are expected to use this information to support them in planning relevant learning experiences for the students in their classes and school.

Students' Work

Good tasks can provide information about the extent of the student's knowledge being assessed and give information about a number of mathematical ideas, and the extent to which the student has integrated them and is able to use them in new situations (Webb & Briars, 1990). Open-ended questions generate a variety of mathematically valid responses, which differ only in the quality of understanding displayed (Clarke, Clarke, & Lovitt, 1990).

With richer responses from students given open-ended questions, teachers need to develop skills in interpreting evidence and using the results (Bryant & Driscoll, 1998). Assessing student's written work on a problem could be done using analytic scoring, focused holistic scoring or general impression scoring. Analytic scoring involves the use of a scale to assign points for certain aspects thus it becomes possible to identify specific areas of strength and weakness. Focussed holistic scoring enables a numerical score to be assigned to the total solution based on specific criteria. It is most appropriate when requiring a general rating of the processes used and where reliability of the scoring procedure is important. General impression scoring is based on implicit criteria to rate a total solution numerically (Charles, Lester, & O'Daffer, 1987).

Open-ended tasks have the potential to enrich assessments as they reveal different information from closed tasks. Teachers are able to apply scoring rubrics to open-ended tasks. The scoring allows the tasks and student performance to be evaluated (Sullivan, 1999). Well-designed tasks with appropriate scoring rubrics can elicit information about students' deeper understanding of mathematical concepts and provide reliable and valid data for measurement purposes (Callingham, 1999).

Stephens and Sullivan (1997) conducted research on the viability of using open tasks for system-wide assessment. Teachers were able to consistently apply a scoring rubric to the tasks, which allows the tasks and student performance to be evaluated and independent judgements to be validly made.

In analysing students' responses to performance assessment tasks, Peressini and Bassett (1996) found that inadequacies in test construction are more explicit in open-ended tasks than traditional mathematical tasks, enabling the teacher to enhance the quality of the open-ended tasks through refinement based on evidence from student responses. Unexpected responses to open-ended tasks can be examined to ascertain whether the unexpected response is due to the construction of the task.

The Research Project

How can extended response tasks be scored to provide appropriate, reliable and valid information to teachers on the numeracy skills of each student?

A three-phase research project is being developed. Phase One involves the development of the tasks and accompanying scoring rubrics. A single embedded case study provided an appropriate methodology for research into how extended response tasks could be scored. The phenomenon under study, the development of the means of scoring extended response tasks, is not readily distinguishable from its context, which is the processes of developing the Secondary Numeracy Assessment Program (Yin, 1994). The case is one complete cycle of development for the pilot in 2000 involving selecting two tasks from over twenty tasks. The researcher is an active participant and leader within the team developing the assessment. Triangulation of data has been achieved through collecting data on the tasks and development process from different periods of time (trialling, retrialling, and review) and from different persons involved in the process. Peer examination, involving asking colleagues to comment upon findings as they emerge occurs at each stage of the development process.

The Rasch model has been used to inform decisions about the appropriateness of tasks and criteria during the first phase.

The Rasch model is different from other models including latent models. It is primarily designed to aid test constructors in the process of constructing measurement variables. The primary difference is

> that it is a model of intent. That is, it has evolved from a theory or theoretical position. As such, it has come before any data are collected. The data are then collected and compared to the model, which is a mathematical rendition of the intention. If the data do not accord with the model, then this is evidence that the data do not reflect the intention, and further qualitative work is required. (Tognolini, 1996, p. 29)

Phase Two is the marking of the tasks using these scoring rubrics. Markers in both schools and a central location will use the marking procedures to mark the students' work. An evaluation involving survey and focus group discussions will be conducted with the school based marking process. The central marking operation involves monitoring through analysis of markers' responses. QUEST and RUMM are the two software programs used for Rasch analysis of dichotomous and polytomous data within this project.

Phase Three is the reporting of the students' results to schools and parents. The reports for schools will include a table indicating the scores for each student on each criteria. Both qualitative and quantitative research methodologies will be used to investigate the use and implications of the report data for the extended response tasks.

The Process for Developing Tasks and Criteria

Developing, panelling and trialling initial tasks. Contract and team writers developed more than twenty five tasks. A panel of numeracy, curriculum and equity experts critiqued the tasks in terms of quality and appropriateness for a Year 7 numeracy assessment. Tasks were subsequently modified or rejected. Sixteen of the accepted tasks were trialled. About two hundred students, representing various demographic and ability groups, attempted each task.

The Triangles task, as shown in Figure 2, is located within the mathematics key learning area. This task is a good question (Clarke, Sullivan, & Spandel, 1992) that requires more than recall or repetition of a fact or procedure to complete the task, is open-ended and has the possibility for the student to learn about mathematics in doing the task. The skills required to attempt the task are identified in the *Mathematics K-6* (1989) and *Outcomes and Indicators for Mathematics K-6* (1998) as being appropriate for students in Stage 3 (Years 5 and 6). It is expected that all students would be able to start the task as some of the required skills are appropriate for Stage 1(Years 1 and 2).

Extended Response Task 2: Triangles – Part A	*Extended Response Task 2: Triangles – Part B*
Draw at least 3 different types of triangles. Label your triangles. Explain how your triangles are different from each other. *Remember* • *Use appropriate mathematical terms in your labels and explanations.*	If a triangle has a large perimeter, will it have a larger area? *Remember:* • *Show how you worked out your answer, including calculations.* • *You could use the triangles in part (a) to help you answer this question.* • *You may need to draw some more triangles to test whether your answer is correct.*

Figure 2. Trial version of Triangles task.

Observing students completing tasks. Personnel with expertise in numeracy and/or assessing open-ended tasks observed students attempting the tasks. These observations were conveyed to the development team. Some tasks were rejected based on these observations. In trialling, all students attempted the Triangles task and only a few students asked for clarification about the task.

Developing draft criteria. Eight tasks were selected for developing draft criteria. The numeracy skills needed to complete each task were identified. Student work samples were examined for evidence of those skills. While students would have had to use certain skills to complete the task, there was not always evidence in the student's work that the skill had been used, the extent to which the skill had been used, or how the skill had been used. There were other skills that were evident only in some students' work, even though other students had provided work that indicated a higher level of numeracy achievement. Draft criteria could be developed for five tasks.

An extensive list of skills (Figure 3) could be generated for the Triangles task and evidence of these skills was found in the student work samples.

Drawing triangles	Checking answer to problem	Solving problems
Labelling triangles	Using calculations to solve problem	Use terminology
Measuring angles	Give reasons for solution	Use centimetres
Measuring area of triangles	Use square centimetres	Identifying angles
Measuring perimeter of triangles	Using drawings to solve problem	Definitions of triangles
Use a ruler to measure length	Know the units for length are cm	Properties of triangles
Use a grid to measure area	Know the units for area are cm2	

Figure 3. Skills for Triangle task.

Marking criteria were developed which markers were able to use consistently. Criteria A1 and B7 are shown in Table 1.

Table 1

Triangles: Trial Marking Criteria for Criteria A1 and B7

Part	Criteria	Scoring	Descriptor
A1	Construction of triangles	0	No triangles drawn
		1	1 triangle drawn matches label
		2	2 different triangles match labels
		3	3 different triangles match labels
B7	Uses drawings	0	None drawn
		1	Drawings with errors, irrelevant or lacking information
		2	Drawings accurate, relevant, appropriate information

Marking trialled tasks using draft marking criteria. Four tasks were marked. Only one group of four markers marked each task. Each group had a team leader who provided training on the criteria and practice marking to check consistency of application of criteria. The markers worked with one team leader referring any questionable work samples to the team leader. At the conclusion of marking, the markers provided comments and recommendations on the wording and appropriateness of the tasks, the appropriateness and quality of the criteria, and the marking procedures utilised. One task was rejected as a result of observations during marking.

Analysing data of student achievement on tasks. The results for three tasks were analysed using QUEST and RUMM analysis. The data for each criterion for the three tasks was examined for reverse thresholds, inappropriate item fit, spread of thresholds and mean ability of groups for each score of each criteria. The analysis of the marking data revealed that the majority of the criteria satisfied the measurement requirements. For Triangles, Table 2, Figure 4 and Figure 5 show that criteria B7 satisfied these requirements while criteria A1 has reverse thresholds and an unacceptable fit. Suggestions were made for the other criteria, which indicated that a valid set of criteria could be developed for this task. Numeracy and measurement aspects were considered in selecting the two tasks for the pilot.

Table 2

Percentages of Students Achieving each Score for Criteria A1 and B7

Criteria	Score 0	Score 1	Score 2	Score 3
Criteria A1	32%	18%	18%	32%
Criteria B7	37%	38%	25%	0%

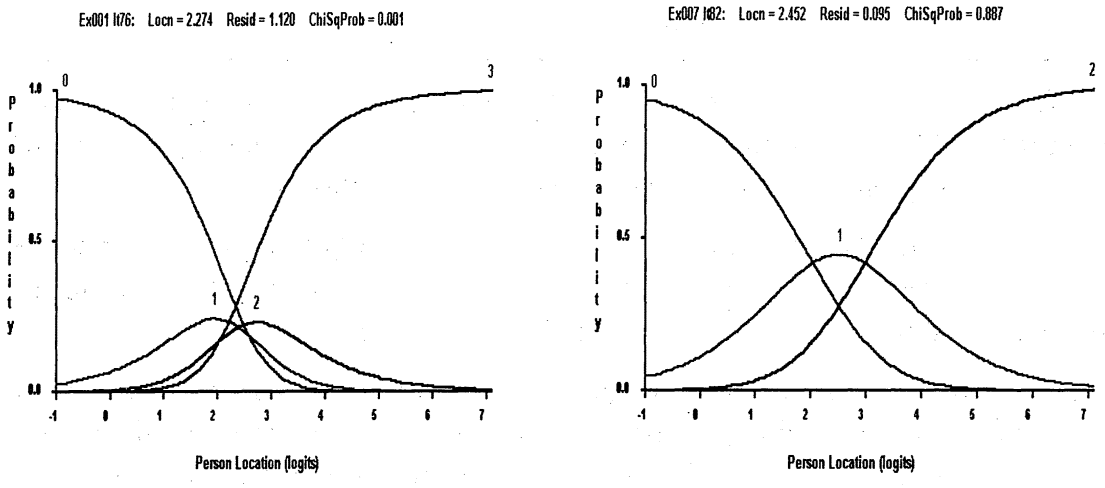

Figure 4. Category probability curves for trial criteria A1 and B7.

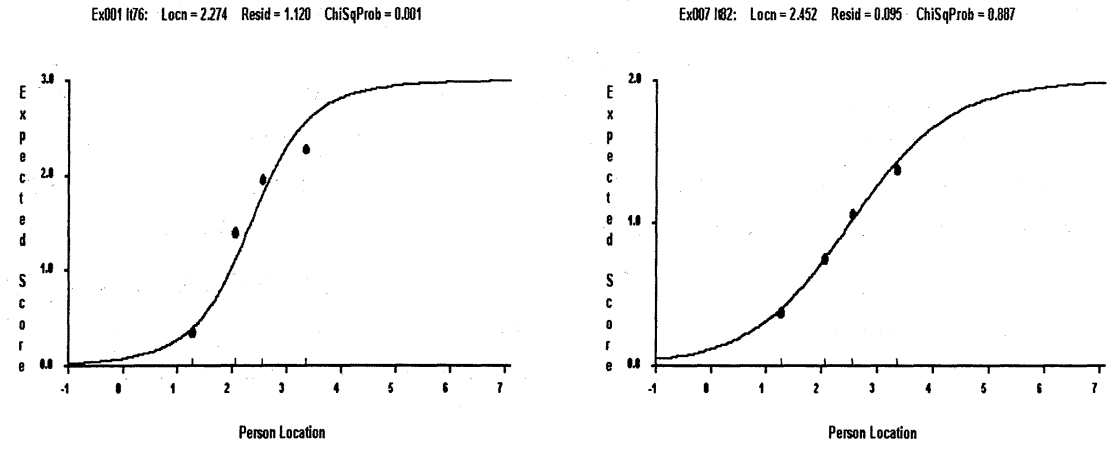

Figure 5. Item characteristic curve for trial criteria A1 and B7.

Refining tasks and criteria. Changes were made to the layout and wording of the tasks based on the information from the trialling and analysis of the trialling data. The retrialled (and final) version of the Triangle task is shown in Figure 1. Revised criteria were developed for marking the retrial. Table 3 shows the redefined A1 criterion.

Table 3
Triangles: Retrial Marking Criteria for Criteria A1 and A4

Part	Criteria	Scoring	Descriptor
A 1	Drawing different triangles	0	No triangles or only 1 triangle drawn or draws triangles which are similar
		1	Draws 2 triangles which are not similar or draws two different triangular shapes
		2	Draws 3 dissimilar triangles
A4	Uses knowledge of shape definitions	0	No evidence of shape definitions
		1	One shape definition correct
		2	Two shape definitions correct
		3	Three shape definitions correct

Retrialling tasks. Both tasks were retrialled with more than one hundred and sixty students in four schools. The four schools were diverse in their student demographics and student ability based on past performance of Year 7 students at the schools.

Marking of retrialled tasks. Again, each task was marked with only one group of markers using similar procedures to the first trialling process. Extensive notes were taken during the marking process for use in developing the detailed marking procedures.

Analysing retrial data and student work samples. As for the first trialling process the marking data was analysed using QUEST and RUMM. The criteria were modified and recoding or rescoring occurred with the criteria. Acceptable criteria have an appropriate spread of thresholds, no reverse thresholds, the mean ability increasing with score within the criterion, and appropriate fit for the criterion.

Analysis of the retrial data indicated problems with some of the criteria for Part A of the Triangles task. Figure 6 shows the reverse thresholds for criteria A4. These criteria were redefined and all the work samples rescored. Discussion of the work samples led to further modifications and rescoring. The revised criterion for A4 is shown in Table 4. The data was analysed to reveal acceptable statistics, as shown in Figure 6, for the criteria. The validity of the criteria for a numeracy assessment was paramount in making the modifications to the criteria. The skills being assessed by the criteria needed to be appropriate numeracy skills and the criteria needed to be valid instruments to measure the numeracy skills indicated.

Developing draft marking procedures. The agreed criteria were elaborated with detailed explanations, examples and references to actual student work samples. A draft was developed and critiqued through the use of contentious work samples and questions of clarification.

Table 4

Triangles: final version of marking criteria for criterion A4

Part	Criteria	Scoring	Descriptor
A4	Explanation of different types of triangles	0	Incorrect descriptions/definitions
		1	Writes 1 or 2 descriptions/definitions of geometrical properties of their triangles which are correct
		2	Writes or indicates 3 descriptions/definitions of geometrical properties of their triangles which are correct and consistent with both the labels and drawings

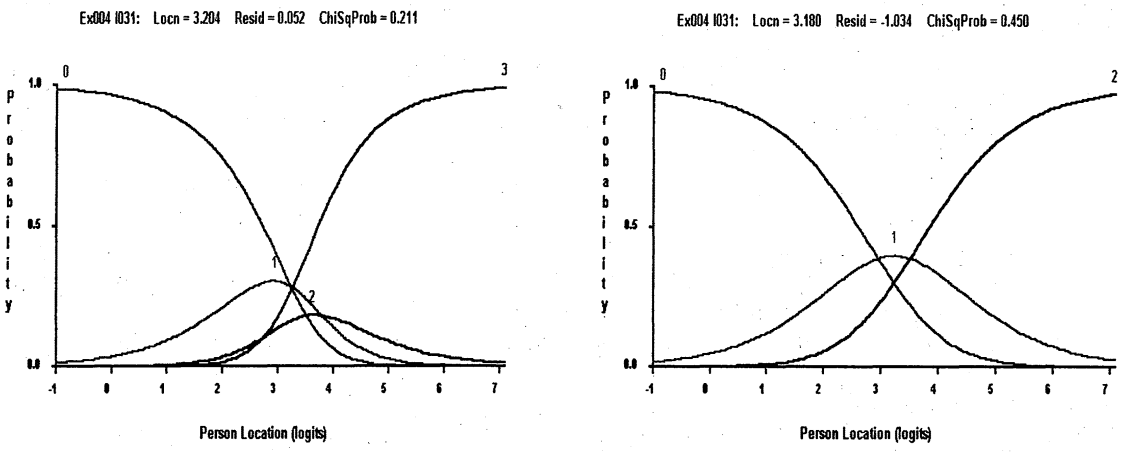

Figure 6. Characteristic probability curves for two different versions of criteria A4.

Refining marking procedures. Experts in numeracy, literacy, curriculum, assessment and measurement reviewed the marking procedures. The document is currently being finalised with modifications based on three review cycles.

Marking of extended response tasks. The marking procedures will be used to mark the extended response tasks in the pilot of SNAP. Teams of markers, both in selected schools and at a central location will be trained. The marking process will include vertical and horizontal audits for reliability and consistency of marking.

Recommendations for Future

The data from this case study indicated that the process up to the first trialling phase was useful in selecting and refining the tasks. As indicated in other research, the students' responses provided information that could be used to enhance the quality of the tasks. The process of retrialling enabled the enhanced tasks to be used for refining the criteria.

Further research is needed on the aspects of task design which impact on student achievement. Research on the differences in the teaching and assessing contexts would provide valuable information on why student achievement in assessments may not be consistent with teachers' expectations. The data from this project revealed a difference between the perceived difficulty of these tasks and the actual difficulty of the tasks. There were indications that

teachers may actively assist students when doing numeracy assessment tasks in the classroom. In assessment situations students are expected to work without this assistance. This has possible implications for classroom-based assessments of what students are able to do.

References

Bryant, D., & Driscoll, M. (1998). *Exploring Classroom Assessment in Mathematics: A Guide for Professional Development*. Reston, VA: National Council of Teachers of Mathematics.

Callingham, R. (1999). Developing performance assessment tasks in mathematics: A case study. In J. Truran & K. Truran (Eds.), *Making the Difference*. (Proceedings of the 22nd annual conference of the Mathematics Education Research Group of Australasia, pp. 135-142). Sydney: MERGA.

Charles, R., Lester, F., & O'Daffer, P. (1987). How to Evaluate Progress in Problem Solving. Reston, VA: National Council of Teachers of Mathematics.

Clarke, D., Clarke, D., & Lovitt, C. (1990). Changes in mathematics teaching call for assessment alternatives. In T. Cooney & C. Hirsch (Eds.), *Teaching and Learning in Mathematics in the 1990's : 1990 Yearbook*. (pp. 118-129). Reston, VA: National Council of Teachers of Mathematics.

Clarke, D., Sullivan, P., & Spandel, U. (1992). Student response characteristics to open-ended tasks in mathematical and other academic contexts. In B. Southwell, B. Perry & K. Owens (Eds.), *Space - The first and final frontier*. (Proceedings of the15th annual conference of Mathematics Education Research Group of Australasia, pp. 209-221). Sydney: MERGA.

NSW Board of Studies (1998). *Mathematics K-6 Outcomes and Indicators*. Sydney: NSW Board Studies.

NSW Department of Education (1989). *Mathematics K-6*. Sydney: NSW Department of Education.

Peressini, D., & Bassett, J. (1996). Mathematical Communication in Students' Responses to a Performance-Assessment Task. In P.C. Elliott & M.J. Kenney (Eds.), *Communication in Mathematics, 1996 NCTM Yearbook*. (pp. 146-158). Reston, VA: National council of Teachers of Mathematics.

Stephens, M., & Sullivan, P. (1997). Developing Tasks to assess mathematical performance. In F. Biddulph & K. Carr (Eds.), *People in Mathematics Education*. (Proceedings of the20th annual conference of Mathematics Education Group of Australasia, pp. 470-476). Waikato: MERGA

Sullivan, P. (1999). Seeking a rationale for particular classroom tasks and activity. In J. Truran & K. Truran (Eds.), *Making the Difference*. (Proceedings of the 22nd annual conference of Mathematics Education Research Group of Australasia, pp. 15-28). Sydney: MERGA

Tognolini, J. (1996) Rasch modelling: Advantages and Limitations. In National Meeting on Assessment and Reporting.

Webb, N. & Briars, D. (1990.) Assessment in mathematics classrooms, K-8. In T. Cooney & C. Hirsch (Eds.), *Teaching and Learning in Mathematics in the 1990's: 1990 Yearbook*. (pp. 108-117). Reston, VA: National Council of Teachers of Mathematics.

Yin, R. (1994). *Case Study Research: Design and Methods*. Second Edition. Thousand Oaks: Sage Publications.[1]

[1]*Acknowledgment*. This case study would not have been possible without the support and involvement of the staff in the NSW Department of Education and Training School Assessment and Reporting Unit. During the case study they were involved in developing the materials and analysing the student data.