Developing tasks to assess mathematical performance

Max Stephens and Victorian Board of Studies

Peter Sullivan[i] Australian Catholic University

There is a need to find ways to assess performance of students in mathematics which also provide some assurance to the community of the quality of the work of the students. As part of a larger project which sought to develop assessment tasks, we found that tasks could be used to assess the level of students' mathematical knowledge, that teachers could make holistic judgments on students' responses. We also argue that it is necessary to use a range of tasks to assess mathematical performance on particular aspects of the curriculum.

Introduction

In many English speaking countries there is considerable attention to assessment of students in mathematics, and particularly in using such assessment as measures of accountability within school systems. Within Australia, most States and Territories have implemented large-scale data collection on student performance in mathematics through compulsory tests.

The model used in these large scale assessments has been typically that of the standard achievement test, with response types often being constrained by a requirement for electronic correction. However, there is growing dissatisfaction with this form of assessment. Some criticisms have centred on limitations of this form of testing mathematical understanding (e.g. Clements & Ellerton, 1995). Other criticisms have been directed at the ineffectiveness of large-scale testing in supporting curriculum planning and informing instruction (Clarke, 1996).

It seems that there is a need to identify assessment tasks which satisfy community demands for accountability and which are also educationally meaningful and valid. One approach which has the potential to do this is called performance assessment. This paper describes some characteristics of performance assessments and presents some data from a project which sought to identify tasks which could support teachers in assessing mathematical performance.

Performance assessments

Performance assessments may take the form of practical tasks, interviews with students, project work or specially designed written tasks. Judicious use of performance assessment tasks taken over time is intended to assist teachers to make overall judgments of whether a student is performing at, below, or beyond a given level. This focus on performance assessment is a direct outcome of recent work in North America and other countries (see for example National Council of Teachers of Mathematics, 1989; Hong Kong Department of Education, 1996) on curriculum frameworks and assessment standards.

Performance assessment is generally linked to a statement of a standard derived from content frameworks developed by state or national agencies or by professional associations, such as the Curriculum and Evaluation Standards for School Mathematics (National Council of Teachers of Mathematics, 1989). A program of performance assessment typically contains two further elements: performance descriptions and annotated work samples. In the words of the New Standards Project (1995, p2), the former are "descriptions of what students should know and the ways they should demonstrate that knowledge and the skills they have acquired", and the latter are examples of student work chosen "for their capacity to illustrate the meaning of the performance descriptions together with commentary that shows how the performance descriptions are reflected in the work sample".

The nature of performance assessment is well summarized by the following description:

> Performance assessment requires that students actively respond to questions. Success in producing a response depends on reasoning, problem solving, and communication skills that students bring to bear on the problems that are posed. In contrast, success in traditional tests depends substantially on the ability of students to recall knowledge. In the course of performance assessment, students are

> required to apply their knowledge to situations to closely resemble "real life" circumstances. Performance assessment tasks require that students respond by completing one or more complex tasks or activities, such as developing an argument, performing an analysis, formulating a policy, or solving a problem that included several component problems. (Wisconsin Center for Education Research, 1993, p. 7)

One of the key aspects of performance assessments is the ways in which they differ from standard achievement tests. In the use of standard achievement tests, links between assessment and teaching are difficult to identify. The progressive nature of assessment over time marks a further key distinction, as does the inclusion of a wide range of types of assessment tasks, ranging from practical tasks, projects, observations of student performance, oral interviews, discussions, portfolios of selected samples of students' work, as well as specific assessment tasks (Hong Kong Department of Education, 1996). Further, the importance of teacher judgment in the interpretation of assessment marks an obvious break with standard achievement testing. What is also needed is a way to satisfy the legitimate community interest in the standard at which the students are performing.

Defining standards of performance

Proponents of global tests presumably justify their use by claiming that the tests measure performance against a clear standard, they are free from teacher bias, and are reliable. The inference is that teacher assessments, while being useful for informing teaching, are not robust in terms of objectivity or consistency. Performance assessment have the potential to address such concerns directly. One of the advantages of performance assessments is that they provide teachers with tools to define, measure and report the standard of performance of the students. Sometimes referred to as benchmarks, targets or indicators of achievement, outcomes, or pointers, these standards describe the knowledge, skills and understanding that students are intended to exhibit at a given stage or level of schooling. Different levels of students' achievement are described by reference to such standards.

For example, in the Victorian Curriculum and Standards Framework - Mathematics (CSF) (Victorian Board of Studies, 1995), the curriculum is described in six major strands: Space; Number; Measurement; Chance and data; Algebra, and Mathematical tools and procedures. Each strand is then further elaborated by Substrands. The Number Strand, for example, is considered in terms of the Substrands*: Number, Counting and Numeration; Mental Computation and Estimation; Written Computation; Applying Numbers; and Number Patterns and Relationships.* For each substrand, seven levels of performance are given for reporting student achievement over the eleven years of schooling covered.

For example, in the *Applying Number* Substrand levels 3 and 4 read as follows:

> Level 3: Representing and solving problems involving up to two of the four operations, including situations involving whole numbers and simple fractions of objects, money, and other quantities within the student's experience. (Level 3 describes what the majority of students might be expected to achieve by the end of Year 4.)
>
> Level 4: Choosing and using appropriate operations to solve problems which may involve whole and fractional numbers and more than one operation (Whole-number multipliers and divisors only). (Level 4 describes what the majority of students might be expected to achieve by the end of Year 6.)

In the CSF these statements are illustrated by indicators of performance sometimes referred to as outcomes. The statements also describe an increasing complexity of performance from one level to the next. A similar approach is taken in the Targeted Oriented Curriculum (Hong Kong Department of Education, 1994), whereas the NCTM Curriculum and Evaluation Standards for School Mathematics (1989) uses a broader framework for describing standards and the curriculum. Thirteen standards are used involving mathematical content and thinking for Grades K - 4, fourteen standards for Grades 5 - 8, and the same number for Grades 9 -12. The fourteen standards relating to Evaluation are intended to "help teachers better understand what students know and make meaningful instructional decisions ... As the curriculum changes, so must the tests" (p. 189).

Performance standards have several important features which distinguish them from behavioural objectives. First, performance standards are intended to reflect key emphases in the development of the mathematics curriculum, and their achievement is to be taken as

evidence of student progress through that curriculum. Behavioural objectives were developed to indicate a sequence of learning, usually in tiny chunks, and are not necessarily linked to a curriculum or program of teaching. Performance standards reflect a tendency to view learning and achievement in more complex terms:

> The tasks must have multiple objectives and require higher order levels of thinking than is demanded by most pencil-and-paper examinations. (Foster, 1991, p. 35)

In the New Standards Project (1995), for example, where for the High School years performance descriptions are elaborated in eight areas of the mathematics curriculum, a single assessment task might be designed to elicit performances from more than one area. Embodying more complex statements of achievement, performance assessments have a different level of aggregation from items on a standard achievement test. Several performance assessments, considered together, can provide a guide to the level of achievement of a particular student or group of students in a given area of mathematics. In other words, teachers make judgments on student responses to rich tasks designed to elicit different levels of achievement. If it can be established that teachers do make such judgments validly and reliably, then performance assessment will serve the same accountability requirements as mandatory tests, as well as providing better instructional support.

Preparing tasks for assessing mathematical performance

This is the report of some data which were collected as part of the Exemplary Assessment Materials Project which produced assessment tasks to support individual teachers in monitoring performance of their students in mathematics (Beesey, Clarke, Clarke, Stephens & Sullivan, 1997)[ii]. The project prepared 200 tasks which teachers could use to assess student achievement of specific aspects of content as specified within curriculum frameworks, which give insights into how students approach mathematical tasks, and which provide opportunities for students to give high level responses, and which are rich teaching activities as well. The tasks were a mix of extended, medium length or short tasks drawn from the Number, Measurement, Algebra and Chance and Data strands of the curriculum.

Data were collected on a variety of aspects of the assessment tasks in order to establish whether the tasks were suitable and whether the recommended scoring was manageable. In particular, the data collection sought to determine whether:

- the tasks were able to be completed by students at the levels at which they were posed;
- the tasks were able to be compared with each other;
- teachers could make valid and reliable judgments on the performance of the students at the tasks as a whole and whether responses of the students were consistent across the tasks.

Task administration

All tasks were piloted to adjust wording and difficulty level. For this data collection, the tasks were grouped together into components of particular Substrands of the mathematics curriculum. For example, the tasks which addressed Written computation were completed by the same students.

The tasks within each CSF Strand were completed by at least two complete classes of students. The responses to the tasks were scored by their teacher and analysis done on that scoring. For one sub-strand in each strand of the curriculum a member of the research team scored one sub-strand of tasks independent of the scoring done by the teacher.

Scoring the tasks

The tasks generally focused on a particular aspect of mathematical content. Each tasks consisted of several related questions, often in increasing order of difficulty, set in a particular context. The project took the view that providing a direct numerical scoring (e.g. 3 out of 5) was not an appropriate way for recording information on student responses to the tasks. A more global approach was used to assist teachers to form on-

balance judgments of the student's performance on the task as a whole. Each of the tasks were scored on a range from 1 to 4 using the following general rubric:

Score	Summary/Description (Used for Content Strands: Space, Number, Measurement, Chance & Data, Algebra)	Mathematical Tools and Procedures Equivalences
Goes Beyond	Fully accomplishes the task, but uses methods and/or makes interpretations significantly beyond those specified for this level.	Strategies /mathematical communication /reasoning significantly beyond those specified for this level.
4	Task accomplished. Central mathematical ideas clearly demonstrated and understood.	Appropriate plan. Clear communication of strategies and mathematics used.
3	Substantial progress towards completing the task; indicative of understanding of relevant knowledge, concepts and skills, but some key ideas may be missing.	Some evidence of planning; some communication of strategies and mathematics used.
2	Attempt at the task makes some progress; partial but limited grasp of the central mathematical ideas; reeals gaps in knowledge, understanding and/or skills.	Little evidence of effective strategies/communication/ reasoning.
1	Little progress or understanding evident.	Ineffective strategies.

In addition to the rubric, teachers were give about a particular task, including comments on the particular mathematical focus, guidelines for administration, and samples of student work.

Results

Data were collected and analyzed across a range of strands and Substrands of the mathematics curriculum. Just one aspect of these data are presented to illustrated the style of tasks, the data collected, and form of analysis. Data are presented here on one substrand of the Number in the CSF, *Number Counting and Numeration*, for Level 4, which is applicable to upper primary students (ages 11/12). The specific descriptor for this substrand at the level was:

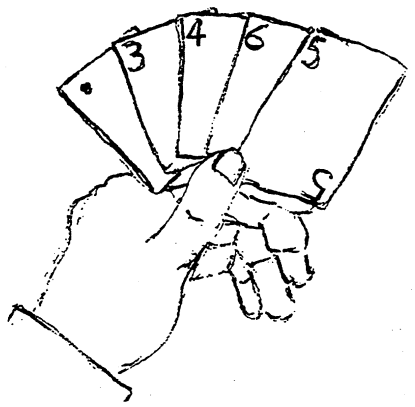
Counting, ordering, estimating and describing with large numbers and common and decimal fractions. Using place value (thousandths to millions and beyond)

The following is one of the tasks from this substrand presented to show the format in which the tasks were posed.

Five cards

You are playing a game using cards with numbers and decimal points. When it is your turn, you can choose to use all five of your cards, or just 4, or just 3, or just 2 or just 1.

You have these 5 cards:



1. Using these cards, what is the smallest possible number you can make?
2. Using these cards, what is the largest possible number you can make?
3. What numbers can you make that are between 3.5 and 5? (Give as many answers as you can. Remember that you can use as few or as many cards as you wish.)
4. Circle the number which is closest to 5.4?

   5.3 5.46 5.6

   5.364 5 5.463

5. One of your friends asks you to explain the best way to decide which number is closest to 5.4. How would you explain how to work out the number closest to 5.4?

Note that the task includes some closed items including some which require analytical thinking, an open-ended item, and an item which requires communication of thinking processes.

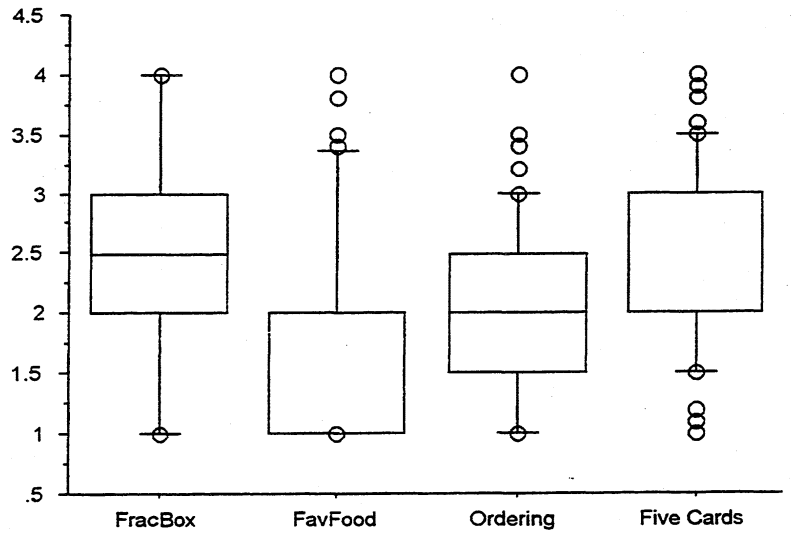
Three other tasks from this substrand are also part of the data presented. *Fraction Boxes* sought information on different forms of fraction representations, *Favourite Foods* was about the application of fractions, decimals and percentage to a menu context, and *Ordering* compared numbers written in different forms, such as common fractions, decimals and percentages.
The tasks were completed by 4 classes of grade 5/6 students. All tasks were scored on a scale from 1 to 4. A score of 1 indicates the student showed no evidence of appreciation of the task requirements, while a score of 4 suggests a high level response with a clear understanding of the task requirements. Fractional scores are possible.
Table 1 is a description of the scores given to the tasks by the teachers of the classes.

Table 1: Summary of task scores

	Fraction Boxes	**Favourite Foods**	**Ordering**	**Five Cards**
No. of Responses	132	133	127	137
Mean	2.4	1.9	2.1	2.4
Std. Deviation	1.0	0.9	0.8	0.8

Our first goal of was to determine whether the tasks could be completed by students at the level at which the content standard is specified. If the mean scoring for a task was well above 3 to well below 2, it would suggest that the task was either too easy or too difficult. It seems that although *Favourite Foods* is slightly more difficult for the students, the tasks seem to be within an acceptable range.
While the above summary gives some indication of the difficulty and spread, it is not so clear how the tasks can be compared with each other. Table 2 uses an alternate format, in which box plots compare the task scores (in this case with the teacher and researcher scores combined). Each box shows the middle 50% of the scores, and the whiskers show the spread from the 10th to 90th percentile. Basically it is a graphical representation of the spread of scores.

Table 2: Comparison of task results (teacher and researcher combined).



These box plots provide a clear indication of the comparative difficulty of the tasks and the spread of responses. The median of *Favourite Foods* is 2 and *Five Cards* is 3. The responses of the students to *Five Cards* and *Fraction Boxes* were scored higher than those of the other tasks, with a greater spread of the scores for *Fraction Boxes*. It seems that, overall, these tasks were able to be completed by students at this year level, with at least some students giving comprehensive responses. The tasks are able to be compared with each other.
The project also sought to explore whether teachers could make consistent judgments on the performance of students on the tasks. The researchers reported on the ease with which the teachers scored the tasks, and interpreted the rubric and additional information. As

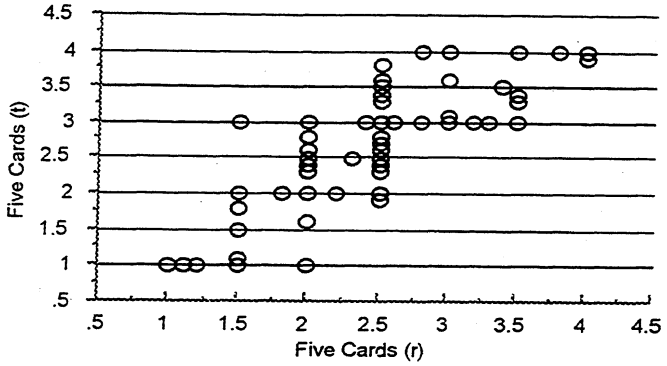
another measure of the validity of the judgments of the teachers, for a selected substrand a researcher scored each task independently of the teacher scoring. Table 3 presents the correlation between the teacher and researcher scores for each of the tasks discussed.

Table3: Correlation of teacher and researcher scores

Task	Correlation Value	Number of Scores
Fractions in Boxes	.93	132
Favourite Foods	.87	133
Ordering	.90	127
Five Cards	.80	137

All of these are highly significant but it is difficult to interpret the educational significance of such figures. To illustrate the strength of the relationships, Table 4 is a scattergram containing teacher and researcher scores on the *Five Cards* task, the lowest correlation figure.

Table 4: Scattergram of teacher and researcher scores for Five Cards.



This is clearly indicative of a strong relationship between the two scorings. There are quite few instances of marked divergence between the scores. On balance, the degree of association suggests that teachers and researchers applied the rubric to the individual student responses in similar ways.

A further aspect of the data collection was to explore whether students give consistent responses across the tasks. It is difficult to find a way to analyze, present or interpret the range of responses given by individuals to the related tasks. Basically we are interested in the extent to which the scores given to individual students over the four tasks are spread. To present this information simply, the number of students whose scores ranged from 0 to 1 were counted, and likewise for 1 to 2, and 2 to 3. Table 5 shows the number of students with scores within these ranges.

Table 5: Range of scores of individuals over the four tasks

Range of individual's scores	Frequency
$0 < x \leq 1$	57
$1 < x \leq 2$	54
$2 < x \leq 3$	34

The first impression is that these scores are spread quite widely. One possible inference is that the tasks do not provide reliable measures of achievement.

Another interpretation is possible. Even though the tasks the *Number, Counting and Numeration* substrand, they cover a broad range of outcomes drawn from ordering of decimals, ordering of fractions, comparison of different representation of fractions, and interpretation of the meaning of fractions. It may be that this spread in scores illustrates both the need to use a range of tasks to assess achievement even within a narrow component of the curriculum and also the need to exercise care in forming judgments from responses to a single question.

Summary

This has been some concern expressed about the educational value of mandatory testing. It seems useful to explore whether systems can be devised which have positive instructional implications and which also provide a measure of the overall standards of achievement. Data are presented here from a project which sought to compile assessment tasks and to develop a process by which teachers assess students' achievement on mathematical tasks by making holistic judgments on performance which are practical, meaningful, and valid. It is suggested that teachers are able to apply the scoring rubric to the tasks, that the scoring allows the tasks and student performance to be evaluated, and that independent judgments can be validly made. It appears that there is variation in the responses of individual students. This makes the need for performance assessments more critical. More research need to be undertaken into the type of tasks which provide the best source material for teachers' assessment of the mathematical performance of their students, and how best to enhance the quality and consistency of teacher judgments in using performance assessments.

References

Beesey, C., Clarke, B., Clarke, D., Stephens, M., & Sullivan, P. (1997). Exemplary assessment materials project - Mathematics. Carlton: Victorian Board of Studies.

Board of Studies (1995). *Curriculum and standards framework - mathematics.* Carlton, Victoria: Author.

Clarke, D. (1996). Assessment. In A. Bishop, K. Clements, C. Kietel, J. Kilpatrick, C. Laborde (Ed.) *International handbook of mathematics education* (pp 327-371). Dordrecht: Kluwer Academic Press.

Ellerton, N., & Clements, M.A. (1995). Challenging the effectiveness of pencil and paper tests in mathematics. In J. Wakefield & L. Velardi, *Celebrating mathematics learning.* Melbourne: Mathematical Association of Victoria.

Education Department of Hong Kong (1994). *General introduction to target oriented curriculum.* Hong Kong: Author.

Education Department of Hong Kong (1996). *Target oriented curriculum assessment guidelines: Key Stage 1 (Primary 1-3).* Hong Kong: Author.

Foster, J. D. (1991). The role of accountability in Kentucky's Educational Reform Act of 1990. *Educational Leadership, 48*(5), 34-36.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

New Standards Project (1995). *Performance standards: English Language Arts, Mathematics, Science, Applied Learning.* National Center on Education and the Economy, Pittsburgh, PA: Author

Wisconsin Center for Education Research (1993). Wisconsin performance assessment development project: Annual report. School of Education, University of Wisconsin-Madison: Author.

---

[i] The paper is a report of an aspect of a collaborative project whose members included Cathy Beesey, Barbara Clarke and Doug Clarke as well as the authors

[ii] The project was funded by National Professional Development Program, through the Victorian Board of Studies