Validity is the extent to which the inferences drawn from scores on a test or assessment can be justified empirically and theoretically. Establishing the validity of an assessment can be seen as an ongoing process of judgment, using different forms of evidence for substantiation. Traditionally validity has been defined within an evidential framework of three interrelated aspects: criterion, content and construct validity. Messick’s (1989) integrated view of validity led to inferences drawn from the information gathered from the assessment being pivotal to establishing validity and extended the understanding of validity to the consequences of test or assessment use, and led to a consideration of construct validity as the over-arching evidential basis.

Performance assessment, in which students provide some form of product or performance, typically takes place under less standardised conditions than traditional test forms. It has been suggested that validity standards should be different for this form of assessment, relating more directly to the specific performance. Students’ observed performances on a complex performance task in a numeracy context, as well as on tests of mathematics and mathematical problem solving, were analysed using Rasch modelling techniques. Six criteria suggested by Messick for construct validity evidence, were applied to the data. Findings indicated that the performance assessment information validity was high, and provided data useful for multiple purposes. The implications of these findings for performance assessment are discussed.
Background

In recent years, in Australian school systems there has been a shift away from traditional subject areas, such as mathematics and English, towards an emphasis on cross-curriculum competencies, such as numeracy and literacy. In some instances, this has been accompanied by changes in the nature of the enacted curriculum, for example, the New Basics project in Queensland (Education Queensland, 2000) and the Essential Learnings in Tasmania (Department of Education, 2002). These new conceptions of the curriculum are expected to provide opportunities for students to exercise higher order thinking skills, such as analysing, justifying and conjecturing.

These shifts provide a challenge for assessment developers. In Australia, numeracy is described in terms of numerate behaviour:

To be numerate is to use mathematics effectively to meet the general demands of life at home, in paid work, and for participation in community and civic life.

In school education, numeracy is a fundamental component of learning, performance, discourse and critique across all areas of the curriculum. It involves the disposition to use, in context, a combination of:

- underpinning mathematical concepts and skills from across the discipline ( numerical, spatial, graphical, statistical and algebraic);
- mathematical thinking and strategies;
- general thinking skills; and
- grounded appreciation of context.

(Australian Association of Mathematics Teachers, 1997, p. 15.)

Although the importance of mathematical skills is acknowledged, the contextual and situated nature of numeracy is also of consequence. Students are expected to apply a range of mathematical skills and understandings across different contexts. In addition, there are demands to include higher order thinking through the application of mathematical thinking and general thinking skills. Assessment of numeracy thus needs to consider these various aspects.

Context-based assessment has been referred to as “authentic” (Newmann & Associates, 1996) or, sometimes, as requiring “rich tasks” (Education Queensland, 2000). More generally, the term performance assessment has been used to describe assessment of this kind, where students produce a performance or product about which judgments are made (Airasian, 1994). This type of assessment has been advocated to bring teaching and assessment closer together (Shepard, 2000) and to address outcomes relating to higher order thinking (Linn & Baker, 1996). Performance assessment, however, has been criticized as lacking reliability (Linn & Baker, 1996), being task specific (Linn & Burton, 1994), and for not always providing information about the target domain (Moschkovich, 1998). Although performance
assessment would appear to be appropriate for the assessment of numeracy, in that it can take account of context and provide opportunities for higher order thinking to be demonstrated, the technical and measurement problems are not trivial. In particular, the validity of the assessment information must be established if performance assessment is to provide a real alternative to “objective” assessment methods such as tests.

In this study, students undertook discrete performance assessment tasks in their own classrooms, with their usual teacher providing appropriate support and intervention. The assessments were not timed tests under standardised conditions. All tasks included the use of some form of concrete materials, and typically consisted of a set of eight to ten activities that required various kinds of response. Some required students to undertake a short investigation; others asked specific questions, requiring only a written answer. All tasks included activities that required students to explain or justify their thinking. Each of the component activities had a scoring rubric with an associated numerical code, and the set of activities, rubrics and codes made up a task. This paper presents findings from a validity study of one particular task, *In a Spin*, based on convergent and discriminant evidence (Cronbach & Meehl, 1955).

**Validity**

Validity could be summarised as being concerned with “…what the test measures and how well it does so” (Anastasi, 1988). Traditionally validity has been defined within an evidential framework of three interrelated aspects: criterion, content and construct validity (Cronbach, 1971). Criterion validity was, in essence, the extent to which the scores on the assessment could be related to specified criteria (Anastasi, 1988). Any assessment process can sample only a small part of the potential knowledge domain, but to have content validity a test or assessment should be representative of the wider domain addressed by the assessment (Anastasi, 1988; Cronbach & Meehl, 1955). This form of validity is generally established by experts in the domain. The term “construct validity” was introduced by Cronbach and Meehl (1955). They suggested that construct validation should be investigated “… whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured” (p. 2). This appears to be particularly pertinent to numeracy.

Messick (1989), however, suggested that validity as defined by the three traditional ideas of content, criterion and construct validity was fragmented and incomplete. Despite a long tradition of use of content and criterion validity, he argued that construct validity was over-arching, subsuming other forms of validity. He proposed six general standards for evidence of construct validity:

1. content, which should be relevant and representative;
2. substantive, that is underpinned by a rationale or theoretical basis;
3. structural, where the responses should be internally consistent across different parts of the test;
4. external, that is the scores should relate to other measures or background variables;
5. generalisable, both within and across populations and time; and
6. consequential, the social aspects of the interpretation of the test scores should be considered.
This integrated view of validity placed the meaning of the information gathered from the assessment at the centre, and extended the understanding of validity to the consequences of test use. Without construct validity, the assessment information had no evidential basis, and this compromised all other aspects.

Messick (1989) further suggested that there were two major threats to validity: construct under-representation and construct irrelevance. An assessment that does not include critical aspects of the construct lacks validity because it is too narrow in its perspective. The opposite is true of assessment that has construct irrelevance—it is too broad in its perspective and may include inappropriate aspects, such as test-wiseness or a method effect. These threats are of major importance when considering performance assessment. Performance assessment, however, is not immune to the requirement to establish construct validity, especially where the information will be used as a basis for decision making about future teaching or the students’ potential for further work (Messick, 1994). Any performance assessment that addressed numeracy, therefore, had to meet Messick’s (1989) six criteria for construct validity evidence.

**Establishing construct validity**

To establish construct validity two components are required: a conceptual model and a measurement model (Medical Outcomes Trust Scientific Advisory Committee, 1995). Although the description of numeracy (AAMT, 1997) provided some indication of how numerate behaviour could be considered, it did not provide a continuum of numeracy development against which assessment could be designed, and students’ growth measured. There was also the problem of performance task specificity (Linn & Burton, 1994) that was a threat to construct validity because of the limitation on domain coverage. Earlier use of performance assessments in numeracy contexts, however, indicated that they could be underpinned by a generalised continuum of competence (Callingham & Griffin, 2000). This continuum provided an overall holistic sequence in which a sequence of pattern → rule → generalisation was identified, and is shown in Table 1. This was used as a basis against which performance tasks were designed, and provided the conceptual model.
Table 1: Continuum of competence (Callingham & Griffin, 2000)

<table>
<thead>
<tr>
<th>Level</th>
<th>What the student does</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjecturing</td>
<td>At this level the student can suggest extensions of the original problem, or change the task parameters to create new situations by posing 'What if…’ questions. Problem solutions are expressed in appropriate symbolic or technical language. The student is ready to learn how to identify assumptions and develop further hypotheses.</td>
</tr>
<tr>
<td>Generalisation</td>
<td>At this level the student can express the generalisation of the problem solution in symbolic form, and apply it to new situations, justifying this by reference to the generalisation. The student has mastered the particular problem type and is ready to learn how to change the problem type to explore new ideas.</td>
</tr>
<tr>
<td>Rule extension</td>
<td>At this level the student can form a generalisation of the solution strategy and express this generalisation in words. The student is ready to learn the use of symbolic forms and technical language that will allow transfer of the generalisation to other settings.</td>
</tr>
<tr>
<td>Rule or process use</td>
<td>At this level the student’s own rule is applied to extensions of the initial task. The student can extend the solution obtained to a limited range of other tasks having a similar structure to the initial task, and is ready to learn how to form a generalisation that could be transferred to other settings.</td>
</tr>
<tr>
<td>Rule or process recognition</td>
<td>At this level the student recognises a rule underpinning the structure of the task and is ready to learn how to apply that rule consistently and extend the use of the rule. The rule is likely to be expressed in words or diagrams, using non-technical language that summarises the student’s own approach to the problem.</td>
</tr>
<tr>
<td>Pattern or structure use</td>
<td>At this level the student recognises the underlying principles in the structure of the task, and can apply these in a familiar setting, such as a straightforward extension of the initial task. The student is ready to learn how to identify a rule that links the repeating elements together.</td>
</tr>
<tr>
<td>Pattern or structure recognition</td>
<td>At this level the student recognises the repeating elements in the structure of the problem, and is ready to learn how to recognise the underlying principles. Problem solutions are likely to be presented as incomplete diagrams or oral explanations.</td>
</tr>
<tr>
<td>Element identification</td>
<td>At this level the student recognises individual elements of the task and is ready to learn how to combine these into a pattern or structure. Problem attempts are presented as single drawings or phrases that relate to one element only of the task.</td>
</tr>
<tr>
<td>No apparent understanding</td>
<td>At this level there is not enough information to describe the student’s work. If the task was attempted, it is likely that the student did not recognise the elements of the underlying structure of the task.</td>
</tr>
</tbody>
</table>

The measurement model was provided by an application of the Rasch (1960) model. It has been claimed that this model is a useful approach to establishing construct validity (Fisher, 1994). Rasch measurement models use the interaction between persons (cases) and items to place both persons and items on a single measurement scale. Construct validity can be examined by considering the fit to the model of both items and cases (Wright & Masters, 1982). If the items can be shown to be systematically and predictably related to each other along the variable this is confirmation that a single construct is being measured, and provides evidence of construct validity. In this study the Partial Credit Model (PCM) (Masters, 1982) was used as the measurement model for the examination of construct validity of a complex performance assessment task in a numeracy context.

The partial credit model is an extension of the Rasch simple logistic model (Rasch, 1980). In the PCM, the probability of a student responding in the $x^{th}$ category as opposed to the $x-1^{th}$ category is dependent on the difficulty of the $x^{th}$ level
This model has been used extensively with assessments based on developmental continua (Wilson, 1992, 1999).

The PCM can be expressed as:

\[
\frac{\pi_x}{\pi_{(x - 1)} + \pi_x} = \frac{\exp(\beta - \delta_x)}{1 + \exp(\beta - \delta_x)}
\]

where

- \(\pi_x\) is the probability of a person responding in category \(x\) (\(x = 1, 2, \ldots, m\)) of item \(i\);
- \(\beta\) is the person’s ability in the domain being measured by this set of items; and
- \(\delta_{ix}\) is the difficulty of the step threshold that governs the probability of the response occurring in category \(x\) rather than category \(x - 1\).

This model could be used to evaluate the validity of summarising performances on different activities in a single measure (Wilson, 1999). By using an estimate of “step difficulty” within each item in the assessment, the PCM locates a person on an underlying variable through a consideration of the number of steps that the person has made beyond the lowest level of performance. Using assessment items that addressed increasing competence on an underlying variable, and that were scored with more than a right/wrong response, the model could provide information about a student’s level of understanding on the target variable. The points at which the likelihood of a higher-level response became greater than that of a lower level response were called thresholds.

Wright and Masters (1982) provided a rationale for establishing validity based on Rasch models. The first step was to define the variable by separating the items. Items that clustered together could not provide sufficient information about the variable to allow interpretation. Since interpretation of the information was central to validity arguments, any threat to the interpretation also compromised the validity (Messick, 1989). However, items that spread out along the variable could provide a basis for interpretation of the underlying construct. Once items had been calibrated along the variable, these then needed to be interpreted in terms of the item writers’ intentions. Examination of model fit could provide information about how justifiable it was to measure the underlying construct with the particular set of items chosen (Wilson, 1992). Good fit to the model suggested that the items were measuring the same unidimensional construct, that is, the assessment had construct validity.

Once it had been established that the measured variable had direction and could define several statistically separate levels, the degree to which the variable met the intentions of the assessment developer needed to be determined. The extent to which the set of items defined the variable, in this study the continuum of competence, provided construct validity, that is measured a single, dominant and interpretable underlying trait.

Although this approach provided convergent evidence of construct validity, and could be used to establish construct under-representation, it was also necessary to consider divergent evidence, that is the extent to which the construct was related to other irrelevant constructs. The classic method is the multi-trait multi-method matrix (MTMM) (Cronbach & Meehl, 1955). Through a consideration of correlations among different traits assessed by different methods, hypotheses about method effects, or other construct irrelevancies may be tested.
A full MTMM requires a minimum of three traits and three methods. This was not possible in this study. However, a smaller 2x2 matrix could be constructed. The traits hypothesised were higher order thinking in numeracy contexts and mathematical skills, and the methods were teacher judgment and objective tests. Higher order thinking was measured using teacher judgment by the performance assessment task, *In a Spin*, and using an objective test by the Collis-Romberg Mathematical Problem Solving Profiles (Collis & Romberg, 1992). Mathematics skills were obtained by teacher judgment using a teacher judgment scale of mathematics ability and through a multiple-choice objective test of mathematics skills. The smaller matrix limited inferences drawn from this method, but it could nevertheless provide useful validity evidence.

**Method**

**Study design**

The study had two components. A cross-sectional design provided data about one performance task, *In a Spin*, and several related and unrelated variables. The purpose of this component was to establish convergent and discriminant evidence of construct validity. Students in Year 10 from 14 different government high schools in Tasmania undertook an assessment battery that included the performance task, *In a Spin*, a multiple choice test of mathematics skills, and an objective test of mathematical problem solving, the Collis Romberg Problem Solving Profiles (Collis & Romberg, 1992). Teachers were also asked to rate their students’ mathematics ability on a Likert scale instrument with 10 items to provide an additional measure based on teachers’ judgments. A summary of the assessment battery instruments and the variables produced is provided in Table 2.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Assessment method</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>In a Spin</em> performance assessment task</td>
<td>Teacher judged performance task</td>
<td>IAS</td>
</tr>
<tr>
<td>Maths teachers’ judgments of students’ maths ability</td>
<td>Teacher judgment on Likert scales</td>
<td>MTR</td>
</tr>
<tr>
<td>Collis-Romberg Mathematical Problem-Solving Profile</td>
<td>Student constructed response, correct/incorrect scoring</td>
<td>CRPC</td>
</tr>
<tr>
<td>Test of Mathematics Ability</td>
<td>Multiple choice test of mathematics skills, machine scored</td>
<td>MAT</td>
</tr>
</tbody>
</table>

A second component had a longitudinal focus. Data from five earlier performance tasks undertaken at different points in time, by two separate student samples, were linked using Rasch equating techniques (Kolen, 1999). The *In a Spin* task included four items in common with an earlier task, *Come In Spinner*, and this provided a means of linking previous administrations of five performance tasks that had all been designed against the same underpinning continuum. This component of the study was intended to provide substantive evidence of construct validity by showing that the same construct underpinned all tasks. A summary of the equating design is shown in Figure 1.
Table 1: Performance tasks

<table>
<thead>
<tr>
<th>Time</th>
<th>Student cohort</th>
<th>Long Table</th>
<th>Come in Spinner</th>
<th>Magic Beans</th>
<th>Keep Australia Beautiful</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Group P₁, Year 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group P₂, Year 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>Group Q, Year 9</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>Group Q, Year 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Summary of longitudinal study design showing links among groups

Performance task development

Performance tasks were developed to address the underlying continuum of competence in different numeracy settings. As a starting point for task development, a particular mathematical idea was targeted, and a teaching sequence outlined that would be similar to the order in which the underpinning concepts would be dealt with in a teaching unit. A series of activities was devised that addressed increasingly higher levels of understanding of the target mathematical concept, together with appropriate mathematical skills. At the same time, a “story shell” was developed to ground the concept within a context considered to be appropriate for the target group of students, and to bring the mathematical ideas into a numeracy framework. Concrete manipulative materials were included as part of each task to ensure that the tasks were genuine performance tasks rather than worksheet-based investigations. Each task had separate question and answer sheets, and was designed to be undertaken by students working in pairs, although each student had to complete an individual answer sheet. Each task had an estimated administration time of approximately 40 minutes (Callingham, 1999). Each task was named according to its story shell. An example of the performance task, In a Spin, is provided in Appendix A together with the associated scoring rubrics. These rubrics were analytical in nature, but were designed with reference to the underlying continuum of competence, although worded in such a way that teachers could use them with minimal training. They provided a set of scores on the different activities that made up the task that considered the quality of the students’ performances, and that could be scaled using the PCM. In this way, the tasks linked developmental assessment processes to a competency-based approach to assessment by operationalising the underlying continuum of competence in different numeracy contexts.

The sample

The student sample for the cross-sectional component of the study is shown in Table 3. All students were in Year 10, the final high school year in Tasmania, before moving to a senior secondary college for post-compulsory schooling. The sample size ranged from 45 in school H to 132 in school J, and was reasonably evenly split across males and females. The 14 schools were all government high schools, from all parts of the state and included schools in high and low socio-economic areas. Some teachers in these schools were involved in a project to improve indigenous students’ numeracy performance (INISSS; see Department of Education, Science and Technology (DEST), 2002) and the sample was opportunistic rather than random since the assessment was being used to provide evaluation data about the project (Callingham & Griffin, 2002). The tests for the cross-sectional component were all...
undertaken within a time period of three weeks (T3 in the longitudinal component), towards the end of the school year in November.

Table 3. Number of schools and students

<table>
<thead>
<tr>
<th>School</th>
<th>M</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>48</td>
<td>62</td>
<td>30</td>
<td>38</td>
<td>59</td>
<td>41</td>
<td>36</td>
<td>27</td>
<td>66</td>
<td>59</td>
<td>26</td>
<td>46</td>
<td>52</td>
<td>42</td>
<td>26</td>
<td>632</td>
</tr>
<tr>
<td>Percentage</td>
<td>7.6</td>
<td>9.8</td>
<td>4.7</td>
<td>6.0</td>
<td>9.3</td>
<td>6.5</td>
<td>5.7</td>
<td>4.3</td>
<td>10.4</td>
<td>9.3</td>
<td>4.1</td>
<td>7.3</td>
<td>8.2</td>
<td>6.6</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>No</td>
<td>43</td>
<td>35</td>
<td>35</td>
<td>29</td>
<td>62</td>
<td>33</td>
<td>49</td>
<td>18</td>
<td>66</td>
<td>64</td>
<td>32</td>
<td>52</td>
<td>47</td>
<td>46</td>
<td>611</td>
</tr>
<tr>
<td>Percentage</td>
<td>7.0</td>
<td>5.7</td>
<td>5.7</td>
<td>4.7</td>
<td>10.1</td>
<td>5.4</td>
<td>8.0</td>
<td>2.9</td>
<td>10.8</td>
<td>10.5</td>
<td>5.2</td>
<td>8.5</td>
<td>7.7</td>
<td>7.5</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Total Students</td>
<td>91</td>
<td>97</td>
<td>65</td>
<td>67</td>
<td>121</td>
<td>74</td>
<td>85</td>
<td>45</td>
<td>132</td>
<td>123</td>
<td>58</td>
<td>98</td>
<td>99</td>
<td>88</td>
<td>1243</td>
<td></td>
</tr>
<tr>
<td>Percentage of total</td>
<td>7.3</td>
<td>7.8</td>
<td>5.2</td>
<td>5.4</td>
<td>9.7</td>
<td>6.0</td>
<td>6.8</td>
<td>3.6</td>
<td>10.6</td>
<td>9.9</td>
<td>4.7</td>
<td>7.9</td>
<td>8.0</td>
<td>7.1</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Because of missing data across all components, through a process of listwise deletion a sub-sample of 440 students who had undertaken every component of the assessment battery, and for whom there were no missing data, was established. This sub-sample was used to calibrate the items using the Partial Credit Model (Masters, 1982) with Quest software (Adams & Khoo, 1996). The purpose of calibrating the items using a subset of students for whom there were no missing data was to provide more stable item estimates by eliminating the effects of items which had been undertaken by small groups of students. The item difficulties from this procedure were written into an anchor file which was then used in the scoring procedure with the full sample of 1243 students. Student ability estimates were obtained for the different components of the test battery.

Equating process

The equating processes used in this study to link the different performance assessments among the different student samples and time periods are summarised in Figure B. This shows a common person equating model used within time, T3, the cross-sectional component of the study since the same students undertook all tests at one point in time. Assessments given in time T3 were linked to assessments undertaken by a different group of students at time T1 through common items between *Come in Spinner* (CIS) and *In a Spin* (IAS) performance tasks. The assessments from time T1 were equated with those given in time T2 through the common performance task, *Long Table*. There were thus direct links between T3 and T1, and between T1 and T2, and an indirect link between T3 and T2. This established connectedness (Linacre, 1997) among all items and activities on each assessment undertaken.

The tasks were linked through common items using Quest (Adams & Khoo, 1996) by the following steps:

1. The group of students from T1 who had no missing data from *Come in Spinner* and *Long Table* was identified (393 students). This sub-sample was written into a data file that was used to calibrate these two tasks, from which an anchor file of CIS and LT was produced.
2. A second data file was prepared by merging the raw data from T1, T2 and T3 into a single data file that had the structure shown in Figure F above. This file contained 3412 individual responses.

3. In a single operation, the complete data set was calibrated and equated, anchored to the common items through the anchor file obtained from the T1 calibration of CIS and LT.

![Equating process to connect all performance assessment tasks](image)

**Results**

Results are presented for each component of the study.

**Cross-sectional component**

Overall item and case fit and reliability statistics were obtained for each variable. Fit values are considered adequate if the infit mean square value lies between 0.77 and 1.3 (Adams & Khoo, 1996; Keeves & Alagumalai, 1999), and these values were used throughout this study. These are summarised in Table 4. Fit and reliability was acceptable for both items and persons across all scales, suggesting that each variable consisted of a single, unidimensional dominant construct. The task variable, IAS, appeared to perform in much the same way as the objective measures.

<table>
<thead>
<tr>
<th>Scale</th>
<th>$R_I$ (a)</th>
<th>IMSQ (b)</th>
<th>zIMSQ (c)</th>
<th>$R_P$ (d)</th>
<th>PIMSQ (e)</th>
<th>zPIMSQ (f)</th>
<th>Cr alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAS</td>
<td>0.98</td>
<td>0.99</td>
<td>0.01</td>
<td>0.83</td>
<td>0.96</td>
<td>-0.12</td>
<td>0.82</td>
</tr>
<tr>
<td>MTR</td>
<td>0.93</td>
<td>1.00</td>
<td>-0.10</td>
<td>0.96</td>
<td>1.14</td>
<td>-0.13</td>
<td>0.97</td>
</tr>
<tr>
<td>CRPC</td>
<td>0.99</td>
<td>0.96</td>
<td>-0.22</td>
<td>0.72</td>
<td>0.98</td>
<td>-0.08</td>
<td>0.74</td>
</tr>
<tr>
<td>MAT</td>
<td>0.98</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.79</td>
<td>0.99</td>
<td>-0.12</td>
<td>0.79</td>
</tr>
</tbody>
</table>

(a) $R_I$ Item Separation Reliability  (d) $R_P$ Case (Person) Separation Reliability
(b) IMSQ – infit mean square (Items)  (e) PMSQ – infit mean square (persons)
(c) zIMSQ – infit t (Items)  (f) zPMSQ – infit t (persons)

Rater and school effects

Although the task did appear to be behaving as expected, it was possible that this was not so across all schools and all raters. Students in 13 different schools undertook the task, and 31 different teachers marked the responses. The technique of considering fit against ability across several groupings provided a strategy for identifying irregularities or differences among the groups (Postlethwaite, Ross, Griffin & Vinh, 2002).

The mean fit and mean student abilities across schools, bounded by ±1 standard error, are shown in Figure B. Acceptable limits of infit mean square values, 0.77 to 1.3 logits (Adams & Khoo, 1996; Keeves & Alagumalai, 1999), are shown by the horizontal dotted lines, and the display has been organised by mean student ability from highest to lowest.

In only two schools, school E and school L, was there any misfit, and both showed overfit, suggesting over-discrimination. In general, however, there was nothing to suggest that the task was measuring a different construct in different schools

![Figure 3: Fit and ability across schools](image)

The same analysis was undertaken across all teachers. Teachers were identified according to the school letter identifier, followed by a single number that
related to the initial identification code assigned to the teacher. Figure 4 shows the data organised by school and teacher, and sorted by student ability. No results are shown for school H as no teacher identification was allocated by the coordinator, for unexplained reasons, so that it was not possible to differentiate between classes. When individual schools were considered they tended to show different mean ability levels among individual classes. This most likely reflected the streamed, or semi-streamed structure that existed in most Tasmanian high schools by Year 10.

There was evidence of randomness in judgments in the data from three teachers, and two of these, F1 and F3, were in the same school. Two teachers, A3 and L1, showed systematic overfit. One of these, L1, was a teacher of English teaching “out of area” to a very small group of only six low ability students, and this may have contributed to the misfit. Overall, however, these results suggested that the scoring rubric was robust over the range of teachers and schools, and adequately defined the expected responses so that teachers could make reliable judgments. At school and teacher level, the performance task results appeared reliable and consistently fitted the model. It was noticeable that the ability estimates did not fall into school-based clusters—high and low ability students were spread across schools and teachers. It appeared that the same variable was being measured across schools and across teachers within schools regardless of the ability estimates obtained.

Overall, the findings supported the notion that the performance assessment task appeared to measure a single dominant construct in a consistent fashion across all schools and teachers. This inference was emphasised in Figure 4, where streaming within schools did not appear to have affected the rating patterns of teachers nor the interpretation of the underpinning variable.
Figure 4: Mean infit and student ability by school and teacher.
Rater and school effects of associated variables

For comparison, analyses by teacher and school were also undertaken for the MTR, MAT and CRPC variables. These are provided in Figures 5, 6 and 7, where the mean student ability and infit measures are shown bounded by ± 1 standard error of the mean.

The MTR variable (Figure 5) is presented by teacher sorted by ability within school. It showed considerable underfit, indicating that some teachers’ judgments of their students’ mathematics ability were affected by randomness. The streamed nature of the classes, however, was also evident, and the ability measures coincided with similar patterns for the same teachers within the IAS variable (Figure 4). It appeared that teachers made similar overall judgments about their students’ ability to that determined through the performance assessment task but that these judgments were more consistent when made against a scoring rubric, rather than made as an holistic judgment on a rating scale.

The CRPC variable (Figure 6) is presented sorted by ability, rather than within schools. It showed a number of instances of overfit, apparently randomly across ability and teachers. There were also some relatively large standard errors of measurement, most likely due to small numbers of students.

The MAT variable (Figure 7) is also presented sorted by ability. In contrast to the MTR and CRPC variables, the fit of the MAT variable was reasonable across all ability levels. There was, however, a trend towards underfit, or randomness, as ability fell. This was likely to be due to less able students guessing answers in the multiple-choice test.
Figure 5: Mean infit and student ability by teacher and school for MTR variable
Figure 6: Mean infit and student ability by teacher for CRPC
Figure 7: Mean infit and student ability by teacher for MAT
Multi-trait multi-method analysis

Although the evidence from Rasch modelling suggested that a single trait was being measured consistently, there were indications that the performance task, IAS, and problem solving test, CRPC, allowed for higher order thinking to a greater extent than the test of mathematics ability, MAT. Considering two traits, higher order thinking and mathematics skills, provided a convenient approach to developing a MTMM matrix, as described above.

The correlations among the variables are shown in Table 5, corrected for attenuation, organised according to trait and method. Values in the validity diagonal are italicised and reliabilities are shown in brackets. All correlations were significant at the 0.01 level.

<table>
<thead>
<tr>
<th></th>
<th>IAS</th>
<th>MTR</th>
<th>CRPC</th>
<th>MAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRJUDG IAS</td>
<td>(0.82)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTR</td>
<td>0.45</td>
<td>(0.97)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEST CRPC</td>
<td>0.57</td>
<td>0.56</td>
<td>(0.74)</td>
<td></td>
</tr>
<tr>
<td>MAT</td>
<td>0.46</td>
<td>0.56</td>
<td>0.78</td>
<td>(0.79)</td>
</tr>
</tbody>
</table>

In an ideal matrix, the values in the validity diagonal, shown in italics, which show correlations between the same trait measured by different methods, would be higher than values in all other cells in the same rows or columns except those showing the reliabilities. With respect to the task variable, IAS, this was true—it correlated more highly with the same trait, higher order thinking, measured in a different way (CRPC) with a correlation coefficient, R, of 0.57, than did different traits measured in the same way (IAS/MTR, R = 0.45), or the different traits measured in different ways (IAS/MAT, R = 0.46). This was not true of the other relationships, however. Mathematics ability measured by different methods (MTR/MAT, R = 0.56), had the same correlation as that between different traits measured by different methods, CRPC and MTR (R = 0.56), and was considerably less than the hetero-trait, mono-method value (CRPC/MAT, R = 0.78). This could imply a method effect. However, the relatively high correlation between the mathematical problem solving variable objectively measured (CRPC) and the mathematics ability measured through teacher judgment (MTR) suggests that the problem solving skills as measured by the objective test and mathematical skills are confounded. There does not seem to be a method effect associated with the performance task variable, IAS.

Longitudinal component

Data were available from five performance tasks designed to address the same underlying continuum of competence within different numeracy contexts. These were linked using the equating process described above.
Equating to earlier tasks

As a result of the equating process, all performance assessment tasks, as well as the MAT and CRPC variables were placed on a single scale. Summary mean item and person infit statistics are shown in Table 6. Fit and reliability figures appeared satisfactory for the overall scale. All assessments appeared to work together to define a single dominant construct.

Table 6: Fit and reliability statistics for equated scale of all tasks

<table>
<thead>
<tr>
<th></th>
<th>R_i (a)</th>
<th>IMSQ (b)</th>
<th>zIMSQ (c)</th>
<th>Rp (d)</th>
<th>PMSQ (e)</th>
<th>zPMSQ (f)</th>
<th>Cronbach α</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.99</td>
<td>0.99</td>
<td>-0.53</td>
<td>0.89</td>
<td>0.99</td>
<td>-0.10</td>
<td>0.81</td>
</tr>
</tbody>
</table>

(a) R_i Item Separation Reliability  
(b) IMSQ – infit mean square (Items)  
(c) zIMSQ – infit t (Items)  
(d) Rp Case (Person) Separation Reliability  
(e) PMSQ – infit mean square (persons)  
(f) zPMSQ – infit t (persons)

Interpretation of the variable

In order to interpret the single scale formed a process of “back-translation” (Griffin & Forwood, 1990) was used to compare the performances with the underlying intended construct, a continuum of competence operationalised in different numeracy contexts. A content and skills audit of the variable was undertaken based on the identification of item clusters along the variable. The process of interpretation was as follows:

1. The variable map produced by Quest was examined to establish an initial grouping of items along the variable.
2. The content and skills addressed by each item in a cluster were considered.
3. Common skills or understandings required were identified, and appropriate cut scores identified.
4. An overall summary of these skills and understanding was developed for each item cluster, based on the common themes identified.

This process provided an interpretation of the variable that could be compared, or “back-translated” onto the underlying construct. Nine item clusters were identified, as shown by the horizontal lines in Figure 8. Although the MAT and CRPC variables are included in this map, this is for comparison only and the item clusters were identified only from a consideration of the performance task variables. Link items between IAS and the earlier task, Come in Spinner, are shown in italics. The entire Long Table provided the additional link items.
<table>
<thead>
<tr>
<th>Case</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>CRPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
<tr>
<td>0.00</td>
<td>MB1</td>
<td>MB2</td>
<td>MB3</td>
<td>MB4</td>
<td>MB5</td>
<td>MB6</td>
<td>MB7</td>
<td>CRPC</td>
</tr>
</tbody>
</table>

Figure 8: Map of task variables from the equating process
When the distribution of items along the variable was considered, the performance tasks generally provided items that addressed a wide range of difficulty. Those from the earlier administration, T1, *Come in Spinner* (CIS) and *Magic Beans* (MB) had fewer items at the highest difficulty level, and more at the lower difficulty levels, whereas the later tasks *Keep Australia Beautiful* (KAB) and *In a Spin* (IAS) both provided a cluster of items at the high levels of difficulty. In contrast, the mathematics ability variable, MAT, tended to cluster around the middle levels of the variable. The objective problem solving test, CRPC provided some difficult items but lacked items at the easier end of the variable. The performance tasks appeared to provide for a wider range of performance than did either of the objective tasks.

The interpretation of the variable was related to the continuum of competence on which the tasks’ design was based, described above. The demands of the task activities within the identified levels did seem to allow for an interpretation consistent with the levels of the continuum. However, an additional level was included at the lowest level of difficulty. Skills required for a successful response to activities appearing at the lowest levels of difficulty relied on simple counting and repetition, whereas at the next level up there appeared to be greater demands on categorising or grouping elements together. The inclusion of five different tasks provided additional information on which to base the interpretation, and appeared to suggest an additional level. A summary of the interpretation of the skill demands of each item cluster level is provided in Table 7, explicated by the expectations of the scoring rubric of a typical item at each level of the identified hierarchy.
<table>
<thead>
<tr>
<th>Activity</th>
<th>Logit</th>
<th>Skill audit</th>
<th>Level</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB7.4</td>
<td>4.37</td>
<td>Explanation uses relationships and symbols to show a general solution applicable to any situation.</td>
<td>Conjecturing</td>
<td>At this level the student can suggest extensions of the original problem, or change the task parameters to create new situations by posing ‘What if…’ questions. Problem solutions are expressed and justified in appropriate symbolic or technical language. The student is ready to learn how to identify assumptions and develop further hypotheses.</td>
</tr>
<tr>
<td>LT8B.4</td>
<td>3.10</td>
<td>Detailed explanation of answer and arrangement includes symbols or equations that relate the table arrangement to the symbolic expression for the arrangement shown</td>
<td>Generalisation</td>
<td>At this level the student can express the generalisation of the problem solution in symbolic form, and apply it to new situations, justifying this by reference to the generalisation. The student has mastered the particular problem type and is ready to learn how to change the problem type to explore new ideas.</td>
</tr>
<tr>
<td>LT10.1</td>
<td>2.34</td>
<td>General relationship described that addresses only one element of the relationship e.g. the number of people seated depends on the length of the table</td>
<td>Rule extension</td>
<td>At this level the student can form a generalisation of the solution strategy and express this generalisation in words. The student is ready to learn the use of symbolic forms and technical language that will allow transfer of the generalisation to other settings.</td>
</tr>
<tr>
<td>KAB9.3</td>
<td>1.90</td>
<td>Recognises a range of similarities and differences but lists these rather than developing a coherent argument.</td>
<td>Rule or process use</td>
<td>At this level the student’s own rule is applied to extensions of the initial task. The student can extend the solution obtained to a limited range of other tasks having a similar structure to the initial task, and is ready to learn how to form a generalisation that could be transferred to other settings.</td>
</tr>
<tr>
<td>Activity</td>
<td>Logit</td>
<td>Skill audit</td>
<td>Level</td>
<td>Interpretation</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>LT9.2</td>
<td>1.50</td>
<td>At least two rules described that rely on single patterns only e.g. It goes up by 4’s this way and by 2’s if it is arranged differently (for appropriate table arrangements)</td>
<td>Rule or process recognition</td>
<td>At this level the student recognises a rule underpinning the structure of the task and is ready to learn how to apply that rule consistently and extend the use of the rule. The rule is likely to be expressed in words or diagrams, using non-technical language that summarises the student’s own approach to the problem.</td>
</tr>
<tr>
<td>KAB5B.2</td>
<td>0.64</td>
<td>Data set presented but no explanation or simple explanation based on context only e.g. lots of people hang out in the shade area.</td>
<td>Pattern or structure use</td>
<td>At this level the student recognises the underlying principles in the structure of the task, and can apply these in a familiar setting, such as a straightforward extension of the initial task. The student is ready to learn how to identify a rule that links the repeating elements together.</td>
</tr>
<tr>
<td>KAB4A.1</td>
<td>-0.19</td>
<td>Single format showing one kind of litter or one place only with no attempt to summarise more generally e.g. table about the canteen area</td>
<td>Pattern or structure recognition</td>
<td>At this level the student recognises the repeating elements in the structure of the problem, and is ready to learn how to recognise the underlying principles. Problem solutions are likely to be presented as incomplete diagrams using inefficient methods or oral explanations.</td>
</tr>
<tr>
<td>CIS2B.1/ IAS1.1</td>
<td>-1.44</td>
<td>Explanation based on 50:50 chance only e.g. Because it's a 50:50 chance</td>
<td>Element identification</td>
<td>At this level the student recognises classes of task elements and is ready to learn how to combine these into a pattern or structure. Problem attempts are presented as single drawings or phrases that relate to one element only of the task.</td>
</tr>
<tr>
<td>MB1</td>
<td>-2.68</td>
<td>Find the remaining totals for a, b and c (addition single digits).</td>
<td>Classification of elements</td>
<td>At this level the student recognises similarities in different elements and is ready to regroup these according to personal ideas.</td>
</tr>
</tbody>
</table>

This interpretation indicated that the various performance tasks could all be construed within the framework of the continuum of competence that was identified through the early tasks. These results suggested that the performance tasks were interpretable within a common framework. They also provided substantive evidence of construct validity.
Discussion

Messick’s (1989) integrated approach to construct validity provided a framework within which this proposition could be addressed, and it is the six criteria proposed for construct validity standards that will be used here. Each of these criteria was examined in the light of the evidence presented in the study. They were applied primarily to the data from the performance assessment task, *In a Spin*, but also, where appropriate to the other variables used to provide convergent and discriminant evidence for the validity of the assessment process.

**Content criterion**

Content validity has been an issue for performance assessment because the specificity of particular tasks has tended to limit the inferences drawn from the assessment to a particular situation (Linn & Burton, 1994).

The domain of numeracy as described in this study had multiple aspects: mathematical skills and knowledge, general thinking skills including higher order thinking, recognition of context, motivation to apply mathematics within a particular context, and ability to communicate (AAMT, 1997). The *In a Spin* task addressed a range of mathematical understanding, including chance and data concepts of sample space, variation and probability, the skills of graph reading, and calculating. It also demanded general thinking skills, such as planning and carrying out an investigation, interpretation of findings and the ability to generalise findings to new and unusual situations. The task allowed students to use concrete materials but also required them to explain their investigation, and to justify their thinking. The *In a Spin* task addressed key aspects of numeracy as understood in this study.

Within the limitations imposed by any practical assessment, this task had numeracy content validity, insofar as that can be described, since it drew on a range of mathematical and general thinking skills in an appropriate context for the students involved.

**Substantive criterion**

This criterion depends on the interpretation of item thresholds and their coherence in defining a single dominant construct. This relationship between the skills demanded at different item thresholds and the underlying construct is at the heart of item response modelling, and has led to the Rasch model being described as an instrument of construct validity (Fisher, 1994). However, as Messick (1989) pointed out, establishing this relationship alone provided convergent evidence only, and was not sufficient to establish construct validity. Discriminant validity was also needed—the target construct should not show undue association with constructs with which it was not theoretically coherent. As such, this aspect of Messick’s (1989) framework was the most demanding.

The item separation reliability and fit indices of the *In a Spin* task at the centre of this study were acceptable indicating that the task activities defined a single dominant scale that had direction and magnitude. However, IAS was one of a series of performance assessment tasks that had been developed to address a generalised underlying continuum. It was thus expected to show similar characteristics and properties to those demonstrated by other tasks in the series. A direct comparison was made by equating the tasks to IAS through a mixture of common person equating within an administration and common
items across time periods using item response modelling techniques. When the tasks were equated they showed a similar relationship between activity thresholds and the underlying variable. All tasks could be interpreted from the single perspective of the underlying continuum of competence, operationalised in different numeracy contexts. The empirical data obtained from the Rasch modelling of students’ responses to the performance tasks related closely to the theoretical underlying construct, thus providing evidence of construct validity.

**Structural criterion**

The behaviour of the test takers provided evidence of the internal structure of a test. This should be consistent, and was evaluated by a consideration of the fit of the students’ responses to the model, and the case separation reliability statistics, that is the person statistics as opposed to the item statistics, when item response models were used.

The case separation reliability of the performance task, *In a Spin*, was acceptable at 0.82, indicating that students’ responses to the task were consistent. In addition, misfit to the measurement model used was minimal across schools and teachers, suggesting that the task, *In a Spin*, had robust structural validity, behaving as expected across a range of situations.

**External criterion**

The external criterion of construct validity refers to the ways in which the test or assessment relates to other measures. The classic approach to this is the multi-trait multi-method matrix (MTMM) (Campbell & Fiske, 1959), which provides both convergent and discriminant evidence. In particular, this approach allows examination of construct irrelevance due to method effects.

The MTMM matrix analysis indicated moderate correlations between the task variable, IAS, and the other mathematically based variables, CRPC and MAT, indicating that the task was associated with related constructs of mathematical problem solving and mathematics skills. The mathematical problem solving variable, CRPC, was, however, relatively highly correlated with both measures of mathematics skills, MAT and MTR. This suggested that the problem solving ability was associated with mathematical skills to a greater extent than was the performance task variable, IAS. There did not appear to be a method effect associated with the task variable. Construct irrelevance due to method effects appeared to be minimal for the performance task, even though this was assessed through application of teachers’ judgments.

**Generalisability criterion**

This form of validity required the variable to behave consistently across different sub-groups of test takers, and over time, and has been a particular problem for performance assessment. Validity would have been compromised if the different tasks did not measure the same construct across all schools, students and teachers. The IAS variable did appear to behave consistently across different situations showing little misfit. In addition, equating to earlier tasks indicated that the same underlying variable was being measured in all situations. Since the tasks were administered to different groups of students at three different points in time, the variable underlying the tasks was stable over time, and could be used to track students’ development against the underlying continuum,
providing evidence of generalisability across time and place.

**Consequential validity**

The nature of this form of validity is somewhat different from the others. The basis is less statistical, and more judgmental, depending on values and appropriateness of test use.

The performance assessments were used to provide information for accountability purposes of the progress of the INISSS project. They were explicitly designed to match teaching approaches advocated by the project and were accepted by the project managers and teachers as suitable for this purpose (Callingham & Griffin, 2002). As a result the project was continued and extended to other schools. In this sense, the tasks had high consequential validity for all stakeholders.

Overall, for the *In a Spin* task, construct irrelevance was low. Although associated with mathematical skills and problem solving, this was to a moderate degree as would be expected of a construct realised in a numeracy context. There also did not appear to be a method effect associated with teachers’ judgments. Although the *In a Spin* task largely addressed mathematical content from the domain of the chance and data component of the curriculum, it also drew on generalised skills that required higher order thinking, as required by the description of numeracy used in this study (AAMT, 1997). It provided for a wide spread of ability along the variable, allowing all students to demonstrate their numerate behaviour. This suggests that the other major source of invalidity, construct under-representation, was also low. Since the same construct underpinned all of the five performance assessment tasks used in this study, the validity inferences could be applied not only to the *In a Spin* task, but also to the other tasks. These findings suggested that the performance assessment tasks met Messick’s (1989) criteria for construct validity, despite the unusual nature of the assessment process, which was based on teaching practice.

**Conclusion**

The performance tasks had good construct validity and reliability and provided for generalisation across contexts and time. The interpretive framework provided by the continuum of competence, rather than one grounded in a specific content area, provided a means of interpreting students’ performances that appeared to have construct validity evidenced by the application of Messick’s (1989) criteria. This construct did not relate to the content of mathematics, or numeracy, but to the students’ competence as it was demonstrated within a particular situation. This generalized construct allowed for the interpretation to be linked to a range of mathematical content and different numeracy contexts.

This approach to numeracy assessment provided for development, took account of formal mathematical knowledge and informal approaches, and different settings. Although numeracy competence was not defined, that is what standard was required for someone to be considered numerate, the means of doing this was inherent in the notion of a generalised framework that allowed for the operationalisation of the construct in different contexts. Overall, the process appears to have met some of the concerns about performance assessment, particularly in relation to validity standards.
References


