The Longitudinal/Cross-Sectional Study of the Impact of Teaching Mathematics using *Mathematics in Context* on Student Achievement

---

Monograph 4

2005

Measures of Student Performance

Thomas A. Romberg

David C. Webb

Lorene Folgert

University of Wisconsin-Madison

Mary C. Shafer

Northern Illinois University

WCER

Wisconsin Center for Education Research
School of Education - University of Wisconsin-Madison

The Longitudinal/Cross-Sectional Study of the Impact of Teaching Mathematics using *Mathematics in Context* on Student Achievement was carried out by the staff of the Wisconsin Center for Education Research, University of Wisconsin-Madison with the support of the National Science Foundation Grant No. REC 0553889. The analysis of the data gathered in this study was conducted by the staff of the Wisconsin Center for Education Research, University of Wisconsin-Madison and funded by the National Science Foundation Grant No. REC 0087511. Additional support for completing the monograph series was provided by Northern Illinois University.

Romberg, T. A., Webb, D. C., Shafer, M. C., & Folgert, L. (Editors). (2005). *Measures of student performance* (Longitudinal/Cross-Sectional Study of the Impact of Teaching Mathematics using *Mathematics in Context* on Student Achievement: Monograph 4). Madison, WI: University of Wisconsin-Madison.

Webb, D. C., Romberg, T. A., Shafer, M. C., & Wagner, L. (2005). Classroom Achievement. In T. A. Romberg, D. C. Webb, M. C. Shafer, & L. Folgert (Editors) *Measures of student performance* (Longitudinal/Cross-Sectional Study of the Impact of Teaching Mathematics using *Mathematics in Context* on Student Achievement: Monograph 4), 7-26. Madison, WI: University of Wisconsin-Madison.

Turner, R., & O'Connor, G. (2005). The development of a single scale for mapping progress in mathematical competence. In T. A. Romberg, D. C. Webb, M. C. Shafer, & L. Folgert (Editors) *Measures of student performance* (Longitudinal/Cross-Sectional Study of the Impact of Teaching Mathematics using *Mathematics in Context* on Student Achievement: Monograph 4), 27-66. Madison, WI: University of Wisconsin-Madison.

The L/CSS Monograph Series

This is the fourth of eight monographs derived from the NSF funded Longitudinal/Cross-Sectional Study of the impact of teaching mathematics using *Mathematics in Context* (National Center for Research in Mathematical Sciences Education & Freudenthal Institute, 1997-1998) on student achievement.

In 1992 the National Science Foundation funded several projects to develop new sets of instructional materials that reflected the reform vision of school mathematics espoused by the National Council of Teachers of Mathematics (NCTM, 1989). One of the funded projects was to the National Center for Research in Mathematical Sciences Education (NCRMSE) at the University of Wisconsin-Madison. The project was organized to develop a comprehensive mathematics curriculum for the grades 5-8 (NSF Grant No. ESI-9054928). Assisted by the staff of the Freudenthal Institute (FI) at the University of Utrecht in The Netherlands the *Mathematics in Context* (MiC) curriculum materials were created and field-tested prior to being published in 1997-1998 by Encyclopaedia Britannica.

In 1996, as the development of the MiC materials was nearing completion, a proposal was submitted to the National Science Foundation to investigate how teachers were changing their instructional practices in schools whose staffs were using *Mathematics in Context*, and how such changed practices affected student achievement. Two NSF grants were awarded to the University of Wisconsin-Madison: first, to conduct a three-year study of the impact of *Mathematics in Context* on student mathematical performance (NSF Grant No. REC-9553889); and second, to analyze the data gathered in that study (NSF Grant No. REC-0087511). This monograph series presents the rationale, development, and conduct of the study of the implementation of the MiC materials in classrooms across the nation, and the results portray the impact of the use of that curriculum on student achievement.

As students and teachers begin to use any of the new mathematics materials, district administrators, mathematics educators, teachers, parents, and funding agencies express cogent needs to demonstrate that the curricula have a positive impact on students' understanding of mathematics. They often want to know the bottom line—the results on measures of achievement that confirm improved student mathematical performance. However, while improved student performance is critical, we contend that just relying on outcome measures to judge the impact of a standards-based program is insufficient. In fact, it is not enough to consider student outcomes in the absence of the effects of the culture in which student learning is situated, the instruction students experience, and their opportunity to learn comprehensive mathematics content in depth and with understanding. The dynamic interplay of all these variables has an impact on student learning, and as such, these variables must be considered in making judgments about the impact of any standards-based curriculum.

This monograph series tells the complex story of the variations in how the MiC materials were used by teachers and students in classrooms that vary in location and ecological culture, and the impact of that variation on the achievement of their students. The story unfolds in eight monographs.

L/CSS Monograph Series on the Impact of Teaching *Mathematics in Context* on Student Achievement

Monograph 8 Implications and Conclusions

Chapter 1. Implementation Stories

Chapter 2. Insights about Implementing a Standards-Based Curriculum in Schools

Chapter 3. What we have Learned.

Introduction to Monograph 4

This monograph contains two chapters. Their purpose is to summarize the sources of student performance data for the variety of different studies carried out in this project. The measures of student performance were used in relationship to Research Question #1. What is the impact of the MiC instructional approach on student performance? In Chapter 1 the development of two assessment systems that included eight mathematics tests, two for each of four grades, is described. The Problem Solving system was developed via a subcontract with the Freudenthal Institute, University of Utrecht, The Netherlands. Then, in Chapter 2, the creation of a single scale that encompasses data from all eight of the mathematics tests is described. This scale was developed by researchers at the Australian Council for Educational Research.

CHAPTER 1: CLASSROOM ACHIEVEMENT

David C. Webb, Thomas A. Romberg, Mary Shafer, and Lesley Wagner

The purposes of the longitudinal/cross-sectional study of the impact of *Mathematics in Context* (MiC) on student performance were (a) to determine the mathematical knowledge, understanding, attitudes, and levels of student performance as a consequence of studying MiC for over three years; and (b) to compare student knowledge, understanding, attitudes, and levels of performance of students using MiC with those using conventional mathematics curricula. The research model for this study is described in Chapter 2 in Monograph 1). The outcome variables for this model are - knowledge and understanding, and application. For analytic purposes we assumed that variation in classroom achievement based on the measures developed for these variables can be captured in a single scale or progress map for classroom achievement (CA), aggregated by strand, grade, or total performance as described in Chapter 2 of this monograph.

Assessment Pyramid

The development of a single scale for the study involved the data gathered for the two outcome variables - knowledge and understanding, and application - in the original full model. The selection of assessment tasks used in developing these measures were based on an adaptation of a three dimensional model developed for the National Dutch option of TIMSS (Verhage & deLange, 1997), and illustrated by the pyramid shown in Figure 1-1. The first dimension involves domains of mathematics. For *Mathematics in Context* (MiC*)* there are four domains included in the curriculum - algebra, geometry, number, and statistics and probability. The second dimension reflects the pedagogical notion of "progressive formalization." This assumes that student responses to tasks progress from informal, to pre-formal, to formal responses as the tasks progress from simple to complex. The third dimension is about the level of thinking. Level I thinking involves performing specific calculations, solving a given equation, or reproducing memorized facts. Level II thinking requires integrating information, making connections within and across mathematical domains, and solving non-routine problems. Level III thinking involves recognizing and extracting the mathematics in a situation and using that mathematics to solve problems, analyze and develop models and strategies, and make mathematical arguments and generalizations. Thus, any assessment task can be located in the "assessment pyramid" in terms of mathematical domain, expected student response, and level of reasoning required. Furthermore, tasks involving Level III thinking often will involve concepts or procedures from more than one domain (as illustrated by the lack of lines separating domains at this level).

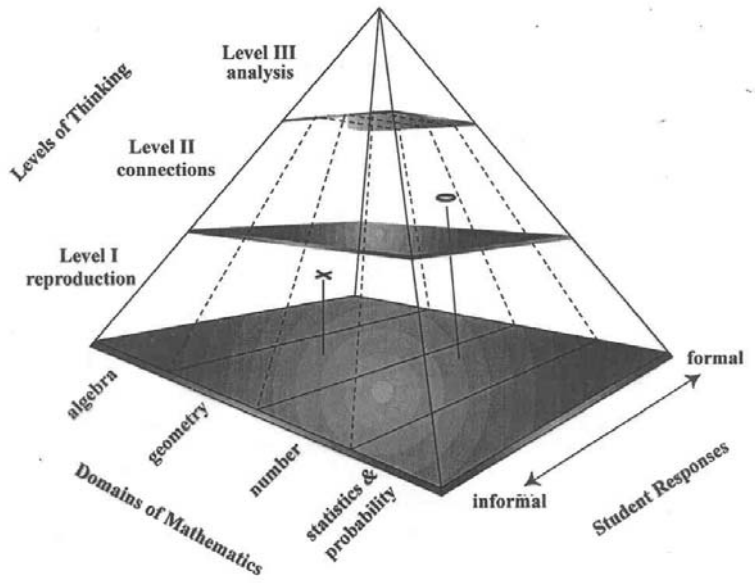


Figure 1-1. Assessment pyramid. Source: Verhage, H. & deLange, J. (1997, April). Mathematics Education and Assessment. *Pythagoras, 42,* 14-20.

The assessment tasks used in the study come from two assessment systems created for the study, identified as "External Assessment System" and "Problem Solving Assessment." In the External Assessment System (EAS), four different instruments, one for each grade, were used to assess different aspects of students' understanding of mathematics and to document the impact of variables related to curriculum and instruction on students' mathematics achievement. Similarly, four problem-solving assessments

(PSA), one for each grade, were constructed for the study by the Freudenthal Institute. The format for all of these tasks was constructed response with all answers coded with respect to the quality of the answer (partial credit) and strategy used.

External Assessment System

Each of the four EAS instruments made use of publicly released tasks from the 1992 NAEP, 1996 NAEP, and the 1996 TIMSS. Although student performance on such tests offers limited views of student understanding (Greeno, Pearson, & Schoenfeld, 1996), student performance on such tests is often used as evidence of a program's effectiveness. Thus, the EAS we created allowed us to relate student performance in this study to representative national and international samples of students. Details of the development of the EAS instruments can be found in (Webb, Romberg, & Shafer, 2000). Each instrument contained items evenly divided among four strands: number, geometry, algebra, and statistics and probability. In order to examine growth over time, a selection of items of moderate difficulty was repeated on each assessment.

EAS Pilot Assessment. Five anchor items and three non-anchor items were chosen for each content strand, for a total of 32 items on each assessment. National 8th grade *p*-values (i.e., percentage of correct responses in the sample tested) were used to classify items as easy, anchor, and difficult. The general difficulty level of each assessment was raised per increase in grade level by distributing easy, anchor, and difficult items in the manner described in Table 1-1. In the Grade 5 EAS, three easy items and five anchor items were selected for each of four strands. For each successive grade level, one easy item was replaced by a difficult item in each strand.

Table 1-1
Difficulty Rating and Number of Items by Grade-Level Exam, Pilot Assessment

Rating	Mean *p* -value	Number of items			
		Grade 5	Grade 6	Grade 7	Grade 8
Easy	80	12	8	4	0
Anchor	60	20	20	20	20
Difficult	40	0	4	8	12

The criteria used to select the items were as follows:

- within each strand items should reflect a range of strand content (e.g., for the number strand —computation, number theory, number representations, proportions, percents, estimation, and so on);
- three items in each strand, non-anchor items, should increase in difficulty from fifth to eighth grade (see Table 1-1);
- the ratio of multiple choice items to constructed-response items should be similar to the NAEP and TIMSS (i.e., 70% multiple choice, 30% constructed-response or extended-response).

In spring 1997, 10 classes (265 students) in nine different schools participated in a pilot study of the EAS. Three of the nine schools served students from large urban school districts. Test conditions allowed for use of calculators, and students were given two class periods to complete the assessment. Students were asked to record beginning and ending times each day. Assessment items were scored by three project assistants according to guidelines provided with the NAEP and TIMSS items. Project assistants later visited pilot sites to discuss results with teachers that participated in the pilot.

As shown in Table 1-2, mean pilot *p*-values were generally one standard deviation above mean national 8th grade *p*-values. Students from fifth through seventh grade had little difficulty in completing eighth-grade items.

Table 1-2
P-Value Means by Problem Difficulty Rating (National vs. Pilot Results)

	5th grade		6th grade		7th grade	
Item type	**National**	**Pilot (N=73)**	**National**	**Pilot (N=65)**	**National**	**Pilot (N=78)**
Easy	80.6 (8.0)	87.7 (8.6)	78.2 (7.4)	92.6 (6.1)	72.5 (3.9)	84.2 (6.4)
Anchor	60.7 (9.4)	71.5 (11.8)	62.5 (7.4)	81.0 (12.1)	60.8 (8.8)	78.8 (12.0)
Difficult	N/A	N/A	40.1 (10.3)	63.8 (14.5)	36.1 (9.9)	51.4 (14.7)
Overall	69.3 (13.3)	78.4 (13.2)	63.5 (14.3)	81.3 (16.3)	55.3 (15.6)	71.3 (17.8)

Note: Values are given as Mean (Std. Dev.)

Although students were given 90 minutes to complete the assessment, mean completion times for fifth-, sixth- and seventh-grade assessments were 31, 27, and 40 minutes respectively.

Of particular interest to us was the relatively strong performance of the fifth-grade students. Even when taking into account probable demographic differences between the sample of students tested in this pilot and the sample of students that were tested for the NAEP and the TIMSS, we did not expect our fifth-grade pilot sample to exceed the eighth-grade *p*-values. Because we were designing the EAS for a longitudinal study, we had some concern that a ceiling effect might occur in the first year of administering the assessment.

Revision of the EAS. Given the results and our interest in following student performance on NAEP and TIMSS items over several years, each assessment was revised to reflect a new standard for selecting easy, anchor, and difficult items. Because the most difficult items on the NAEP and TIMSS were constructed- and extended-response items, more of these items were included in the anchor and difficult categories. The *p*-values for easy, anchor, and difficult items on the revised assessment were 64.0, 40.0, and 24.2, respectively, reflecting a decrease in mean *p*-value of approximately 20 %. To reduce the time needed to take the assessment, the number of non-anchor questions was reduced to 2 per strand, leaving a total of 28 items on each assessment (see Table 1-3).

Table 1-3
Difficulty Rating of Items by Grade-Level for the External Assessment System

Rating	Mean *p* -value	Number of Items			
		Grade 5	Grade 6	Grade 7	Grade 8
Easy	64	8	6	2	0
Anchor	40	20	20	20	20
Difficult	24	0	2	6	8

Fifth-Grade External Assessment. The fifth-grade assessment contained 28 items, with an equal distribution of seven items across the four content strands. Five anchor items in each content strand were used on each grade level assessment to monitor growth in student achievement over time. The questions in the number strand required students to identify appropriate operations, use whole numbers, fractions, percents and decimals to solve problems, interpret a ratio, and calculate absolute and relative comparisons. The questions selected for the geometry and measurement strand required students to visualize a three-dimensional object from a two-dimensional net, use proportional reasoning to solve problems (e.g., estimate distance on a map using a scale), draw a rectangle with specific dimensions and calculate its area, and calculate the area of a square and circle. The questions in the algebra strand required students to reason about variables, simplify algebraic expressions and solve equations, identify a point that would lie on a line given

two other points, extend patterns, and model situations that involve patterns. The questions selected for the statistics and probability strand required students to interpret ratios in a probability context and find the probability for a given situation, construct a sample space, interpret a line graph, and critically analyze graphical representations of data. With the exception of one statistics item (i.e., Metro Rail), all of the items used in this assessment addressed Level I and II reasoning. The design of the fifth-grade EAS is summarized in Table 1-4. The number of items in each content domain and the number of items designed to elicit the "level-of thinking" are provided.

Table 1-4
Fifth-Grade EAS Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	7	3	4	0
Geometry	7	4	3	0
Algebra	7	5	2	0
Stats/Prob	7	5	1	1
Totals	28	17	10	1

Sixth-Grade External Assessment. The majority of items used on the sixth-grade EAS were the same items used on the fifth-grade EAS. Two of the easier number and probability questions used on the fifth-grade EAS were replaced by two items that were more difficult. One of these new items required students to use place value concepts to reason about the best combination of numbers for a subtraction problem. The other new item required students to interpret a scatter plot to determine the median. The design of the sixth-grade EAS is summarized in Table 1-5.

Table 1-5
Sixth-Grade EAS Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	7	3	4	0
Geometry	7	5	2	0
Algebra	7	5	2	0
Stats/Prob	7	5	1	1
Totals	28	18	9	1

Seventh-Grade External Assessment. Four items from the sixth-grade EAS, one from each strand, were replaced by more difficult items to create the seventh-grade EAS. These new items required students to evaluate an expression using the order of operations, demonstrate their understanding of variables, compare the cost of two different rates, and use properties of similar triangles to find an unknown side length. The design of the seventh-grade PSA is summarized in Table 1-6.

Table 1-6
Seventh-Grade EAS Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	7	3	4	0
Geometry	7	5	2	0
Algebra	7	4	3	0
Stats/Prob	7	4	2	1
Totals	28	16	11	1

Eighth-Grade External Assessment. The items selected for this assessment are nearly equally distributed among Level I and II reasoning, with one item at Level III. Two of the easier items used on the EAS for previous grades were replaced by two new items. One of the new questions required students to model a situation and apply geometry concepts to describe a particular region. The other new item asked students to interpret, extend, and generalize a given pattern. The design of the eighth-grade EAS is summarized in Table 1-7. Note the increasing level of reasoning expected in contrast with the EAS items used at Grades 5 and 6.

Table 1-7
Eighth-Grade EAS Items by Content Domain Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	7	3	4	0
Geometry	7	4	3	0
Algebra	7	3	4	0
Stats/Prob	7	4	2	1
Totals	28	14	13	1

Problem Solving Assessments

Four end-of-year Problem Solving Assessments (PSAs) were designed for the study, one for each grade 5, 6, 7, and 8. The assessments were designed to align with the general curricular goals of problem solving, communication, reasoning, and connections advocated by NCTM's *Curriculum and Evaluation Standards* (1989). Assessment items were presented in contexts; multiple items were associated with each context. Items were written with the intent that students' strategies would progress from informal models and calculations to more formal symbolic notations over the grade levels. Details about the construction, scoring, and coding of all items can be found in Shafer, Romberg, & Wagner (2005).

Tasks related to four strands of mathematics (number, algebra, geometry, data/statistics) were included in these assessments, with decreasing emphasis on number concepts as the grade levels increase. Although the problem solving assessments were designed as part of a study evaluating the impact of MiC, the assessments are not curriculum specific. The PSAs contained items that all students, regardless of the curriculum studied, should be able to solve successfully. Moreover, PSA items were accessible on a variety of levels so that students who relied on direct modeling or drawing strategies and experienced success with these problems. At the same time other students used more sophisticated strategies, such as the application of an invented algorithm, to solve these same problems. To the extent that a paper-and-pencil assessment can be used as an indicator of student thinking, the PSAs were designed to elicit students' thinking processes. To facilitate this elicitation, the general directions to the test requested that students demonstrate how they arrived at their answers.

The PSAs, aligned with the Assessment Pyramid in Figure 1-1, assessed students' thinking at all three levels and across all mathematical domains with easy and increasingly difficult items. Researchers at the Freudenthal Institute in The Netherlands wrote the initial PSA items and provided the point values for each item. The project staff then adjusted the language, names, and contexts for use with American students without altering the mathematical intent of the items (Dekker et al, 1997-98). The project staff added items on fraction operations (often deemed a benchmark topic in middle-school instruction in the U.S.) on the sixth- and seventh-grade assessments. Following the revisions draft versions of the assessments were pilot-tested in classes using MiC and in classes using conventional curricula. Availability of four-function calculators and other tools including rulers and compass cards was encouraged. Students were given two class periods to complete the assessment. The pilot tests were scored by three graduate project assistants according to initial rubrics provided by the Dutch. The PSAs were then revised for use in the study. In cases in which students experienced little success, items were reworded or items were added as scaffolding in order to lead students toward meaningful solutions of more difficult items. In the PSAs, score points differentiated correct, partially correct, and incorrect responses. Refinements were made in partial-credit scoring by incorporating responses and strategies that appeared frequently but were not identified prior to the pilot test. Codes were assigned for items in which differences in students' solution strategies and explanations were sought and identified. Strategy codes were assigned when the item required that students show and/or explain their work or provide justification for their conclusions. Codes were created for (a) computational strategies including algorithms, flexible use of numbers, drawings, and tables; (b) explanations and descriptions; (c) use of patterns; (d) algebraic strategies; (e) geometric and

measurement strategies; and (f) justifications. Codes were also used to classify incorrect strategies such as arbitrary calculation in which numbers given in the item were used without demonstrating an understanding of the mathematics necessary to successfully complete a solution. Strategy codes captured the variation in student responses, yet at the same time, provided consistency in coding across grade levels.

Fifth-Grade Problem Solving Assessment. The fifth-grade assessment contained twenty-two items organized around a class field trip to the zoo. Seven sub-sections provided specific contexts related to the trip. The first section, A Visit to the Zoo, had of four questions about the cost of the trip and transportation from school to the zoo. The mathematical content included multiplication with whole numbers and decimals; calculation of percent discount; interpretation of a product in the given context; and interpretation of a map using cardinal directions. The second section, The Monkeys, contained three questions, involved feeding the monkeys at the zoo and included multiplication of whole numbers and fractions; interpretation of a mixed number in a context; and expressing the result of calculation as a fraction, decimal, or ratio. The third section, Buildings, contained the map of the zoo with top views of the buildings as well as their location; and on an adjacent page, side views of the same buildings were shown. The two items in this section had students interpret two-dimensional representations of three-dimensional objects by matching front or side views of buildings with top views of the same buildings, and determine that two representations of the zoo were created with different scales. The fourth section, Pools, contained three items that included reading decimal numbers from a scale to determine the height of a building, addition and subtraction of rational numbers, and placing a decimal accurately on a scale. The fifth section, Snail, was a single item that involved forming a valid conclusion about a mathematical statement based on information gleaned from a photograph. The sixth section, Playground, contained a set of four items on algebraic reasoning set in the context of the construction of a playground at the zoo. These items were part of a set of items on all four grade level assessments used to investigate growth over time in students' understanding of geometric and algebraic concepts. On the fifth-grade PSA, The Playground, the dimensions of which remained undecided, was a square composed of white square tiles surrounded by a ring of shaded tiles. Students are asked to reason about the design as the playground increased in size and to form a generalization of the pattern with respect to odd and even numbers of tiles. The final section of the assessment, Questionnaire, composed of five items, asked students to interpret the results from a questionnaire designed to determine the success of the field trip, analyze statistical information, construct an appropriate graphical representation of the data, and form conclusions based on their analysis. The design of the fifth-grade PSA is summarized in Table 1-8. The number of items in each content domain and the number of items designed to elicit the "level-of thinking" are provided. Also, note that no items solely assessed algebraic concepts at this grade level. Rather, four items involved some algebraic notions in relationship to number and geometry.

Table 1-8
Fifth-Grade PSA Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	9	8	1	0
Geometry	3	1	2	0
Algebra	0	0	0	0
Stats/Prob	5	2	3	0
Num/Geo	1	0	1	0
Num/Alg	1	1	0	0
Alg/Geo	2	0	1	1
Alg/Geo/Num	1	0	0	1
Totals	22	12	8	2

Sixth-Grade Problem Solving Assessment. The sixth-grade PSA was designed to concentrate on number concepts but include questions from all strands of mathematics. This assessment contains approximately equal distribution of items on Levels I and II and considerably fewer Level III items. The overriding context for this assessment was a local park and the responsibilities of two students who worked as volunteers in the park's ranger station. The assessment contained twenty-four questions in six sub-sections. The first section, The Ranger Station, contained six questions that included calculation and interpretation of the mean, solving multiple-step problems that involved fractions and integers, and multiplication of a fraction by a fraction. The second section, A Patio, composed of three items, related to the construction of a patio to be built in the park. These items were part of a set of items on all four grade level assessments used to investigate growth over time in students' understanding of geometric and algebraic concepts. On the sixth-grade PSA, Students interpreted and extended a pattern, and generalized the pattern using reasoning about odd and even numbers. The third section, Fly One Day, contained four items concerning the flying habits of birds. The mathematical content included multiplication with whole numbers and decimals, and solving for an unknown using a formula. In the fourth section, Bird Watchers' Bulletin, which was composed of two items, students compared two subscription plans for a magazine published by the volunteers at the ranger station. The fifth section, Selling Tickets, consisting of one item, students solved two equations with two unknowns as they determined costs of admission to the park for adults and children. The final section of the assessment, Birds of All Sizes, which contained a set of eight items, was set in the context of the weights, wingspreads, and surface areas of the wings of several birds. Students drew a triangle with given specifications, calculated the surface area of an irregular shape, read and plotted coordinate points on a graph, determined the unknown in a given formula, and referenced specific points in a graph to support or refute a given statement.

The design of the sixth-grade PSA is summarized in Table 1-9. Note that no items solely assessed concepts in statistics or probability at this grade level. Rather, four items involved some statistical notions in relationship to number and algebra.

Table 1-9
Sixth-Grade PSA Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	7	6	1	0
Geometry	2	0	2	0
Algebra	2	0	2	0
Stats/Prob	0	0	0	0
Num/Alg	7	2	4	1
Alg/Geo	2	1	1	0
Stats/Num	2	1	1	0
Stats/Alg	2	2	0	0
Totals	24	12	11	1

Seventh-Grade Problem Solving Assessment. The seventh-grade PSA included twenty-six items from all strands of mathematics and an increasing concentration on concepts other than number. The seventh-grade PSA is organized around five distinct contexts. In the first context, Baby Feeding, consisting of seven items, students investigated amounts of formula for feeding newborns and the packaging of a specific brand of formula. They interpreted information from a table and graph, calculated with whole numbers, fractions, and decimals, made metric conversions, and calculated volumes of rectangular prisms. In the second section, The Pentagon, composed of three items; students reasoned about the relationship of angle measures and side lengths of a given triangle, analyzed the effect of the sun's rays on the shadow of a building, and determined the measure of an interior angle of a regular pentagon. The third section of this assessment, Airships, contained four items that required students to reason about the relative nature of percent and construct and interpret a graph of exponential decrease. The fourth section, Pyramids, which contained seven items, asked students to complete a drawing and write a description of a pentagonal pyramid, identify vertices, faces, and edges of various pyramids, generalize a pattern using variables, and simplify expressions involving variables. The final context, Playgrounds, was composed of five items that were an extension of the Playground and Patio contexts in the fifth- and sixth- grade PSA, all of which were designed to assess growth in algebraic reasoning over time. Students interpreted and extended a pattern of square tiles, generalized the pattern using variables, and used the generalization to find an unknown. The final item assessed students' abilities to calculate the area of a rectangle, the dimensions of which were mixed numbers.

The design of the seventh-grade PSA is summarized in Table 1-10. Note that no items solely assessed concepts in statistics or probability at this grade level. Rather, three items involved some statistical notions in relationship to algebra.

Table 1-10
Seventh-Grade PSA Items by Content Domain and Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	3	1	1	1
Geometry	7	4	2	1
Algebra	7	2	4	1
Stats/Prob	0	0	0	0
Num/Alg	3	3	0	0
Num/Geo	1	1	0	0
Alg/Geo	1	1	0	0
Alg/Stats	3	2	1	0
Num/Alg/Geo	1	0	1	0
Totals	26	14	9	3

Eighth-Grade Problem Solving Assessment. The eighth-grade PSA included twenty-one questions from all content strands and a decreased emphasis on number concepts as compared with the assessments at other grade levels. Although this assessment contained a majority of items to elicit thinking at Levels I and II, the items were more difficult to answer. In addition, a number of items were designed to elicit reasoning at Level III, which reflects more formalized mathematical ideas. The eighth-grade PSA was organized around seven distinct contexts. The first context on this assessment, Club Members, contained one question in which students critically analyze a bar graph and recognize the incorrect scaling of the y-axis. The second section, Lopsided, contained three items that involved three structures that have begun to lean. Students solved for the unknown in a proportional situation, calculated using a given formula, completed metric conversions, and compared calculated ratios. In the third section, Keys, which contained three items, asked students to determine all possible combinations given a sample space and constraints, and form conclusions about the validity of given statements, The fourth section, Seesaw, included two items in which students used properties of similar triangles to determine missing dimensions. The fifth section, Stretch, was composed of five items in the context of an experiment conducted in an eighth-grade science class. Students identified a pattern based on data in a table or plotted in a graph to determine the y-value of a data point given the x coordinate, calculated slope of a given line based on values read from graph, wrote a formula for the equation of the line in the graph, and formed conclusions about the validity of an argument related to the graph. The sixth section, Parking, consisting of three items, was organized around a parking lot paved with circular stones. Students calculated the area of the circle and completed multiple-step calculations involving whole numbers and decimals. The final section, Cubes, containing three items, extended the Patio

and Playground contexts in earlier assessments to assess growth in algebraic reasoning over time, with the exception that the eighth-grade context involved a series of cubes. Students analyzed the pattern, and extended and generalized the pattern using variables,

The design of the eighth-grade PSA is summarized in Table 1-11. Note that only one item solely assessed number concepts at this grade level although six items involved some number in relationship to algebra and statistics. Also, note the increasing level of reasoning expected in contrast with the PSAs at Grades 5. 6. and 7.

Table 1-11
Eighth-Grade PSA Items by Content Domain Level of Reasoning

Domain	**N**	**I**	**II**	**III**
Number	1	0	1	0
Geometry	5	1	2	2
Algebra	4	1	2	1
Stats/Prob	1	0	1	0
Num/Alg	3	2	1	0
Alg/Geo	4	0	3	1
Num/Stats	3	0	2	1
Totals	21	4	12	5

Scoring and Coding Student Responses

Twenty-six scoring institutes were conducted from 1998 to 2000 to score the Problem Solving Assessments and External Assessments administered to students as part of the longitudinal/cross-sectional study. During the scoring institutes, each student response was scored by two raters who were experienced elementary- and middle-school teachers. Interrater reliability was calculated to assess the scoring procedure and the quality of the scoring. Interrater reliability is the frequency at which the two raters who scored a student response agreed with one another.[1] The purpose of this section is to describe the scoring procedure at these scoring institutes, to summarize interrater reliability, and to report factors that influenced interrater reliability.

Problem Solving Assessments (PSAs) and External Assessments (EAs) were administered to all study students in each of the four districts in the study. Ten of the twenty-six scoring institutes, held in spring of each study year, were conducted in the districts; study teachers were the raters, providing them an opportunity to participate in the scoring process, learn about the assessments, and examine student work from a variety of teachers. The other sixteen scoring institutes were held at the University of Wisconsin–

[1] If there was a discrepancy between the scores, a third rater adjudicated. Occasionally more adjudications were necessary. When two raters agreed upon a score, it was considered final.

Madison; teachers (Grades 3–12) from schools in the Madison area were the raters. Seven of the Madison institutes were week-long sessions and were held during the summer. The remaining nine Madison institutes were held on Saturdays during the fall of 1998. These fall scoring institutes were added because more assessments were received after the first scoring institutes and changes in scoring rubrics required rescoring items on many assessments.

The number of PSAs and EAs varied by the number of study students at each grade level and by the number of absentees when the assessments were administered. In the first year of the study, the PSA contained eighteen contexts, each of which was scored separately. In contrast to the PSA, five EA anchor items (constructed-response items) were repeated on each grade-specific assessment. In addition, three other constructed-response items were scored (Contexts 1, 7, and 8). The items on the PSA were scored in clusters, according to problem context.[2] The number of contexts scored at each institute varied from one to thirteen depending on the number of raters, the number of assessments, and the number of days the institute lasted. On average, two PSA or EA contexts were scored each day.

Rater Training.

On average, raters and adjudicators were trained 0.5 to 1 hour for each PSA context and 0.25 to 0.5 hour for each EA context. After the first two Madison scoring institutes, the majority of the raters were veteran raters, needing less training time during the rest of the institutes. The training included raters solving the problems in a particular context, presentation and discussion of the scoring rubric and strategy codes (if any) for that context, and examination of scored student work samples that clarified each portion of the rubric or each strategy code for each item. The context-specific training was followed by instruction on the general procedures for scoring (explained below). This context-specific training alternated with periods of scoring. For example, during a typical day at one of the Madison scoring institutes, all raters were trained in the scoring of "A Visit to the Zoo" (the first context, items 1–4, on the Grade 5 PSA). After raters and adjudicators received training for one context, they were directed to begin scoring student responses on that context. When all of the Grade 5 PSAs were scored and adjudicated for that context, the raters were trained in the scoring of "Monkeys" (the second context, items 5–7, on the Grade 5 PSA). Raters then scored "Monkeys."

Description of Rubrics

Item-specific scoring rubrics were used in the PSA. Scores ranged from X (no response) or 0 (incorrect response) to 4, depending on the complexity of the problem. Correct answers for less complex items, for example, were scored 1; answers for the most complex items could receive as many as 4 points (see Table 1-12). The complexity of these rubrics was reflected in the discussion during training. For many of these items, raters were also asked to determine the strategy evident in the student's solution

[2] The context numbers apply only to the context groupings in this paper. (The context groupings are numbered differently in the 1998–1999 and 1999–2000 Interrater Reliability papers (Folgert & Shafer, 2000, 2001a, 2001b) because some of the items [e.g., Item 6 from EA5-Context 1 in this paper] do not apply in the other papers.)

from a predetermined list of codes specific to the item. Some of the scoring rubrics evolved during the scoring process. Two factors influenced this development. First, PSA items were pilot-tested with groups of 75–100 students. Rubrics and strategy codes were created and revised based on those student samples. However, because during the study the PSA was administered to hundreds of students, additional types of student responses and solution strategies were detected. These newly discovered cases were integrated into existing scoring and coding schemes. Second, as a result of the pilot test, some items were rewritten, and new items were included. Because student work for changed and new items was unavailable prior to the administration of the assessments, rubrics and strategy codes were based on anticipated student responses. As student responses were examined during the rating process, rubrics and lists of strategy codes were refined to better represent the variety of responses actually demonstrated on specific items. When rubrics or lists of strategies were changed, items scored prior to the changes were rescored.

Item-specific rubrics were also used with the EA. The rubrics used in scoring EA items were identical to rubrics used in the NAEP and TIMSS assessments. Scoring involved assigning point scores (scoring for most contexts was based on partial-credit rubrics) and strategy codes when appropriate. Interrater reliability was determined only for point scores. These rubrics were generally less complex than PSA rubrics, and, because EAs were designed to yield comparisons with national and international samples of students involved in the NAEP and TIMSS, these rubrics could not be changed.

Table 1-12

Sample Scoring Rubric Used with PSA Items

Points	Response
4	*Student must have both to receive four points:* Correct answer: 72 (tiles) *with* Clear explanation: 9 x 8 = 72 or "you multiply the ring number by 8" or demonstration of how student found answer (e.g., extends table to ring #9) NOTE: if a student uses the doubling strategy in question 15, the student should find 2048 tiles as the answer here.
3	Clear and complete demonstration of valid strategy, but one minor error in computation leads to incorrect answer, e.g., Completes drawings to ring #9 but makes minor mistake in counting (69-75) or Completes drawings to ring #9 but makes minor error in playground #9 (69-75 tiles in ring) or Extends table to ring #9 w/ one minor computational error *or* Student gives correct answer* for "the total number of tiles" (361) instead of the answer for "number of tiles in ring"
2	Clear and complete demonstration of valid strategy, but two minor errors in computation lead to incorrect answer
1	Correct answer only, no explanation or strategy evident *or* *Valid strategy evident but incomplete (student finds the number of tiles in ring #6, #7 or #8 only)*
0	Incorrect answer
X	Nonscorable or no response

EA rubrics were less complicated than PSA rubrics, but, because these contexts involved anchor items repeated at each grade level in the study, in most cases, larger sets of assessments were scored for each EA context.[3] (Multiple-choice items were also scored by two raters, but did not require analysis of interrater reliability). All items, regardless of complexity, were assigned 1 point (see Table 1-13). Complexity of scoring is reflected in the breakdown of that point. Some items were scored with a fraction of a point. Some items also included codes for student strategy. Scoring, at times, was complex.

[3] Some contexts were scored at more than one scoring institute because additional assessments were received after the first round of scoring was completed.

Table 1-13

Sample Partial-Credit Scoring Rubric Used with EAS Items

Scoring Guide	
1	Correct answer. (Must state 420; must tie step 20 back to beginning of pattern in some specific form of generalization)
0.75	Correct explanation of pattern but does not include or omits the correct number of dots (420).
0.50	A partial (incomplete) correct explanation (i.e., does not tie together well).
0.25	An attempt to generalize OR draw all 20 pictures in the pattern (with a clear understanding of the pattern).
0	The work is completely incorrect, irrelevant, or off task. (A response of just 420 is a score of 0.)
X	No response.

Preparation Prior to Scoring Institutes

To assure anonymity of students, teachers, and districts, names were removed from all student assessments and student scratch papers. At the district scoring institutes, assessments from the different schools and classes were mixed randomly; at the remaining institutes, assessments from different districts were mixed randomly. Assessments were separated into packets of 5–8 assessments, and each packet was scored by two raters. Each assessment contained two rating sheets. The second rating sheet had spaces for a third rating, if adjudication was necessary. Raters recorded their assigned codes on lines next to each context they scored. This procedure allowed us to track interrater reliability by rater and by institute. Raters were typically seated in groups of four. At each table, a table leader was responsible for distribution and collection of packets given by the coordinator. The coordinator used a predetermined routing sequence (based on a Latin square) to determine the flow of packets during scoring. Adjudicators, responsible for rating student responses when the first two raters did not agree, were seated at other tables.

The Scoring Process

Each rater was given a packet of 5–8 student assessments to score. The rater scored the first assessment for a particular context and circled the score and strategy code (if applicable) on the Rater 1 Score Sheet. The rater then placed the Rater 1 Score Sheet at the back of the student assessment and placed the scored assessment at the bottom of the packet. Scoring continued until all student assessments in the packet were rated. The packet was handed to the table leader, who in turn gave the rater another packet. Scoring continued until all packets had been scored once. Packets were then randomly distributed to different tables for the second round of rating. Raters used the same scoring process, but completed the Rater 2 Score Sheet for individual assessments. Scoring continued

until all packets had been scored twice. After each packet was scored a second time, the table leader compared both rating sheets for a given student assessment and marked scores and strategy codes (if applicable) that were not in agreement. These assessments were sent in packets to an adjudicator for an additional rating. If agreement was reached between two of the now three raters, the agreed score or strategy code was used for the student response. If agreement was not reached, another adjudicator scored the response. If agreement was reached between two of the four raters, the agreed score or strategy code was used for the student response. The adjudication process continued until agreement was found between two of the raters. This routing system allowed raters who worked faster to score more assessments than slower raters. For EA multiple-choice items, packets were distributed in the same way as for PSA and EA contexts. Scoring, however, differed in that Rater 1 circled the letter selected by the student and the appropriate point value for the response (X for no response, 0 for an incorrect response, and 1 for a correct response). Rater 2 verified that the scoring was done correctly. Adjudication was unnecessary.

Some of the scoring rubrics evolved during the scoring process. Two factors influenced this development. First, PSA items were pilot-tested with groups of 75–100 students. Rubrics and strategy codes were created and revised based on those student samples. However, because during the study the PSA was administered to thousands of students, additional types of student responses and solution strategies were detected. These newly discovered cases were integrated into existing scoring and coding schemes. Second, as a result of the pilot test, some items were rewritten, and new items were included. Because student work for changed and new items was unavailable prior to the administration of the assessments, rubrics and strategy codes were based on anticipated student responses. As student responses were examined during the rating process, rubrics and lists of strategy codes were refined to better represent the variety of responses actually demonstrated on specific items. When rubrics or lists of strategies were changed, items scored prior to the changes were rescored.

Interrater Reliability by Scoring Institute

The number of assessments rated at the institutes varied depending on which assessment was scored, the number of new assessments received, and number of assessments that needed to be rescored. Interrater reliability was then calculated by scoring institute. The number of student responses given the same score points by two raters was determined and percentages were calculated. For example, of the 2,592 student responses rated for the first scoring institute, 2,069 student responses were assigned the same point scores by two raters. Therefore, the raters agreed on the point scores 79.82% of the time during the first scoring institute.

Interrater agreement was high for all twenty-six scoring institutes, ranging from a low of 73.47% at the third institute to a high of 97.14% at the eleventh institute. Interrater agreement tended to increase over time. This increase could be attributed to a high quality scoring procedure, the increasing experience of presenters and raters over time, the refinement of scoring rubrics, and the general movement from rating harder-to-score PSAs to easier-to-score EAs.

Interrater Reliability by Rater

Interrater reliability was calculated for all raters at each institute. Agreement was then determined between ratings of the both raters on individual student responses, and percentages were calculated. For instance, Rater A agreed with a second rater on 306 of the 370 student responses or 82.70% of the time. (This includes when Rater A was the second rater). Interrater agreement was high for all raters, ranging from a low of 73.47% to a high of 97.14%. Over half of the raters reached over 90% agreement with second raters. The factors that contribute to this high level of agreement can be attributed to clear rubrics, high quality presentation of rubrics and examples, and experience over the set of institutes. There was considerable variation in the number of assessments each rater scored due to the length of the scoring session in which the rater participated, the number of assessments prepared for scoring, and, especially, the speed at which the rater scored assessments.

Conclusion

At each of the scoring institutes, there was high interrater agreement indicating a high quality scoring procedure. The extensive training proved worthwhile because it reduced the number of items that needed to be adjudicated. As experienced elementary- and middle-school teachers, raters provided valuable input for clarifying some of the more complex PSA rubrics and identifying different categories of student responses and solution strategies. Through this process, rubrics became user friendly, which in turn increased interrater reliability. The scoring institutes also provided a significant professional development opportunity for teacher-raters who commented that they would make changes in their pedagogy to emphasize mathematical communication, include lessons that promoted more complex reasoning, and integrate various types of problems designed to elicit student thinking at more complex levels in their classroom assessment practice.

Coding for ACER

Following the scoring of all the instruments the student responses were coded for analysis by the Australian Council for Educational Research (ACER) as described in Chapter 2. The coding involved classifying scoring rubric responses used for PSA and EAS items according to: item format (i.e., multiple choice or constructed response); content strand; degree of formalization expected in student responses (i.e., informal, pre-formal, or formal); and level of thinking (i.e., reproduction, connections, or analysis; see previous discussion of Assessment Pyramid). Comments that described the type of mathematics required for each type of student response were also provided. These tables were developed separately by three members of the project team for each grade level assessment. Discrepancies in the classification of items were deliberated until a consensus was reached. After the coding scheme was stabilized, a seven-digit code was generated for each type of student response per assessment that identified the type of assessment (PSA or EAS), grade level of assessment (5 through 8), problem number, and student response. For example, the seven digit code, 5802018, indicated a student response to the eighth grade PSA (58), for problem two (02), and was classified as response type 18 for that particular problem. This classification scheme resulted in a significant number of response codes for each grade level assessment.

The number of response codes generated for the grade five through eight External Assessments was, respectively, 133, 133, 143, and 148. The number of different response codes generated for the grade five through eight Problem Solving Assessments was significantly higher: 250, 318, 257, and 204, respectively.

Using these seven-digit response codes, student performance data tables for each assessment were created for each grade-by-year participant group (e.g., Grade 8 - Year 2). To produce both individual and group-based performance data the rows for each table included the identification codes for student, district, school, teacher, class, treatment, participation years, gender, and ethnicity (see Table 1-14). The remaining columns of the table marked the sequence of problems for a particular assessment.

Table 1-14
Sample Student Performance Data Table Submitted to ACER

StudentID	District	School	Teacher Class	Treatment	Grade	Gender	Ethnicity	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
254	3	001700	000711000	040	080	1	5	5801001	5802004	5803001	5804001	5805001	5806001	5807011	5808001
256	3	171700	161711000	440	780	1	5	5801001	5802011	5803001	5804001	5805002	5806003	5807006	5808001
262	3	171700	164713000	440	780	2	5	5801003	5802017	5803003	5804003	5805003	5806003	5807011	5808004
265	3	171700	161714000	440	780	1	5	5801001	5802004	5803003	5804004	5805003	5806002	5807011	5808003
266	3	171700	161714000	440	780	1	5	5801001	5802010	5803003	5804003	5805001	5806003	5807006	5808012
269	3	171700	161711000	440	780	2	6	5801001	5802004	5803002	5804002	5805002	5806003	5807011	5808003

Summary

This chapter describes the design and scoring procedures of two assessment systems that were developed to monitor the impact of the MiC instructional approach on student performance. The selection of assessment tasks used in developing these measures were based on an adaptation of a three dimensional assessment pyramid. It is worth noting that the dimension of student thinking (the vertical dimension on the pyramid) also has been used to guide the selection of tasks for a large scale international study, the Program for International Student Assessment. The range of problem types designed for the Problem Solving Assessment and selected for the External Assessment System permitted the development of a single scale for the study. This scale, also known as a progress map, is described in more detail in the chapter that follows.

CHAPTER 2: THE DEVELOPMENT OF A SINGLE SCALE FOR MAPPING PROGRESS IN MATHEMATICAL COMPETENCE

Ross Turner and Gayl O'Connor
Australian Council for Educational Research[4]

Introduction

> When we measure students' [mathematics] achievement we base our measures on students' responses to individual [mathematics] tasks that are grouped together in carefully constructed test instruments. When we come to interpreting the meaning of these measures, however, we look beyond the specifics of the tasks and test instruments used to obtain the measures to the generalities of the underlying variable. That is, we are interested in the particular [mathematics] tasks only in as far as they provide us with samples of behaviors for the purpose of estimating students' [mathematics] achievement (the general variable).
>
> Underlying variables, when described, provide us with conceptual 'maps' against which we can develop assessment tasks. These maps make explicit what it means to progress, or become better in an area of learning. The maps also provide a framework against which we can monitor students' growth. Maps of this kind are continually revised and enriched as we refine our understanding of the underlying variable.[5]

The current authors' intention is to develop a progress map that describes increasing mathematical competence. The main purpose of this chapter is to describe the means by which the research team has been able to use the data from a longitudinal testing program to build and refine a picture of growth in mathematical competence. In this chapter, conceptualisation of the underlying variable to be measured in developing a mathematical progress map is considered, and approaches to its development are discussed. The approach adopted is described and the results of the application of that methodology to the data generated through this research study are presented. Finally, the progress map is presented, and illustrated with sample test items.

In order to measure and describe growth in a variable of interest, it is necessary to first define the variable, which in this case is referred to as 'mathematical competence'. Educational researchers and mathematical educators already have some understanding of

[4] The authors wish to acknowledge the contribution of our colleagues Greg Macaskill and Margaret Forster, of ACER – Greg carried out the data analyses on which the detailed development of the progress maps is based. Margaret contributed to the conceptualization of the project, and provided editorial assistance with the text of this chapter.
[5] Adapted extract from Forster, M. (in progress).

what mathematical competence means. It is possible, therefore, to make some use of a 'top-down' approach to describing mathematical competence, such approaches being characterised by their reliance on the existing knowledge of experts in the field. There exists a well-established set of curricula in mathematics that operates in schools. Teachers expose students to these curricula from their first appearance at school, and continue to work to develop mathematical competence in their students right through school. Tertiary education institutions and others train mathematics teachers and provide them with ongoing professional development in order to maximise the quality of mathematics teaching and learning. Researchers carry out research in mathematics and in mathematics education. Experts write books about mathematics teaching and learning. Teachers and researchers assess student progress in mathematics, and report that progress in various ways and for various purposes throughout schooling. Those activities contribute to, and reflect, a very substantial body of expert knowledge from which a top-down methodology can be built.

The researchers involved in this project have used that accumulated wisdom to construct a set of mathematics test items appropriate for students in the middle years of schooling, described in Chapter 1 of this monograph, that demand the kinds of skills that are part of mathematical competence as this is currently understood. The researchers have developed a battery of test items that require a variety of particular mathematical skills and understandings, related to the range of fundamental mathematical ideas that underpin the middle years mathematics curriculum that is typically provided to students – number, geometry, algebra, statistics and the processes involved with mathematical problem solving. Moreover, having observed that students typically grow in their mathematical understanding, and that there are often wide differences among students in the level of mathematical competence they can demonstrate, it is clear that if researchers or teachers wish to measure the competence of students, they must provide test items that embody a wide range of demands to ensure that the measuring instruments are appropriate to the objects of measurement. The test items therefore vary in the level of mathematical competence they demand. Some of the test items make relatively low level demands which, it could reasonably be expected, students at the early stages of growth in mathematical competence would find challenging, but which students at more advanced stages of growth would be expected to master with little difficulty. Other test items place much higher-level demands. The conception of mathematical competence that underpins the development of these test items would lead the researchers to expect that these items will be mastered by only those students who have progressed relatively far along the part of the continuum of growth in mathematical competence that is relevant to students in the middle years of schooling. Indeed one objective in developing the battery of test items used in this research has been to ensure that items are used that encompass the full range of levels of competence, within each of the mathematical areas of interest, for all students we may wish to be able to measure and describe.

While this work is based on a reasonably strong and widely understood conception of mathematical competence, it is important that at every opportunity, empirical means are used to strengthen this understanding, and to build more detail into the way in which mathematical competence is conceived. This research does that by using data from the test items, which provide observations as to what students can actually do. This is known as a 'bottom-up' approach to describing mathematical competence. Bottom-up conceptualisations are empirically derived descriptions based on the knowledge, skills and understandings that students actually demonstrate. In the case of mathematics, this means observations of what students actually do in response to mathematical stimulus

material. The responses of large numbers of students, of differing mathematical ability, to the same set of mathematical tasks, are analysed and through this process a pattern of typical growth in mathematical competence emerges. Those responses can be used to develop rich descriptions of the continuum of mathematical proficiency. And the descriptions can be exemplified with sample test items that point clearly to steps along that growth continuum. Of course, in practice there is a direct relationship between these top-down and bottom-up conceptualisations. Bottom-up conceptualisations are in effect empirically-based refinements and enrichments of initial top-down conceptualisations. This current study takes another step along the ongoing path of variable definition described by Stone, Wright and Stenner (1999):

> …variable construction never ends because it is never complete; it is ever continuing. Variables require continuous attention for their development and maintenance. The map of a variable is a visual representation of the current status of variable construction. It is a pictorial representation of the 'state of the art' in constructing a variable (Stone, Wright and Stenner, 1999, p. 309).

As well as defining what is meant by mathematical competence, and building a rich description of growth in mathematical competence, it is important that teachers and researchers are able to measure students' mathematical accomplishments so that it becomes possible to reliably locate individuals' on the growth continuum. Knowing where an individual student is on the progress map will allow improvements over time in the mathematical competence of that individual to be monitored. Knowing the location of students on the progress map will permit comparisons among groups of students, will enable the identification of individuals who appear not to be following the expected growth trajectory, and then when it is seen to be required, will lay the basis for attempts to rectify the problem through properly planned and well-targeted interventions.

Approaches to Developing Progress Maps

The approach taken here to developing a progress map for middle school mathematics follows a well established methodology, developed and used successfully in a large number of educational research projects emanating from the Australian Council for Educational Research (ACER), and built on the theoretical and empirical work of researchers at ACER (e.g. Geoff Masters and Margaret Forster). Much of that work has been in the area of developmental assessment – the process of monitoring students' progress through an area of learning so that decisions can be made to facilitate further learning (Masters & Forster, 1996a). A unique feature of developmental assessment is its use of a 'progress map'. A progress map describes the nature of development (progress, or growth) in an area of learning – knowledge, skills and understandings in the sequence in which they typically develop. A progress map provides a picture of what it means to progress in an area of learning. In developmental assessment, it provides the framework against which student development is monitored (Masters & Forster, 1996b).

Frameworks of this kind usually are developed in one of two ways: 'top-down' or 'bottom-up'. Top-down approaches require specialists to use their professional knowledge to develop a picture of the sequence in which the knowledge, skills and understandings of a learning area typically develop. Bottom-up approaches use only observations of students' responses to develop a picture of increasing understanding. (See e.g. Masters and Forster 1996a). Central to this work is an understanding of the fundamentals of objective scientific measurement and the use of Item Response Theory (IRT) to develop objective measures of student progress (see for example Rasch, 1960; Masters, 1982; Bond and Fox, 2001). In IRT, responses to items on a test are accounted for by a single latent trait. This model, developed originally by Danish mathematician Georg Rasch, proposes a mathematical relationship between a person's ability on this trait, and the difficulty of the task being attempted.

> At the "heart" of the theory is a mathematical model of how examinees at different ability levels for the trait should respond to an item (Crocker & Algina, 1986 p. 339).

Using this mathematical model, tasks and students are placed on the same scale—tasks at their difficulty location, and students at their ability location. Only those items (tasks) that conform to the model (the demands of unidimensionality) are selected for use in any one test, and are used in any description of the underlying variable.

Once tasks and students are placed on the same scale, it is possible to analyse the characteristics of ability at particular locations along the scale in terms of the tasks students are more or less likely to be able to complete correctly. If a task is of greater difficulty than a particular student's ability, then that student will have a low probability of completing the task correctly. If a task is of lower difficulty than the student's ability, then there is a high probability that the student will be able to complete the task correctly. This analysis provides an empirically based 'picture' of progress along the underlying variable.

The same methodology has also been used in projects based on comparative assessments. The Organisation for Economic Cooperation and Development's PISA project (Program for International Student Assessment) has made use of this methodology in developing the described proficiency scales it uses to discuss levels of student proficiency in the three test domains that were included in its 2000 round of testing. The PISA 2000 Technical Report (OECD, 2002, Chapter 16) describes the methodology used in the development of those described scales for the major test domain of Reading as well as the two minor test domains of Mathematics and Science. The current project uses a sequence of steps to define and describe a scale of progress in proficiency in middle school mathematics that is very much aligned with the procedures used in the OECD PISA project to define a scale of proficiency in the construct known as 'PISA Mathematical Literacy'. Based on the published PISA mathematics framework (OECD, 1999) that outlined a definition of mathematical literacy and described a number of central elements of mathematical literacy, a large number of test items were developed that would elicit evidence of the relevant mathematical behaviours. These were combined into a number of test instruments that were administered to samples of students in more than 30 participating countries. Two analytic activities then occurred in parallel. One was the generation of data based on student responses to the items, and analysis of those data using item

response modelling of the kind previously mentioned. A key output of this analysis was a set of item difficulty measures for the test items. The parallel activity was a qualitative analysis of the demands of the test items, conducted by expert mathematicians, teachers and test developers. The key output of the qualitative analysis was a set of rich descriptions of the mathematical skills and competencies needed to respond to each item. The outcomes of those two analytical steps were then combined. This led to the creation of a scaled variable, being an expression of mathematical literacy. This scaled variable was conceived as a line, with the test items being located at various points along that line according to their difficulty, and accompanied by descriptions of the mathematical behaviours associated with each item – hence a 'described proficiency scale'. Subsequent refinement of the proficiency scale led to the identification of achievement levels along the continuum of mathematical literacy, development of descriptions of typical aspects of student proficiency at each level, and validation of the levels and their descriptions.

Test equating.

Using IRT it is possible to calibrate a number of tests together, thereby equating the tests and making it possible to place all the tasks in the separate tests on a single scale. This provides a richer set of examples from which to generalize the underlying variable, and allows for a broad range of item difficulties that can then be used to measure the ability of a broad range of students. Test equating is a prerequisite for the reliable comparison of students' performances on different assessment instruments. Because tests inevitably vary in difficulty, and often in length, it cannot be assumed that a score of, say 20, on one test represents the same level of proficiency or achievement as a score of 20 on any other test. The purpose of test equating is to establish score equivalences on different tests (e.g., what score on Test A represents the same level of mathematical competence as a score of 20 on Test B?). Meaningful score equivalences of this kind can be established only between tests measuring the same achievement/proficiency dimension. Petersen, Kolen and Hoover (1989) define test equating as a set of

> …empirical procedures for establishing a relationship between raw scores on two test forms that can then be used to express the scores on one form in terms of the scores on the other form (Petersen, Kolen and Hoover (1989, p. 242).

To establish score equivalences between two tests it is necessary to collect data that allow the relative difficulties of the items in the two tests to be simultaneously estimated.

There are two standard approaches to test equating. The first approach is to ensure that the tests to be equated share some common items. These shared items provide the 'link' required to establish the relative difficulties of two tests and thus to establish score equivalences between the two tests. The second approach is to ensure that some students attempt both tests. The performances of these students provide the data required to simultaneously calibrate the items in the two tests and so to establish score equivalences. The rather complex design of this research study, which includes multiple test instruments, the testing of students at multiple year levels, and testing over multiple years, makes the use of these equating procedures an essential part of the exercise.

The data

The raw test data were generated from testing that took place over a three-year period from 1997/1998 to 1999/2000. Testing was carried out on students across four grade levels (Grades 5, 6, 7 and 8). The students in Grade 5 who were tested in 1997/98 were tested again in each of the following two years when they were in Grades 6 and 7. The students in Grade 6 who were tested in 1997/98 were also tested again in each of the following two years, when they were in Grades 7 and 8. The students in Grade 7 who were tested in 1997/98 were tested once more, in 1998/99 when they were in Grade 8. The approximate number of students tested at each grade level and in each year of the study is summarized in Figure 2-1. The number remaining in the analysis used to generate the progress map given in parentheses. The cohorts that were tested across multiple years are indicated through the use of heavy arrows. One group was too small for use in the analysis, which left eight cohort groups that were tested and used in the analysis.

Test items were of two broad types, and the structure of the tests for each grade level, are summarized in Chapter 1 of this Monograph. The items comprised 93 general mathematical problem solving assessment items (referred to as 'PSA items'), involving constructed responses that were coded by trained expert markers; and 36 standard external examination questions (referred to as 'EA items') with a closed response format (often multiple-choice, or other restricted-response forms), that could be automatically coded and scored. All students tested did two test papers: one comprises 28 of the EA items and one comprising about 20 PSA items, the precise number varying across grade levels.

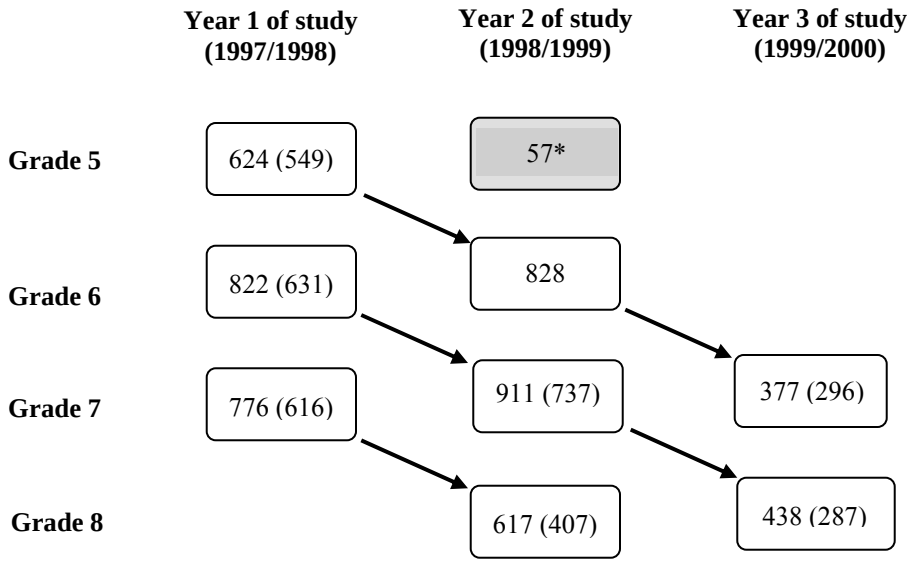


Figure 2-1: Number of students tested (and analysed) in each grade level for each year of the study.

Twenty of the EA items were included as common items across the EA test instruments used at each of the four grade levels, and there was additional overlap of items between the Grades 5 and 6 instruments (four items), the Grades 5, 6 and 7 instruments (two items), the Grades 6, 7 and 8 instruments (two items), and the Grades 7 and 8 instruments (four items). Two items were unique to each of the Grades 5 and 8 instruments. The PSA items were placed into four non-overlapping problem solving assessment instruments, one for each of the four grade levels. The Grade 5 instrument comprised 22 items; Grade 6, 24; Grade 7, 26; and Grade 8, 21 items. The placement of items in test instruments is summarized in Figure 2-2 showing the linking of items across grade levels, and the composition of grade level tests by item type.

This design meant that theoretically at least, the common EA items could be used in a test equating exercise to place all of the EA items on a single scale, and the fact that students in each grade undertook both EA and PSA items could be used to link the EA test items with the PSA test items. It would make practical sense to do so if it could be demonstrated that all these items worked well together to define a single construct of mathematical competence. This gave rise to the possibility that all items, both the PSA items and the EA items, could be placed on a single scale and used to form a single combined mathematics progress map. This would enable

the development of a rich set of descriptions of mathematical competence, using both the PSA items and the more standard EA test items from the four different content domains of mathematics that were encompassed by the tests (number, geometry, algebra and statistics).

After data cleaning, the student numbers able to be included in the analyses were reduced by about 23%. The numbers in each group are shown in Figure 2-1. These data were processed using Quest software (Adams and Khoo, 1996) and applying standard models from modern item response theory, in order to generate data that could be used to empirically investigate the aspects of mathematical competence exposed by the test items, and ultimately for the development of the progress maps.

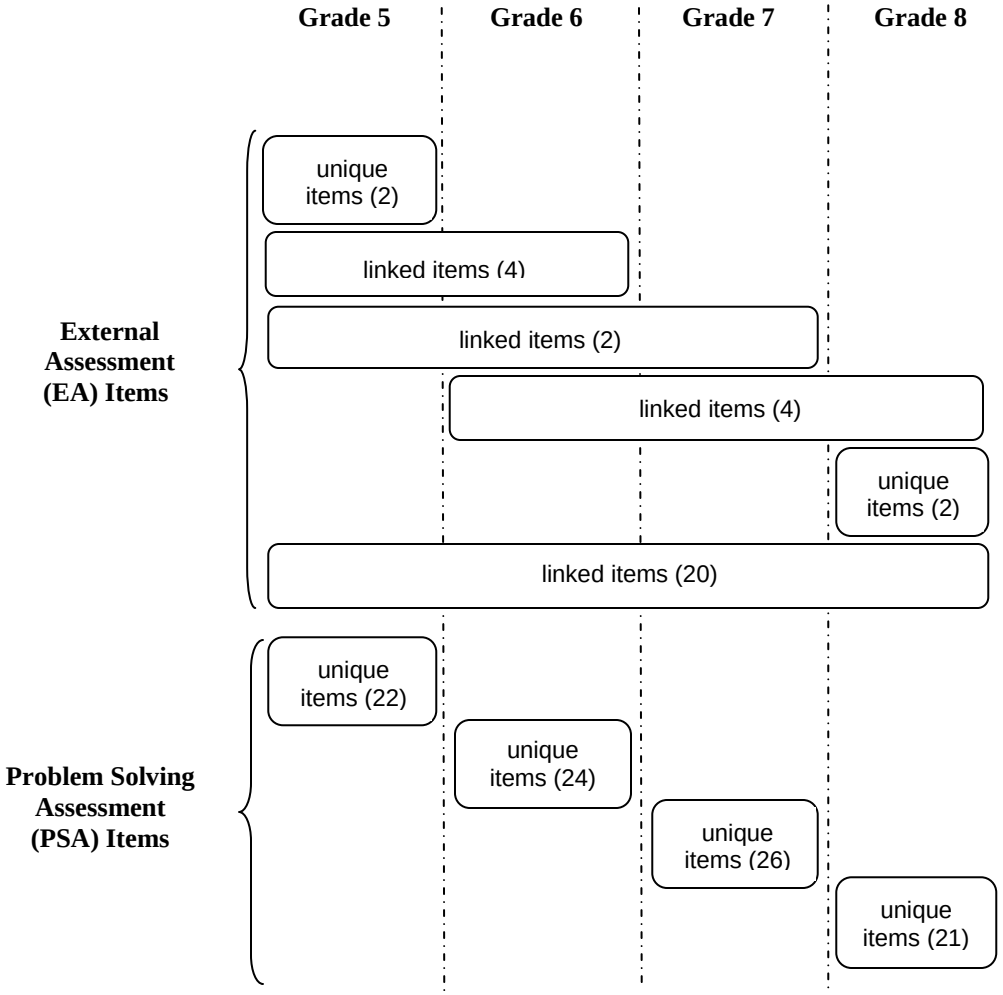


Figure 2-2: Summary of item types for each year level.

As well as the test response data, extensive qualitative analyses of the test items themselves had been carried out by test development experts and others with expertise in mathematics education and mathematics item development. These qualitative

analyses were independently reviewed and expanded as a preliminary step in developing descriptions of the mathematical demands placed on students attempting each item.

In the following sections, the specific methodology used in the present study is described, together with the outcomes of the application of that methodology to the data gathered in this study.

Methodology of Developing Progress Maps

As discussed in Chapter 1 extensive item development work took place over an extended period of time. This resulted in a set of test items for use in instruments designed to provide measures of student competence in various aspects of mathematics. The development of the progress map then proceeded in a number of stages, or steps. These steps are described briefly here and elaborated in subsequent chapter sections. The first step involved an expert qualitative analysis of the knowledge, skills and understandings needed to answer each of the questions used to assess mathematical progress. Each item was categorised according to several variables, and the demands of each item described. On this basis, items were placed into groups in a 'top-down' progress map. Secondly, the test data were analysed using the appropriate psychometric models to obtain item difficulty measures. Third, all items (and item-steps in the case of partial credit items) were ordered according to their difficulty measures (thresholds). The order of the items based on threshold values (measured in logits) was then compared to the proposed groups of items based on the qualitative analysis of item demands previously undertaken. The fourth step involved the identification of proposed achievement bands (i.e. construction of a 'bottom-up' progress map), based on the item analysis. Then, in step 5, for each proposed achievement band, the items situated within the band were scrutinised to identify any common knowledge, skills and understandings students needed to demonstrate in order to correctly respond to the items in the band. Generic descriptions of each achievement level were then formulated based on this common set of knowledge, skills and understandings. In the final stage (step 6), items were selected to appropriately illustrate each band description, and annotated to indicate how the items exemplify the demands of items located within these bands. In the following sections, each of these steps is described in detail, together with the key outcomes of each stage of development.

Development of the Progress Map for Middle School Mathematics

Step 1 Development of 'top-down' progress map

In this step, a qualitative analysis of the knowledge, skills and understandings needed to answer each of the questions used to assess progress in mathematical competence was undertaken. Each item was categorised according to several variables, using classifications derived from OECD (1999) and Romberg (2001). The demands of the individual items comprising the item set (EA and PSA items) were extensively categorised according to the same set of three variables:

- 'Competency Class' (Classes 1, 2, 3);

- ‘Competency Class Indicators’; and
- ‘Degree of Formalization’ (Informal, Preformal, Formal).

Definition of variables

Competency Class

The myriad mathematical competencies that students must demonstrate in order to solve problems, are classified into three larger competency classes, consistent with the definition of competency class in the PISA Mathematics Framework (OECD, 1999). In order to operationalism this mathematical competencies aspect through the construction of items and tests, it is helpful to organise the skills into three larger classes of competency. The three competency classes are:

- Class 1: reproduction, definitions, and computations
- Class 2: connections and integration for problem solving
- Class 3: mathematical thinking, generalisation and insight (OECD, 1999, p. 43).

The classification of the items according to mathematical competency class in this way is consistent with the use of the term ‘Mathematization’ in the PISA framework. Mathematization is defined as being able to ‘recognise and extract mathematics embedded in the situation and use mathematics to solve the problem; to analyse; to interpret; to develop their own models and strategies and to present mathematical arguments, including proofs and generalisations’ (OECD 1999, p. 45).

Romberg (2000) links the PISA framework to his view of mathematical literacy as follows:

> ..Considering mathematics as a language implies that students not only learn the concepts and procedures of mathematics (its design features) they must learn to use such ideas to solve non-routine problems and learn to mathematize in variety of situations (its social functions). The epistemological shift involves moving from judging student learning in terms of mastery of concepts and procedures to judgements about student understanding of the concepts and procedures and their ability to mathematize problem situations (Romberg, 2000 p. 1).

Romberg considers that the PISA framework for assessing mathematical literacy ‘fits the reform epistemology.’ The three competency classes as defined in the PISA framework are seen to be ‘consistent with the reform rhetoric’ (Romberg, 2000, p. 5), and accordingly competency classes have also been used to classify the items in this study.

Competency Class Indicators

To assist in the assignment of items to competency classes, elements of the PISA definitions of the three competency classes have been used to derive indicators that identify specific aspects of knowledge, skills and understandings within each of the three competency classes that students must demonstrate in order to solve a problem successfully. For competency class 1 (C1) the indicators are: standard representation, computations, definition, routine procedures, and one method. For competency class 2 (C2) the indicators are: modelling, problem solving, interpretation/reflection, and multiple well-defined methods. For competency class 3 (C3) the indicators are: problem posing, reflection, original mathematical method, mathematical insight, multiple complex methods, and generalization. Each item in this study has been classified according to which one or more of these indicators apply. Further, for most items more than one indicator was often considered relevant to the item. For each item the most relevant indicator was identified (referred to as the 'primary' indicator) and any 'secondary' indicators thought to be relevant were also identified. In general, items were assigned to the competency class containing the primary indicator.

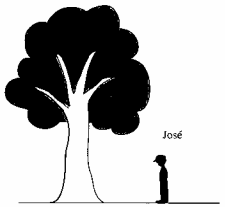
Degree of formalization

The categorisation of items by degree of formalisation is based on a model of Progressive Formalization of Concepts as outlined in Romberg (2001, p. 5). In the context of the algebra content strand, 'the progressive formalization of the mathematics involves, first, having students approach problems and acquire algebraic concepts and skills in an informal way. They use pictures, and/or diagrams of their own invention to describe mathematical situations, organize their own knowledge and work, solve problems, and explain their strategies.' In other units, 'students gradually begin to use symbols to describe situations, organize their mathematical work, or express their strategies. At the preformal level, students devise their own symbols or learn certain nonconventional notation (e.g., arrow language). Their representations of problem situations and explanations of their work are a mixture of words and symbols. In another set of units at the formal level, 'students learn and use standard conventional algebraic notation for writing expressions and equations, for manipulating algebraic expressions and solving equations, and for graphing equations.'

The mathematical content strand associated with each item was also identified. 'Algebra', 'Geometry', 'Number' and 'Statistics' strands were represented by the letters A, G, N and S respectively. A succinct description of the demands of each item was also developed. The two broader categories of competency class and degree of formalization were selected as the key variables upon which a 'top-down' progress map for mathematical competency (based on the qualitative descriptions of the items in the set) was then constructed.

Classification of items – an example

For example, the item Jose's Tree (Figure 2-3) was classified as shown in Table 2-1.



José is 1.5 m tall. About how tall is the tree?

A) 4 m

B) 6 m

C) 8 m

D) 10 m

Figure 2-3: The item - Jose's Tree

Item	**Mathematics content strand**	**Degree of formalization**	**Competency Class**	**Indicator/s**	**Item Description**
José's Tree	Geometry	Informal	2	Problem solving (primary) Computations (secondary)	Interprets a problem situation to determine the height of a tree visually, given a non-standard decimal quantity as the unit for comparison.

Table 2-1: Item classification for Jose's Tree

In this example, Jose's Tree has been categorised as having item demands consistent with a primary focus in competency class 2 (problem solving being the appropriate indicator), with one secondary focus in competency class 1 (computations being the appropriate indicator). The degree of formalization for this item is Informal, and the content strand is Geometry. The item demand is summarised in the column headed Item Description (see Table 2-1).

Grouping items according to item characteristics (creating a 'top-down' progress map)

Having established the characteristics of each item, items were grouped according to their common characteristics to create a top-down progress map. The particular selection of variables used to construct a map based on a qualitative analysis is of course somewhat arbitrary. A map could be devised according to one classification category only, or to one or more in any given priority order. In this instance, the broader categories of competency class and degree of formalization were selected, rather than the finer level of detail contained within the indicator classifications.

A 'Map Code' was assigned to each group, to indicate both the primary competency class and the degree for formalization assigned to the items in that group. For example, the item Jose's Tree was assigned the map code of C2I, indicating a primary focus in competency class 2, and an Informal degree of formalization. Six groups of items were created according to this system. The proposed 'top-down' model formed by these six groups is based on the assumption that the underlying variable of Competency Class will provide the basis for a conceptual continuum of the progress map, with the degree of formalization providing hierarchical subdivisions within each Competency Class (progressively from Informal, Preformal to Formal). The rationale for selecting Competency Class as one variable in developing the draft progress map is consistent with the approach taken in the PISA (2000) framework:

> The [competency] classes form a conceptual continuum, from simple reproduction of facts and computational skills, to the competency of making connections between different strands in order to solve real-world problem, and to the third class, which involves 'mathematization' of real-world problems and reflection on the solutions in the context of problems, using mathematical thinking, reasoning and generalisation (OECD 1999, p. 44).

The degree of formalization was selected as a second variable. Romberg notes that with respect to the model of progressive formalization of concepts, 'movement along this continuum is not necessarily smooth nor all in one direction. Students move back and forth among levels of formality depending on the problem situation or on the mathematics involved' (Romberg 2001, p. 6).

On the basis of the top-down conceptualisation, the most difficult items in the set would, in general, be expected to be those within Group 6 (C2F). Representative items are Town Populations and Playground Q4. The least difficult items would be expected to be those within Group 1 (C1I). Representative items are Flour for Cookies and Baby Feeding Q1. It was anticipated that the quantitative analysis to follow would enable the establishment of a series of achievement bands based on actual relative item difficulties as revealed in the analysis. The bands would take into account the likelihood that there will be some regions of overlap between the six proposed groups of the top-down progress map. For instance, some items with a primary focus in C1 may be more difficult than some items with a primary focus in C2, and so on (perhaps depending on the degree of formalization attributed to the item, or due to the effect of some other variable). The items located within each band would then be examined closely to identify their common characteristics.

Group	Categorisation of items [Map	Number of items in	Example Items	Item description	Predicted relative item difficulty
6	C2F	18	Town Populations Playground Q4	▪ Interpret numerical and graphical (pictographs) for population data of two towns and substantiates claim that one town has grown more over a specified period. Two strategies required: absolute comparison (constant difference) and relative comparison (proportional change) and both applied to determine solution. ▪ Use knowledge of even and odd numbers; draw conclusion based on numerical and geometric patterns; use number-based explanation (or other method).	Most difficult items (Higher-level demands)
5	C2P	24	Radio Station Snail	▪ Interpret spatial data to draw and label a diagram to define a region of overlap of concentric circles, given specific constraints and conditions. ▪ Interpret two-dimensional representation of three-dimensional objects; draw valid conclusion about a mathematical statement based on information gleaned from a photograph and assumptions based on real-world knowledge; use ratio and scale; use gross comparisons (or other method).	
4	C2I	12	Jose's Tree Pyramids Q1	▪ Interpret a problem situation to determine the height of a tree visually, given a non-standard decimal quantity as the unit for comparison. ▪ Complete an accurate drawing of a pyramid with a pentagonal base.	
3	C1F	11	Points on a	▪ Identify a third point on a straight line, given the	

			line Pentagon Q3	coordinates of two points on that line. ▪ Identify properties of angles in a two-dimensional shape; compute the interior angle of a regular pentagon using a number-based strategy; provide correct answer with clear work shown (or other method).	
2	C1P	29	k + 6 Visit Zoo Q2	▪ Interpret algebraic expression; demonstrate understanding of the concepts underlying the terms 'variable' and 'infinity'. ▪ Identify appropriate arithmetic calculation in discount context; use percent and decimal numbers; use calculator (or other method); round answer appropriately; determine correct answer.	
1	C1I	32	Flour for Cookies Baby Feeding Q1	▪ Identify and recognize an appropriate strategy (add or multiply) to solve a quantity problem involving a mixed fraction and a whole number. ▪ Read information correctly from a chart.	↑ Least difficult items (relatively low-level demands)

Note: Some items form part of a contextualized set. Individual items within a set can be distinguished by their unique Question number, for example, Playground Q4.

Table 2-2: Proposed groups of items in a 'top-down' progress map.

Step 2 - Analyze the test data using the appropriate psychometric models to obtain item difficulty measures.

In parallel with the qualitative analysis of items and preparation of the 'top down' progress map, the observed student responses to the test items were analysed with a view to assigning difficulty measures to all items on a single scale, and to permit the

development of a ‘bottom up’ progress map. Standard psychometric techniques of analysis were employed, using Quest software (Adams and Khoo, 1996).

The complexity of the experimental design posed a number of challenges as the quantitative analysis was planned and implemented. For example, what impact might the fact that progressively maturing individual students repeatedly took overlapping sets of items across successive years have on the analysis? Likewise, what might be the effect of specific practice, where some students took the same item over multiple years? Is it safe to ignore these maturation and practice effects? A number of Quest analyses were used to explore aspects of the data. Separate analytic runs were conducted on each combined test group, for example the Grade 5 cohort undertaking EA items in 1997, the same cohort in Grade 6 in cohort 1998 undertaking EA items, the Grade 6 cohorts undertaking PSA items in different years, and so on. The main purpose of these analyses was to ascertain the extent to which the items behaved consistently across student cohorts, and across time, and to check that the items worked well together to define a useful variable. A total of 16 initial runs were carried out to cover all grade/year/item type combinations. By and large, whenever items were administered – to the same grades in different years, and to different grades either in the same year or in different years – there was a very high degree of consistency in item characteristics, including in particular relative item difficulty, item discrimination and ‘fit’ of items to the models used. The graph in Figure 2-4 illustrates an example of this. It is a representation of the difficulty of each of the 20 EA items that were common to the item sets used in the four different grade levels. It shows that the profile of item difficulties for these items is remarkably consistent across years and across grades. On the graph, one item appears to be an exception to this, the item labelled ‘Town Population’. This is a partial credit item, and is the most difficult item in the set. In three of the eight cohort groups undertaking this item, not one student achieved full credit. In fact apart from this feature, the profile of that item was quite consistent across the different groups tested.

All relevant indicators suggested that all items of each type, when administered to each grade level in each year, were behaving consistently and well enough to provide useful information about a single dimension of mathematical literacy. The next step was to use ‘common student equating’ across each cohort tested to link the EA and PSA items undertaken by the students in each grade level and each year to determine whether the two different item types were providing consistent information about the same underlying dimension. At each grade/year combination, the single calibration was carefully examined to investigate item fit. At this level of aggregation of data, once again the statistical indicators generated by Quest suggested that the items worked satisfactorily together to define a single dimension for each group in the analysis.

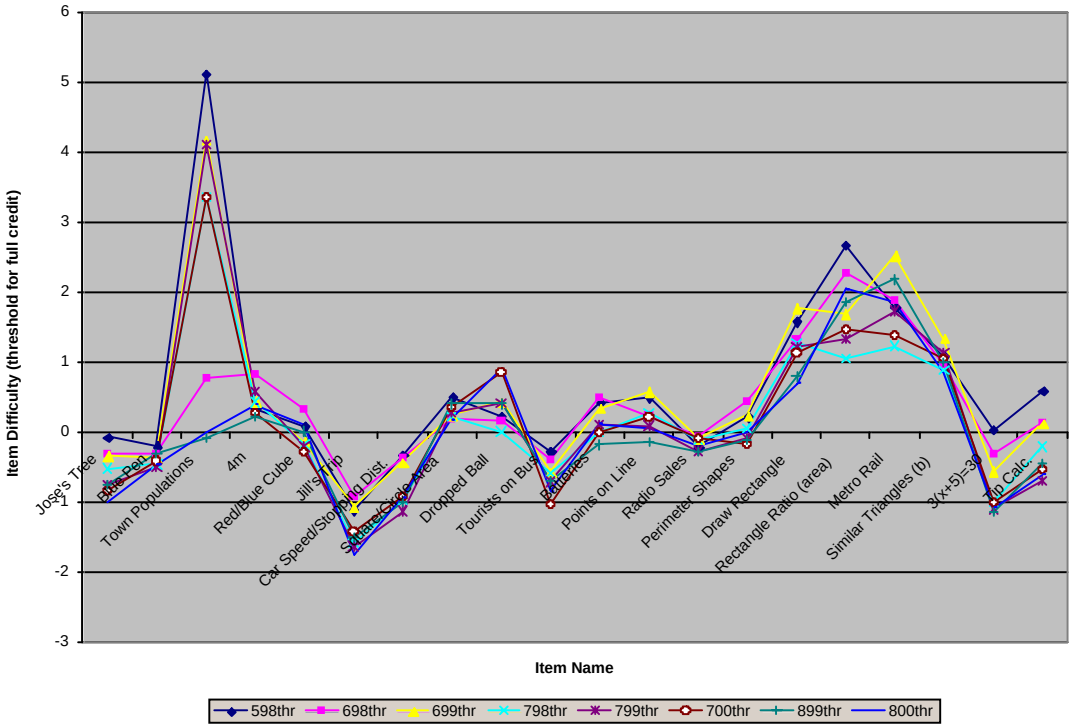


Figure 2-4: EA Item Difficulties for each Year and Grade Cohort.

Finally, common item equating was used in conjunction with the common student equating to link the measures across Grade cohorts and across years. For the purpose of estimating the item difficulty parameters that would be needed to underpin a single combined progress map, a simplifying assumption was made: that all administrations of an individual item to an individual student were treated as simple independent replications, regardless of the year in which the administration occurred, the year level of the students to whom it was administered, and the identity of the student. This simplification meant that all maturation and practice effects were dealt with by treating each successive administration of items to students remaining in the study as if they were different students. The specific practice effects were thus ignored, and the maturation effects were fully accounted for. Data were combined into a single analysis, with a view to deriving ability estimates for all students, using all items, on a single scale of mathematical competence. Two substantive checks were carried out to establish whether it was appropriate to use items of the two different types as

indicators of a single variable of mathematical competence. First, separate ability estimates were calculated for all students across grades and years on the EA and PSA scales, and the scores were correlated. The correlation used was the standard Pearson product-moment correlation, adjusted for the reliability of the two sets of estimates. The correlation was 0.87, which indicates a relatively strong relationship between measures derived from the two item-types.

The second check was to calibrate all items on a single scale and again examine item fit. In general, items fitted the model well. Four PSA items exhibited somewhat poor fit – Zoo Visit Q2 and Playground Q3 (both Grade 5 items), Selling Tickets (a Grade 6 item), and Parking Q3 (a Grade 8 item). Prior to any future use of these items, it is suggested that the scoring rubrics and other features of the items should be reviewed. The outcome of this stage in the analysis was a set of item difficulty measures that permitted the location of all of the items used in the research study on a single line. The line represents mathematical competence, and the location of items along the line indicates the 'amount' of mathematical competence required by each item. The item calibrations are provided in Turner, O'Connor, and Romberg, 2004.

Step 3 - Order all items (and item-steps in the case of partial credit items) according to the difficulty measures, and compare this form of item analysis with the qualitative analysis of item demands previously undertaken.

Ordering the items

The items were ordered according to item difficulty measures (threshold values in logits) as revealed by the data analysis. In the case of multiple-choice items, the threshold for the full credit response to each item was entered. In the case of constructed response items, where applicable, thresholds for full credit (i.e. the highest level of response) and partial credit responses were entered (item steps). For example, the item Carla's & Maria's Tiles appears four times in the ordered set of items: Carla's & Maria's Tiles.4 (4 score points for full credit response); Carla's & Maria's Tiles.3 (3 score points); Carla's & Maria's Tiles.2 (2 score points); and Carla's & Maria's Tiles.1 (1 score point for lowest credited level of response). Considering full credit responses, the items ranged in difficulty from -2.53 logits (least difficult) to 3.97 logits (most difficult). Within the data set, the EA items ranged in difficulty from -1.75 logits (least difficult) to 3.97 logits (most difficult). The least difficult EA partial credit item was located at -1.97 logits. The PSA items ranged in difficulty from -2.53 logits (least difficult) to 2.43 logits (most difficult). The least difficult partial credit item in the data set was located at -2.69 logits. EA and PSA items were distributed across essentially the same range of item difficulties. The resultant item order is provided in Turner, O'Connor, and Romberg, 2004. The items are sorted by threshold size.

Comparison of the top-down groups and the ordered item set

Three subsets of items from the ordered item set were selected for comparison purposes. The ten most difficult items and the ten least difficult items in the ordered set were selected, and compared to Groups 6 and 1 respectively of the top-down progress map. The third set contained ten items located on either side of and adjacent to the mean difficulty of the items. Hence, the third set of items is composed of items in the mid difficulty range in the ordered item set (from -0.12 to 0.18 logits).

On the basis of the top-down progress map, it was predicted that Group 6 would contain the most difficult items. Group 6 items were identified as those with the primary focus of the item in competency cluster 2 (C2), and classified as 'formal' (i.e. items classified as C2F[6]). An examination of the ten most difficult items in the ordered set (Table 2-3) reveals that four of the items are classified in C2, and as formal (*Town Populations*.2; *Questionnaire* Q3.2; *Keys* Q3.2; and *Playground* Q4.2). The most difficult item in the data set (*Town Populations*.2) was located in Group 6, and the inclusion of three other items classified as C2F provides some support for the proposal that items classified as C2F are the most difficult items. However, the other items in the most difficult set include some items classified as C2P (*Radio Station*.4; *Buildings* Q2.2; *Buildings* Q1.3; and *Snail*.3; and two items classified as C1F (*Pyramids* Q7.2 and *Pentagon* Q4.2).

On the basis of the top-down progress map it was predicted that Group 1 would contain the least difficult items. Group 1 items were identified as those classified as C1, and as informal (i.e. items classified as C1I). An examination of the ten least difficult items in the ordered set (Table 2-4) reveals that eight of the items are in fact classified as C1, and as informal (*Baby Feeding* Q1; *Pools* Q1.2; *Ranger Station* Q1; *Playground* Q1; *Nine Chips*; *Jill's Trip*; *Similar Triangles*; and *Stretch* Q1. However, one item is classified as C2I (*Playgrounds* Q1), and one item is classified as C1P ($k + 6$). From these comparisons at each extreme of the top-down progress map, it seems reasonable to conclude that, in general, the least difficult items are indeed those classified as C1I. However, for the most difficult items, only some items are classified as C2F.

A selection of ten items located around the mid difficulty range (-0.12 to 0.18 logits) was also examined (Table 2-5). Whilst all of these items are of similar difficulty, both items classified as C1 and items classified as C2 are represented. For the C1 items, two items are informal, one item is preformal, and three items are formal. For the C2 items, all four items are preformal. Examination of the items from the middle range of difficulty (-0.12 to 0.18 logits) suggests that some items classified in C1 may be more difficult than some items in C2. Further, the degree of formalization does not appear to directly correlate to item difficulty in the way proposed by the top-down progress map conceptualisation.

Therefore, the categories of competency class and degree of formalization do not appear to strictly account for differences in item difficulty as predicted by the 'top-down' progress map model. Closer examination of the specific indicators associated with each item, and the item descriptions, provides further insight into the factors that do in fact contribute to item difficulty in step 5, and can provide explanations as to why some items are not located as predicted by the top-down progress map. Further consideration of factors that may affect item difficulty is also included in the commentary for Step 6 (see *Further examination of items*).

[6] Only items classified in either C1 or C2 were represented in the data set. There were no items with a primary indicator in C3.

In Step 4, achievement bands based on logit ranges are proposed. In Step 5, the common characteristics of the items falling within each band are identified. Indicator/s and item descriptions – in conjunction with competency class and degree of formalization – were examined further to establish the common characteristics determining item difficulty for each achievement band (and hence accounting for observed differences between the actual item difficulties and the groups proposed for the top-down progress map).

Item	**Mathematics content domain**	**Degree of formalization**	**Competency Class**	**Map Code**	**Indicator/s**	**Item Description**
Town Populations.2	Number	Formal	2	C2F	Interpretation/reflection (primary focus) Computations; Problem solving (secondary foci)	Interpret numerical and graphical (pictographs) for population data of two towns and substantiates claim that one town has grown more over a specified period. Two strategies required: absolute comparison (constant difference) and relative comparison (proportional change) and both applied to determine solution.
Radio Station.4	Geometry	Preformal	2	C2P	Interpretation/reflection (primary focus) Problem solving (secondary focus)	Interpret spatial data to draw and label a diagram to define a region of overlap of concentric circles, given specific constraints and conditions.
BuildingsQ2.2	Geometry	Preformal	2	C2P	Interpretation/reflection (primary focus) Computations; Definition; Problem solving (secondary foci)	Compare two-dimensional representations; use visual estimate (or other method) to determine that the representations are on different scales; provide correct answer and correct justification.
PyramidsQ7.2	Algebra	Formal	1	C1F	Standard representation (primary focus) Computations; Interpretation/reflection (secondary foci)	Substitute expressions for variables in a formula; combine like terms; simplify to show equality; provide correct answer and correct explanation.
Questionnaire Q3.2	Statistics	Formal	2	C2F	Interpretation/reflection (primary focus) Definition (secondary focus)	Identify appropriate graph to graphically represent statistical information; explain why the graph is appropriate.

KeysQ3.2	Statistics	Formal	2	C2F	Problem solving (primary focus) Computation; Modelling; Interpretation/reflection; Mathematical insight; Generalization (secondary foci)	Interpret problem situation; determine possible combinations in sample space given constraints; draw conclusion about validity of mathematical statement; draw models and/or count possible combinations; provide correct conclusion with correct explanation.
BuildingsQ1.3	Geometry	Preformal	2	C2P	Interpretation/reflection (primary focus) Definition (secondary focus)	Interpret two-dimensional representation of a three-dimensional situation; match front and/or side views of buildings with top views of the same buildings on a map; provide five correct answers.
Snail.3	Number	Preformal	2	C2P	Problem solving (primary focus) Computation; Interpretation/reflection (secondary foci)	Interpret two-dimensional representation of three-dimensional objects; draw valid conclusion about a mathematical statement based on information gleaned from a photograph and assumptions based on real-world knowledge; use ratio and scale; use gross comparisons (or other method).
Playground Q4.2	Number	Formal	2	C2F	Problem solving (primary focus) Computation; Definition; Interpretation/reflection; Generalization (secondary foci)	Use knowledge of even and odd numbers; draw conclusion based on numerical and geometric patterns; use number-based explanation (or other method).
PentagonQ3.3	Geometry	Formal	1	C1F	Computation (primary focus) Standard representation; Definition; Problem solving; interpretation/reflection (secondary foci)	Identify properties of angles in a two-dimensional shape; compute the interior angle of a regular pentagon using a number-based strategy; provide correct answer with clear work shown.

Table 2-3: The ten most difficult items in the data set

Item	Mathematics content domain	Degree of formalization	Competency Class	Map Code	Indicator/s	Item Description
Baby Feeding Q1	Statistics	Informal	1	C1I	Standard representation (primary focus) Routine procedures (secondary focus)	Read information correctly from a chart.
PoolsQ1.2	Number	Informal	1	C1I	Routine procedures (primary focus) Standard representation; One method (secondary foci)	Read decimal numbers from a scale to determine height of figure; provide correct answer.
Ranger Station Q1	Statistics	Informal	1	C1I	Computation (primary focus) Routine procedures; One method; Interpretation/reflection (secondary foci)	Identify appropriate arithmetic calculation to determine mean; use whole numbers; provide correct answer.
Playground Q1	Number	Informal	1	C1I	Routine procedures (primary focus) One method (secondary focus)	Analyze a pattern; use counting to determine correct answer.
Nine Chips	Statistics	Informal	1	C1I	Standard representation (primary focus) Definition (secondary focus)	Identify appropriate graph to graphically represent statistical information; explain why the graph is appropriate.
Jill's Trip	Number	Informal	1	C1I	Computation (primary focus) Problem solving; Multiple well-defined methods (secondary foci)	Apply appropriate series of arithmetic calculations to solve multi-step problem, using whole dollar amounts to give answer in days or weeks.
Playgrounds Q1	Algebra	Informal	2	C2I	Interpretation/reflection (primary focus)	Interpret pattern demonstrated in diagrams; extend pattern; provide correct answer.

					Modelling (secondary focus)	
k + 6	Algebra	Preformal	1	C1P	Standard representation (primary focus) Definition; Interpretation/reflection (secondary foci)	Interpret algebraic expression; demonstrate understanding of the concepts underlying the terms 'variable' and 'infinity'.
Similar Triangles	Algebra	Informal	1	C1I	Routine procedures (primary focus) One method (secondary focus)	Count number of congruent triangles correctly for three figures in a sequence.
Stretch Q1	Algebra	Informal	1	C1I	Routine procedures (primary focus) Standard representation (secondary focus)	Identify pattern based on data in table or plotted in graph; use pattern to determine y-value for data point given x coordinate; provide correct answer within reasonable range.

Table 2-4: The ten least difficult items in the data set[7]

[7] Full credit responses for each item. Partial credit responses are included in (Shafer, Romberg, Wagner, 2005)

Item	Mathematics content domain	Degree of formalization	Competency Class	Map Code	Indicator/s	Item Description
Pyramids Q2.2	Geometry	Informal	1	C1I	Definition	Correctly name the shape of the base and the faces of a pentagonal pyramid.
Square/ Circle Area	Geometry	Formal	1	C1F	Computation (primary focus) Definition; Interpretation/reflection; multiple well-defined methods (secondary foci)	Apply π r2 to determine area of a circle, find difference from calculated area of square in multi step problem.
Points on a line	Algebra	Formal	1	C1F	Standard representation (primary focus) Modelling (secondary focus)	Identify a third point on a straight line, given the coordinates of two points on the line.
What is 'n'?	Algebra	Formal	1	C1F	Standard representation (primary focus) Definition; Interpretation/reflection (secondary foci)	Demonstrate understanding of abstract variable within context of a series of consecutive terms.
Batteries	Statistics	Preformal	2	C2P	Interpretation/reflection (primary focus) Computation; Definition; Routine procedures; Problem solving (secondary foci)	Use ratio in fractional form to infer sample size.
Perimeter Shapes	Geometry	Informal	1	C1P	Definition (primary focus) Computation; Problem solving (secondary foci)	Estimate perimeter of irregular polygons to identify polygon to fit specified criteria, where not all dimensions given.
Baby Feeding	Statistics	Preformal	2	C2P	Standard representation (primary focus)	Read a graph and a table; connect information from graph

Q3.2					Computation; One method (secondary foci)	with information in a table; identify appropriate arithmetic calculation (multiplication); use information from graph and table to calculate; provide correct answer with correct computation.
Pyramids Q6	Algebra	Preformal	1	C2P	Standard representation (primary focus) Computation; Definition; Routine Procedures (secondary foci)	Use given formula; substitute numbers found in a table for variables in a formula; determine if equation is valid; provide correct explanation.
Stretch Q4	Algebra	Preformal	2	C2P	Interpretation/reflection (primary focus) Standard representation (secondary focus)	Interpret meaning of y-intercept; use graph to draw conclusion; provide correct conclusion with clear supporting explanation.
CubesQ2.2	Algebra	Informal	1	C1I	Standard representation (primary focus) Computation (secondary focus)	Use a formula to compute; calculate with whole numbers; provide correct answer.

Table 2-5: Ten items of mid difficulty (from -0.12 to 0.18 logits)

Step 4 Identify proposed achievement bands based on item thresholds as determine by the data analysis

In order to describe achievement bands for the item set the items were divided into six levels based on item threshold. While it is recognised that 'the number of levels into which a continuum is divided is always somewhat arbitrary' (Masters & Forster 1996a) the six proposed divisions (bands) are consistent with the idea that 'equal distances on a map should represent equal increases in achievement' (Masters & Forster 1996a). The levels are defined as in Table 2-6.

Band	Threshold range categories (logits)
vi	2.0 or greater
v	1.0 to 2.0
iv	0 to 1.0
iii	-1.0 to 0
ii	-2.0 to -1.0
i	less than -2.0

Table 2-6: Proposed bands by threshold

All items (highest level of response where applicable) were allocated to a band according to item difficulty (Table 2-7).

Band	Number of items
vi	8
v	27
iv	45
iii	35
ii	13
i	1
Total	131

Table 2-7: Number of items per band

With the six bands thus defined, descriptions for each band were constructed in Step 5.

Step 5 For each band, examine item demands, and form descriptions of the band.

General trends for variables

General trends for the variables of competency class and degree of formalization for the items were identified as follows:

- Competency Class

A shift from a greater proportion of items with a focus in C1 in the lower/middle bands [i, ii, iii, and iv (with iii and iv being approximately equivalent for the percentage of items in C1 and C2)] to a greater proportion of items with a focus in C2 in the uppermost bands (v and vi). This trend is apparent in the 'Competency class and Indicator' column in Figure 2-4. The predominant competency class indicator/s for each band is also indicated in this column.

- Degree of Formalization

A shift from a predominance of Informal for bands i and ii to Preformal (bands iv and v), with an overlap of Preformal to more Formal items at bands v and then into vi. This trend is apparent in the 'Degree of formalization' column in Table 2-8.

Generic Descriptions for the Bottom-Up Progress Map

For each achievement band, examination of the item demands for the items located within each band revealed a set of common knowledge, skills and understandings students need to demonstrate to achieve at that level. Generic descriptions of each achievement level (see Table 2-8**)** were then formulated accordingly. The generic descriptions include both a general description of the achievement level, and some content domain-specific descriptions – one for each domain of Algebra, Number, Geometry and Statistics.

Achievement Band	Competency class and Indicator	Degree of formalization	*Generic Description*
Band vi	Items situated in this band are categorised as either C1 or C2, with C2 items predominating (85%). This achievement band is best represented by the indicators of *Interpretation/Reflection* (5 items) and to a lesser degree, *Problem Solving* (2 items).	Items situated in this band are categorised as Preformal or Formal (65.5% and 37.5% respectively), with Preformal items predominating. It should be noted however, that the number of items in the pool situated in this band is relatively small, so these figures may be misrepresentative of the band in this case. Informal items are not represented in this band.	A very high level of mathematical competence is required in order to respond fully to items in this band, as students are typically required to translate highly contextualised real world problems (where comprehension of a detailed written stimulus is required) into mathematical terms, and then identify and use an appropriate mathematical strategy and use a range of tools to solve the problems. In general, the demands that have to be met to achieve the highest level of response (full credit) for these items are such that students typically need to identify the key elements of the problem; show an extensive working solution; and provide a summative statement in response to the problem. Students are typically required to identify, compare or combine elements of the problem, draw on assumptions based on real-world knowledge, and provide a fully justified conclusion supported by working, explanation or reasoning. Content domain - specific demands typically required students at this level for: ▪ Algebra problems to substitute values appropriately for several variables in expressions, combine like terms and simplify. ▪ Number problems to interpret numerical data, use ratio and scale, and apply absolute (constant difference) or relative comparison (proportional change). ▪ Geometry problems to interpret spatial data to compare two- dimensional representations of 3D objects, involving different scales and perspectives. ▪ Statistics problems to identify an appropriate graph to represent statistical information and to explain why a particular graph is a suitable choice to support a conclusion.
Band v	Items situated in this band are categorised as either C1 or C2, with C2 items predominating (59%). This achievement band is best defined by the indicators of *Problem Solving* (10 items), and to a lesser degree, *Interpretation/Reflection* (5 items) and *Modelling* (1 item).	Items situated in this band are categorised as Informal (3.7%), Preformal or Formal , with Preformal and Formal co-dominating (44% and 52% respectively).	A high level of mathematical competence is required in order to respond fully to items in this band, as students are typically required to translate contextualised real world problems into mathematical terms, and then identify and use an appropriate mathematical strategy and use a range of tools to solve the problems. In general, the demands required to be met to receive full credit for these items are such that students typically need to provide a correct answer accompanied by a complete explanation of the working needed to arrive at the solution that takes into account the key points identified in the problem. The demands of the items in the band include the ability to interpret and to analyse, but the extent to which these skills are required is generally less compared to band vi. Content domain-specific demands typically required students at this level for: ▪ Algebra problems to interpret a pattern and then generalize the pattern, using variables to represent the pattern; interpret meaning of data points on a graph and use these to identify the slope, or to construct an equation to represent a linear graph; or construct an accurate graph of exponential increase. ▪ Number problems to identify appropriate arithmetical calculations; divide decimal numbers; use both whole numbers and fractions; or compare calculated ratios. ▪ Geometry problems to interpret spatial data to reason about the relationship between angle measures and side lengths; or find missing dimensions using the properties of similar triangles. ▪ Statistics problems to analyse and compare two sets of statistical data, where the data sets not presented in the same units (eg. Cost of x dollars per month, or Cost of x dollars per year).

Band iv	Items situated in this band are categorised as either C1 or C2, with C1 items predominating (62%). This achievement band is best defined by the indicators of *Computations* (15 items), and to a lesser degree, *Standard Representations* (5 items), *One method* problems (5 items), *Definitions* (3 items), and *Routine Procedures* (3 items).	Items situated in this band are categorised as Informal (27%), Preformal or Formal (22%), with Preformal and predominating (51%).	A moderate level of mathematical competence is required in order to respond fully to items in this band, as students are typically required to translate either a contextualised or a non-contextualised generally non –routine problem into mathematical terms. For contextualized problems, the solutions tend to depend on the application of a formula (eg. πr^2 for the area of a circle) or relationship (eg. proportionality of corresponding side lengths) with which it is expected that the student would be familiar. Non-contextualized items also tend to depend on the application of specific knowledge (eg. recognise an algebraic expression). Content domain-specific demands typically required students at this level for: ▪ Algebra problems to recognise the equivalent algebraic form of an expression (m + m + m + m = 4m); identify a third point on a straight line, given the coordinates of two points on the line. ▪ Number problems to use percent; use whole number division with remainder; add and subtract with integers. ▪ Geometry problems to find the area of an irregular shape; complete an accurate drawing of a 2D representation of a 3D object (eg. a pyramid with a pentagonal base). ▪ Statistics problems to construct a bar graph to represent statistical data; describe the concept of the mean.
Band iii	Items situated in this band are categorised as either C1 or C2, with C1 items predominating (63%). This achievement band is best defined by the indicators of *Computations* (10 items) and *Standard Representations* (7 items), and to a lesser degree, *Definitions* (4 items), and *Routine Procedures* (1 item).	Items situated in this band are categorised as Informal, Preformal (40%) or Formal (6%), with Informal predominating (54%).	Mathematical competence in this band is limited to the application of mathematical tools in order to solve predominantly routine problems, and where any contextualized problem-solving tasks also generally require only relatively simple computations for their solution. Content domain-specific demands typically required students at this level for: ▪ Algebra problems to interpret the meaning of the y-intercept; read coordinates of a point from a graph. ▪ Number problems to place a decimal on a number line; use whole numbers to compare metric measures (convert mm to L). ▪ Geometry problems to visualize a 3D object from its 2D net (cube); calculate the area of a circle given its diameter. ▪ Statistics problems to apply a ratio to the whole to determine an unknown quantity.
Band ii	Items situated in this band are categorised as either C1 or C2, with C1 items predominating (92%). This achievement band is best defined by the indicators of *Computations* (4 items), *Standard Representations* (3 items), and *Routine Procedures* (1 item).	Items situated in this band are categorised as Preformal or Informal (8%), with Preformal predominating (92%).	Mathematical competence at this level is limited to the application of routine procedures to standard or familiar contextual representations of problems. Content domain-specific demands typically required students at this level for: ▪ Algebra problems to interpret a simple algebraic expression; interpret a pattern presented in a series of diagrams. ▪ Number problems to read decimal numbers from a scale; use counting to analyse a pattern. ▪ Geometry problems to interpret a scale to estimate distance between two points on a map. ▪ Statistics problems to construct a sample space limited to the possible combinations of two selected items.
Band i	The one full credit item situated in this band is classified as C1, and addresses the indicator *Routine Procedures*.	The full credit items situated in this band is categorised as Preformal.	Mathematical competence at this level is also limited to the application of routine procedures to standard or familiar contextual representations of problems. Partial credit items in this band tend to be of the type in which less credit is given as only some of the required criteria are met (eg. for the construction of a bar graph: only one of `labelled axes, consistent scales, correct bar lengths and widths, labelled bars is met), or where a criterion is met, but an explanation is lacking.

Table 2-8: Generic descriptions of achievement bands

In Step 6, items were selected to illustrate the achievement bands. Items were selected on the basis that they illustrate the typical item demands for each achievement band.

Step 6 Illustrate descriptions of the bands with annotated items.

Items were selected on the basis of being representative of the typical item demands for a particular achievement band. Annotations were developed from analysis of the item characteristics (as described in Step 4) together with the item descriptions, marking rubrics (marking schemes), and direct examination of the stimulus material and associated items. In total, 25 items were selected to illustrate growth in mathematical competence from band i to band iv. The full credit responses to 10 items (PSA items) were selected for inclusion, and a further three partial credit items (PSA) were also selected to further enrich the progress map. To illustrate growth for each of the domain content areas, 12 EA items were selected (four items for each of Geometry, Algebra, Statistics and Number). Figure 2-5 is a pictorial 'snapshot' of the bottom-up progress map. The selected illustrative items are presented, with annotations and scoring rubrics, in Turner, O'Connor, and Romberg, 2004.

Further examination of items

A further small selection of items was also examined in detail, the purpose being to identify some other factors that appear to impact on the difficulty of items. Some recommendations for future refinements of the progress map follow. The majority of the items in band vi are classified as either Interpretation/reflection or as Problem solving (both C2 indicators). For several items, secondary indicators in C2 also apply: Modelling, Problem solving, and Interpretation/reflection (refer to Table 2-3 the eight most difficult items are situated in band vi). For one item, Keys Q3, aspects of Mathematical insight and Generalization (C3 indicators) also apply. For many of the items in this band, two of the competency class 1 indicators: Computation, and Definition also apply. However, the complexity of the computation or definition within the context of the item is challenging for students and/or students are required to synthesize several aspects of each of these indicators. The stimulus material for items in this band is often highly contextualized and requires students to interpret a written description of a complex problem situation (for example, Radio Station).

All of the items in band ii, (with one exception) have a primary focus in one of Standard representation, Computation, or Routine procedures (C1). In general, for these items one or more secondary indicators from C1 also apply. Where the items have a secondary indicator of Interpretation/reflection, the degree of interpretation or reflection is minimal. For example, for the item Blue and Yellow Balls, the required degree of interpretation of the scenario presented is relatively low:

> Steve was asked to pick two marbles from a bag of yellow marbles and blue marbles. One possible result was one yellow marble first and one blue marble second.

Generic Description For Achievement Band[8] (refer to Table **2-8**)	Illustrative items			Growth in Mathematical Competence
	Full credit response	Partial credit items	Content domain-specific items	
vi	**Keys[9] Q3** (2.28[10])	**Buildings Q2.2** (2.43)	**Radio Station** (2.43) *Geometry*	Highest level of competence
v	**Playground Q4.2** (1.82) **Stretch Q5.2** (1.36)	**Pentagon Q2.3** (1.56)	**Metro Rail** (1.52) *Statistics*	
iv	**Keys Q2.2** (0.55) **Playground Q2.3** (0.52)	Pentagon Q2.2 (0.54)	**Carla and Maria's Tiles** (1.08) *Number* **Dots** (0.28) *Algebra* **Square/Circle Area** (0.16) *Geometry*	
iii	**Stretch Q4** (-0.12) **Keys Q1.2** (-0.6)	**Birds of all sizes Q5.2** (-0.17) Pentagon Q2.1 (-0.72) Buildings Q2.1 (-0.91)	**Red/Blue cube** (-0.22) *Statistics* **Tip Calculation** (-0.32) *Number* **3(x + 5) = 30** (0.89) *Algebra*	
ii	**Playground Q1** (-1.4) **Playgrounds Q1** (-1.52)	Birds of all sizes Q5.1 (-1.25)	**Flour for cookies** (-1.03) *Number* **Oxford to Smithville** (-1.08) *Geometry* **Nine Chips** (-1.47) *Statistics* **k + 6** (-1.72) *Algebra*	
i	**Stretch Q1** (-2.53)			Lowest level of competence

Figure 2-5: Snapshot of Bottom-up Middle School Mathematics Progress Map

[8] Including trends in Competency class and Degree of formalization as in Table 2-8.
[9] Items in bold font indicate full credit response locations.
[10] Relative item difficulty (logits)

One item in this band, Playgrounds Q1, has a primary competency class indictor of Interpretation/reflection (rather than Standard representation or Computation). The worded explanation of a pattern in the stimulus material if read in isolation is relatively complex, but in this case it is supported by the inclusion of a series of diagrams, presenting the same information in graphic form. Therefore, the degree of interpretation required is also relatively low. The easiest item in the set, Stretch Q1, has a primary indicator of Routine procedures, and a secondary indicator of Standard representation (both indicators in C1). The procedure is a relatively straightforward one for students. The context presented is one likely to be frequently encountered by students in the classroom.

Items that were more difficult than expected

Some items were more difficult than anticipated according to their classification. For example, Baby feeding Q4 is classified as C1P (refer to Table 2-9), but at 1.76 logits is situated in band v.

Item	Mathematics content domain	Degree of formalization	Competency Class	Indicator/s	Item Description
Baby feeding	Number	Preformal	1	Computations (primary indicator) Interpretation/reflection; Multiple well-defined methods (secondary indicators)	Uses formula; converts one metric unit to another to provide correct answer with correct computation.

Table 2-9: Item classification for Baby feeding Q4

A closer examination of the item demand reveals that this item may be better classified as C2P as explained below. The students were required to use a rule (Each day, a baby needs 150 millilitres of milk per kilogram of its body weight) to calculate how many liters of milk a 13-year-old would need for each meal if milk were her only food. The students were told that the 13-year-old weighs about 40 kilograms.

> Kathleen, Chris' sister, is 13 years old. She wonders how much milk she would need if milk were her only food. Kathleen weighs about 40 kilograms.
>
> Using the rule above, calculate how many liters of milk Kathleen would need for each meal. Show your work.

The solution of this problem requires that several interrelated computations be performed (as shown in the scoring rubric below). The full credit (2 point) response also depends on the students initially providing an adequate assumption as to how many meals a day the 13-year-old would eat.

Scoring rubric: 2 point response
Correct answer: (depends on assumption) e.g. '1.5 L' (for 4 meals) or '2L' (for 3 meals), etc Correct computation shown: 40 X 150 ml = 6000 ml Correct conversion to liters: 6L (Note: correct conversion to liters can be implied from answer given in liters) Assumption for number of meals given: $2 \leq$ # of meals/day ≤ 6 e.g. 'she eats three meals per day' or assumption clear from computation (e.g. 6L divided by 3)

Interpretation/reflection and Multiple well-defined methods appear to characterize the nature of the task demands more appropriately than Computations in this case. Therefore, this item may be better classified in competency class 2, as both the indicators of Interpretation/reflection and Multiple well-defined methods are situated within C2.

Another item, Pyramids Q7, is classified as C1F (refer to Table 2-10). However, the full credit (2 point) response is 2.34 logits (band vi). The partial credit (1 point) response is also relatively difficult, being located at 1.56 logits (band v).

Item	Mathematics content domain	Degree of formalization	Competency Class	Indicator/s	Item Description
Pyramids Q7	Algebra	Formal	1	Standard representation (primary indicator) Computations; Interpretation/ reflection (secondary indicators)	Substitute expressions for variables in a formula; combine like terms; simplify to show equality; provide correct answer and correct explanation.

Table 2-10: Item classification for Pyramids Q7

In a series of related questions prior to Questions 6 and 7, students were required to complete activities related to the relationship between the number of vertices and edges for regular pyramids with bases having various number or sides (n). Students were required to complete the following table, including the last row for the generalization for 'n-gon'.

Pyramid base n-gon	Number of faces	Number of vertices	Number of edges
3-gon	4	4	6
4-gon	5		
5-gon	6		
6-gon	7		
7-gon	8		
12-gon			
n-gon	n + 1		

Students were then introduced to Euler's formula, before being required to complete Questions 6 and 7.

> There is a formula that describes the relationship between the number of faces (F), the number of vertices (V) and the number of edges (E) of figures.
>
> This is called Euler's formula and it is written as: F + V - E = 2.

Q6	Explain if Euler's formula works for the pyramid when n = 5. If Euler's formula does not work for the pyramid when n = 5, explain why it does not.
Q7	Show that Euler's formula works for the expressions in the last line of the table.

The solution of this problem (for full credit) requires that students either substitute formal algebraic terms into the equation directly or provide an explanation in words using algebraic terms. The solution is based on the pattern identified in previous questions, and requires students to show that Euler's formula applies for a generalized 'n-gon' pyramid. It is noted in the scoring rubric, that students

could also use expressions based upon a faulty pattern from a previous question, thus removing this aspect of dependency as a potential contributing factor in increasing item difficulty.

Scoring rubric: 2 point response

Correct equation simplified to clearly suggest a result of 2:
e.g., 'n + 1 + n + 1 – 2n = 2n – 2n +2 = 2'
or
clear and correct explanation in words or with variables:
e.g., 'If you add F and V you have 2n + 2 and E is 2n so 2 is left.'

Scoring rubric: 1 point response

Student uses variables from chart but work has one minor computation error so that equation does not result in the proper calculation of 2
or
Student uses variables from chart to write an equation that = 2 but does not simplify it so that a result of 2 is clearly suggested.

It would appear that the essential element in solving this series of questions is the ability to generalize. In Question 7, the generalization must be tested against a formal algebraic rule. This item (and perhaps several other items in this set) might better be classified in C3, since the indicator of Generalization clearly applies.

Items that were less difficult than expected

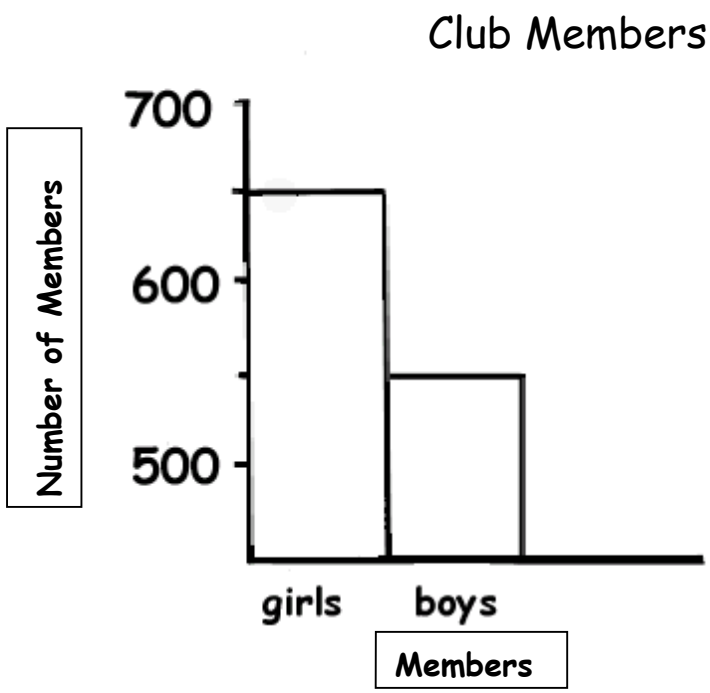
Some items were less difficult than anticipated from their classification. For example, Club members is classified as C2F (refer to Table 2-11), but at – 0.56 logits is situated in band iii (full credit response). The partial credit response is located at – 0.53 logits (band ii).

Item	Mathematics content domain	Degree of formalization	Competency Class	Indicator/s	Item Description
Club members	Statistics	Formal	2	Interpretation/reflection (primary indicator) Computations; Mathematical insight (secondary indicators)	Critically analyse graphical representation of data (bar graph); recognize misleading scale of y-axis; draw correct conclusion.

Table 2-11: Item classification for Club members

A closer examination of the item demand reveals that this item may be better classified as C1F. Students were required to interpret a bar graph; taking into account the nature of the vertical scale (i.e. the scale is discontinuous from zero).

Anika is a member of a swimming club. She made the diagram you see below.



Anika claims, "In our club there are two times as many girls as there are boys.Ó
Use the graph to explain why you agree or disagree with Anika.

The solution of this problem (full credit response) requires that students are able to read the graph correctly and to explain why Anika's statement is incorrect.

Scoring rubric: 2 point response

'If there were twice as many girls there would have to be 1100 girls because there are 550 boys'
or
'650 is not 2 X 550)'
or
'I disagree. There are 650 girls and 550 boys and that's not twice as many.'

or
'I don't agree with her because there are 650 girls and 550 boys and that's only 100 more girls than boys.'
Note: for full credit, students who explain why Anika's statement might be perceived as correct must also explain why Anika did not correctly read the graph.

A partial credit response to this item is one in which the student reads the bar graph correctly but the explanation of why Anika is incorrect is incomplete, inaccurate, or missing. It would appear that the ability to read the scale correctly is a relatively low level skill. Students find it more difficult to explain that the scale used can be misleading. However, the level of understanding required to explain that the scale used can be misleading is also relatively low. This item might better be classified in competency class 1, specifically addressing the primary indicator of Standard representations. A secondary indicator of Interpretation/reflection (C2) may be more appropriate than the indicator of Mathematical insight (C3) in this case.

It is important to note that some full credit items are appropriately located according to the assigned classification, and as expected, the partial credit responses are located in bands below the full credit response. Therefore, when locating items in the progress map, the reader should take care to establish if the location is for a response receiving full credit or partial credit. For example, the item Metro Rail is classified as C2F. The full credit response to this item (Metro Rail.3, scoring 3 points) is appropriately located in band v (refer to Turner, O'Connor, & Romberg, 2004, Illustrative items and annotations). The possible responses to this item span several bands of the progress map. Metro Rail.1 (scoring 1 point) is a partially correct response located in band ii.

Recommendations

In light of these observations, a number of directions for future refinement of the classification system used in developing the progress map are recommended:

- Descriptions of the indicators Computations and Definitions take into account that the degree of complexity of the computation required (or the level of understanding of the definition students need to bring to the item) varies with problem context. Item difficulty will be influenced by these factors.

- Descriptions of the indicator Interpretation/reflection take into account the degree to which the stimulus material is contextualized, and therefore to what extent students are required to interpret a written description of a complex problem situation; and that the degree to which interpretation is required will be reduced if accompanied by supporting explanations or diagrammatic representations. Item difficulty will be influenced by these factors.

- The selection of stimulus material takes into account the familiarity of the context to students. Item difficulty can be reduced for familiar contexts.

Concluding Remarks

The final bottom-up progress map (Figure 2-5) is an empirically-based refinement and enrichment of the initial top-down conceptualization of growth in mathematical competence. As noted earlier, a progress map continually evolves. This study provides a pictorial representation of the variable of 'mathematical competence' at this point in time. This variable itself has been shown to be defined by a unique combination of the three variables acting together to determine item difficulty: mathematical competency class; degree of formalization; and the associated indicators. Through the development of the progress map, the levels of achievement have been derived based on the typical set of knowledge, skills and understandings students demonstrate when they engage with the test items. An examination of the description of the achievement bands reveals the progress map to be consistent with both the PISA definition of Mathematization, and with Romberg's view of mathematical literacy.

References

Adams, R. J. and Khoo, S. T. (1996) *Quest: The Interactive Test Analysis System*. Camberwell, Australia: ACER.

Bond, T. G. and Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Mahwah NJ: Erlbaum.

Crocker, L. and Algina, J. (1986) *Introduction to Classical and Modern Test Theory.* Orlando, FL: Harcourt Brace Jovanovich.

Dekker, T., Querelle, N., van Reeuwijk, M., Wijers, M., Fejis, E., de Lange, J., Shafer, M. C., Davis, J., Wagner, L., Webb, D. (1997–1998). *Problem solving assessment system*. Madison, WI: University of Wisconsin.

Folgert, L., & Shafer, M. C. (2000) *Interrater reliability at the 1997–1998 scoring institutes.* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 21). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Folgert, L., & Shafer, M. C. (2001a) *Interrater reliability at the 1998–1999 scoring institutes.* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 22). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Folgert, L., & Shafer, M. C. (2001b) *Interrater reliability at the 1999–2000 scoring institutes.* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 23). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Forster, M. (in progress). *Text Difficulty and the Measurement of Reading Growth*. PhD Thesis, The University of Melbourne.

Greeno, J. G., Pearson, D. P., & Schoenfield, A. H. (1996). Research in cognition, learning, and development relevant to the National Assessment of Educational Progress. Menlo Park, CA: Institute for Research on Learning.

Masters, G. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika* 47 (2). 149-174.

Masters, G. and Forster, M. (1996a). *Progress Maps.* Camberwell, Australia: ACER.

Masters, G. and Forster, M. (1996b). *Developmental Assessment.* Camberwell, Australia: ACER.

National Center for Research in Mathematical Sciences Education & Freudenthal Institute (Eds.). (1997-1998). *Mathematics in Context*. Chicago, IL: Encyclopaedia Brittanica.

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

Organisation for Economic Co-Operation and Development (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment.* Paris: OECD Publications.

Organisation for Economic Co-Operation and Development (2002). *PISA 2000 Technical Report*. Edited by Adams, R. J. and Wu, M. L. Paris: OECD Publications.

Petersen, N.S., Kolen, M.J. and Hoover, H.D. (1989). Scaling, norming and equating. In: R.L. Linn (Ed.), *Educational measurement (3rd ed., pp. 221-262).* New York: American Council on Education and Macmillan.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute of Education.

Romberg, T.A. (2000) *Changing the teaching and learning of mathematics* in Teaching and learning in world mathematics year 2000: exploring the possibilities. VC 2000

Romberg, T.A. (2001) *Designing Middle-School Mathematics Materials Using Problems Set in Context to Help Students Progress From Informal to Formal Mathematical Reasoning.* NCISLA, Wisconsin Center for Education Research.

Shafer, M. C., Romberg, T. A., & Wagner, L. R. (2005). *Problem solving assessments*. (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 15). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Stone, M.H., Wright, B. D. and Stenner, A.J. (1999). Mapping Variables. *Journal of Outcome Measurement* 3 (4), 308-322.

Turner, R., O'Connor, G., & Romberg, T. A. (2004) *Scale for Mapping Progress in Mathematical Competence.* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Tech. Rep No. 49). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Verhage, H. & deLange, J. (1997, April). Mathematics Education and Assessment. *Pythagoras, 42,* 14-20.

Webb, D. C., Romberg, T. A., & Shafer, M. C. (2001). *Year 1 student performance on an assessment using NAEP and TIMSS items for program evaluation*. (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 16). Madison, WI: Wisconsin Center for Education Research.