The Development of a Numeracy Achievement Scale |

to Assess Progress from Kindergarten Through Year 6

Joanne Mulligan
Centre for Research in Mathematics and Science Education (CRiMSE) Macquarie University, Sydney
joanne.mulligan@mq.edu.au

Juho Looveer & Susan Busatto
NSW Department of Education & Training
Juho.Looveer@det.nsw.edu.au

Abstract

A Numeracy Assessment Instrument K-6 (NAI) and a Numeracy Achievement Scale (NAS) using Rasch measurement were developed to monitor students' numeracy growth in the Numeracy Research in NSW Primary Schools Project. A sample of 1900 students from 20 Case Study schools, 10 Trialling schools and 10 Reference schools was assessed using the NAI. Student achievement as measured on the NAS showed that there were no significant differences in numeracy growth over a six-month period between different school groups. However, significant differences were found between students in different year levels. The NAS was also used to demonstrate that students from eight Trialling schools demonstrated greater than expected numeracy growth compared with their counterparts in Reference schools over an 18-month period.

Key Words: Assessment; Numeracy; Numeracy achievement Scale; Primary school; Rasch measurement

As a national priority, the Australian Government implemented a Numeracy Research and Development Initiative in support of the National Literacy and Numeracy Plan that provided a coherent framework for achieving improvement in student outcomes in literacy and numeracy (1997). Under this initiative the NSW Department of Education and Training, the Catholic Education Commission NSW, and the Association of Independent Schools of NSW participated in a three-year *Numeracy Research in NSW Primary Schools Project* [1] (2001-2003) focussed on two broad research questions:

- what are the educational practices which are 'making a difference' in enabling primary school students to achieve 'outstanding' numeracy learning outcomes?
- to what extent, and in what ways, can such educational practices be successfully applied to other school contexts?

This large-scale broadly focussed research project was designed to identify and describe outstanding numeracy programs, policies, processes and strategies that would support numeracy learning for all students. This involved the analysis of effective practices in numeracy at 45 Case Study schools and the application of some of these practices in ten Trialling schools aimed at improving their numeracy profile (Commonwealth of Australia, 2004). One measure of this process was the development of a *Numeracy Assessment Instrument (NAI)* and a *Numeracy Achievement Scale (NAS)* designed to monitor student numeracy growth. This paper describes the development and implementation of the NAI and NAS using Rasch measurement over three phases of the study.

Background to the Study

Recently, the move to evaluate change through ‘value-added’ educational programs has called for psychometric evidence of reliable and valid ways of describing growth in student achievement between critical points in the evaluation of learning outcomes (Lokan, Doig & Underwood, 2000). Within the NSW Department of Education and Training, Rasch analysis (Rasch, 1980) has been employed to construct scales of achievement for students’ results in the Basic Skills Tests (BST) Numeracy and Literacy in Years 3 and 5 (NSW Department of Education & Training, 2001a), the Secondary Numeracy Assessment Program (SNAP) (NSW Department of Education & Training, 2001b), and the English Language and Literacy Assessment (ELLA) (NSW Department of Education & Training, 2001c).

Methodology

There were six main phases in the design of essentially a two-year study, incorporating 45 Case Study schools, ten Trialling schools and ten Reference schools. Eight of the initial ten Trialling schools were monitored though an additional phase of the project in 2003.

The NAI was developed and trialled in 2001 with 2832 students from Grades K-6 in 51 NSW schools. It was constructed with two forms of seven interview-based assessment schedules (Kindergarten to Year 6). Items incorporated critical aspects of Space, Number, Measurement and Data, which reflected the NSW Mathematics K-6 syllabus (Board of Studies NSW, 2000; NSW Department of Education, 1989).

In order to monitor measures of numeracy growth, two classes of students, matched by grade level, were assessed from each of the 20 Case Study Schools, ten Trialling Schools and ten Reference Schools (to enable a comparative measure). This comprised a sample of 1900 students (K to Year 6) assessed through the NAI in Term 1 and Term 3, 2002.

The project was extended in 2003 to enable the project to determine the extent to which the changes in pedagogical and other practices initiated in the ten Trialling Schools in 2002 were sustained within each school in 2003. Eight of the ten Trialling Schools continued in 2003 where 276 students were reassessed in Term 1 and Term 3.

The Development of a Numeracy Assessment Instrument

In order to develop a numeracy construct that reflected syllabus outcomes and a broad measure of numeracy, critical aspects of mathematical, contextual and strategic knowledge needed to be identified from the outset. Furthermore, it was advantageous to develop a numeracy scale that measured and described students' development of these key aspects as well as representing key developmental features in students' strategies.

The use of Rasch modelling

The Numeracy Research in NSW Primary Schools Project required a methodology that produced measures of students' ability beyond descriptive analyses of raw scores and percentages correct. It also required the development of a Numeracy Achievement Scale that developed appropriately graded items along a continuum for students aged 4.5 to 13 years, representing key aspects of numeracy across all strands of the mathematics curriculum. This required the construction and integration of a large number of items drawn from a number of sources. In order to establish the integrity of these items as a measure of numeracy, it was essential to translate these items onto a linear scale.

The Rasch Simple Logistic Model (1980) is the simplest form, mathematically, of

Item Response Theory (IRT), and was selected as the means to create the NAS. The Rasch model allows the construction of an interval scale that enables assessment against consistent standards from time to time (e.g. the Higher School Certificate of the Board of Studies, NSW) or to assess growth (e.g., Basic Skills Testing by the NSW Department of Education and Training). Rasch analysis produces separate measures for student ability and for item difficulty both on the same scale. This scale of achievement is independent of age and grade.

Consequently students can be located on a scale according to their performance, based on the number of items they answer correctly. The degree to which this score summarises the individual's profile of responses is assessed by identifying the "fit" of the students' response pattern to the model.

The main advantage of using Rasch analysis for constructing the NAS was that it could be used to determine whether any growth had occurred in numeracy achievement between two points of time. In order to measure this growth, ability estimates can be made of students' location on the continuum and changes in students' ability locations provide measures of growth.

The Numeracy Assessment Instrument (NAI)

The intention was to design a numeracy instrument that enabled students to demonstrate conceptual understanding. The interview method was essential for those students who were not yet able to complete a paper and pencil test. Other measures of numeracy achievement had been focused on particular grade levels or stages using pen and paper tests, multiple-choice methods, or were limited to specific content areas. There were no existing suitable assessment tools available to assess numeracy across K-6 using practical materials in a clinical interview and to monitor growth.

The individual interview also allowed for the use of practical equipment, whereby student solution strategies could be observed and recorded. This eliminated the possibility of the response being coded as correct when a student had guessed an answer or had given a response that had an incorrect mathematical basis. For example, in Figure 1 (Trial Form 6.2, Item 8), students were given a drawing of the diagram and squares to cover the rectangle followed by this question.

> *Can you work out how many 2 cm squares would be needed to cover a rectangle 8 cm wide and 10 cm long? It might help to draw the rectangle first.*

Figure 1. Trial Form 6.2, Item 8

Students giving a correct response would need to demonstrate an understanding that they were measuring the inside area of the rectangle by imagining or placing the squares on the rectangle and counting the total.. This would indicate to the interviewer any confusion with the concepts of perimeter and area.

Tasks were also designed in such a way as to determine whether a student had used an appropriate strategy or not. For example, in Figure 2 (Trial Form 5.2, Task 13), the purpose of the task was to assess mental approximation skills and number sense. Where students could only add the numbers and where no attempt to round off or approximate was shown, then the response was coded as incorrect, even if the answer was technically correct. Mental computation was being assessed as reliably as possible.

Show two cards with 799 and 809 on them. (Side by side, do not put underneath each other)

799	809

We don't want you to work out the exact answer. If you added these numbers would the total be closer to 1500 or 1600? How did you know?

Figure 2. Trial Form 5.2, Task 13

Recent advances in mathematics education have encouraged children to develop a range of strategies to solve mathematical problems, and to reflect upon, justify and explain their strategies for solving a problem. This has been widely accepted practice in research studies involving assessment of students' mathematical understandings and recent assessment schedules have included assessment items that require students to justify and explain their responses in writing. This process has been a focus area for curriculum reform in mathematics spanning Years K-12.

Item Specification

Items were designed in accordance with strand organisers that reflected the established mathematics syllabus (NSW Department of Education, 1989): Number, Space and Measurement. Aspects of Data were included in the strand Space. It was not possible to include finely graded examples of each aspect so as to represent sub-strands because each trial form had a limited number of items to be included. It was estimated that trial forms for Kindergarten and Year 1 students should not exceed 16 tasks and could be administered within 10 to 15 minutes. For Years 2 and 3, the number of tasks was increased to 20, and for older students (Years 4, 5 and 6), the number of tasks was extended to 25 in total.

Trial Forms

The Numeracy Assessment Instrument comprised a total of 244 numeracy items categorised by the strands of Number, Space and Measurement. The items were divided into two trial forms per Year level. Link items were created between trial forms within each Year and across adjacent Years. For example, there were four common items (25% of 16 tasks) in test form K.1 and test form K.2, and there were four common items across the Kindergarten and Year 1 trial forms. This was to enable the creation of a common scale.

The NAI Trialling Process

The NAI was trialled from May to December 2001 with 2832 students from Grades K-6 in 51 Government, Catholic and Independent schools. These schools reflected a broad range of categories: socio-economic indicators; proportion of Aboriginal population; language backgrounds other than English (LBOTE); rural, isolated and remote areas and inner and outer metropolitan areas. The interviews were conducted at each school by the same interviewers to ensure consistency in approach.

The Numeracy Assessment Instrument Item design

The NAI aimed to reflect a framework where items assessed specific aspects but which were interrelated with other general problem-solving processes, such as being able to explain and justify one's thinking. This included key aspects of numeracy considered critical to the development of number (e.g. figures 3 and 4).

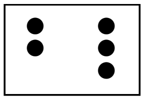
Show the student the number 1067 on the calculator.

Multiply this number by 1000 without showing the student.

Show the new number, 1 067 000, to the student.

Figure 3: Trial Form 5.1, Task 3

Flash, for 1 second, the card which shows a pattern of 5. Turn it over.



How many dots did you see?

Figure 4: Trial Form K.1, Task 2

The ability to identify the next element in a numerical or spatial pattern was also considered important for inclusion in the assessment instrument shown in Figure 5.

Here is a pattern made from some squares. (Show card).

Can you make the next part of the pattern? (Provide cardboard squares if necessary).

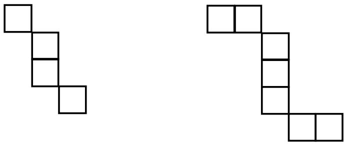


Figure 5: Trial Form 5.1, Task 15

The development of efficient mental calculation strategies, based on informal methods, has emerged as a new priority for numeracy. This is in sharp contrast to a traditional emphasis on written algorithmic procedures. Thus, assessment items were devised to assess mental computation, estimation and approximation skills based on mental computation assessment instruments developed in Australia (Callingham & McIntosh, 2002). This included items requiring students to understand whole number operations, fractions, decimals and percentages. An example is provided in Figure 6.

Show two cards with 799 and 809 on them. (Side by side, do not put underneath each other)

799	809

We don't want you to work out the exact answer. If you added these numbers would the total be closer to 1500 or 1600? How did you know?

Figure 6: Trial Form 5.2, Task 13

In the design of items involving fractions and decimals, an emphasis was placed on the development of conceptual understanding of fractions and the notion of simple ratio and rate. Figure 7 shows an example of partitioning that assesses early fraction concepts.

Show the card with the bars below and say:
How would you share these 2 chocolate bars fairly among 4 children?

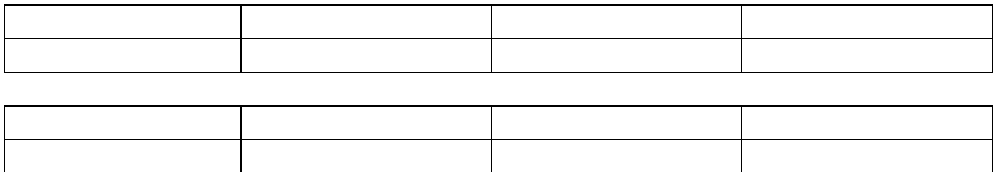


Figure 7: Trial Form 2.2, Task 12

2 balls are placed in a bag (out of view). Ask the student to hold one ball in each hand. Ask students to close their eyes and heft the two amounts.

Say: Pull out the hand which is holding the heaviest ball.
(Ensure that the smallest ball is heavier, and the larger ball is lighter.)

Figure 8: Trial Form 1.2, Task 13

Show 3 cards

cm centimetre	L litre	kg kilogram

Give me the card that you would use to measure how heavy someone is?

Figure 9: Trial Form 3.2, Task 5

Show drawing of a rectangle.

Explain how you could work out the area of this shape
If child says length x breadth, say: *"Tell me what you are actually measuring"*.

Figure 10: Trial Form 6.1, Task 23

Some tasks involving time concepts were included. Young students were required to draw a common time (8 o'clock), on a blank analogue clock face that showed only the numerals 1-12. Other tasks included the sequence of days of the week and understanding basic units of time.

Data analysis and initial construction of the Numeracy Achievement Scale

Each student was assigned a code to ensure confidentiality. Responses were coded as *correct* or *incorrect* by the interviewers on the recording sheets. A partial credit model (Wright & Masters, 1982) was initially applied to allow for polytomous items (i.e. marks were allocated for each response, and questions were thus scored out of several marks). Several of these polytomous items displayed reverse thresholds and lack of fit to the model, hence these items were recoded to series of dichotomous items, and the simple Rasch model applied for all further analyses.

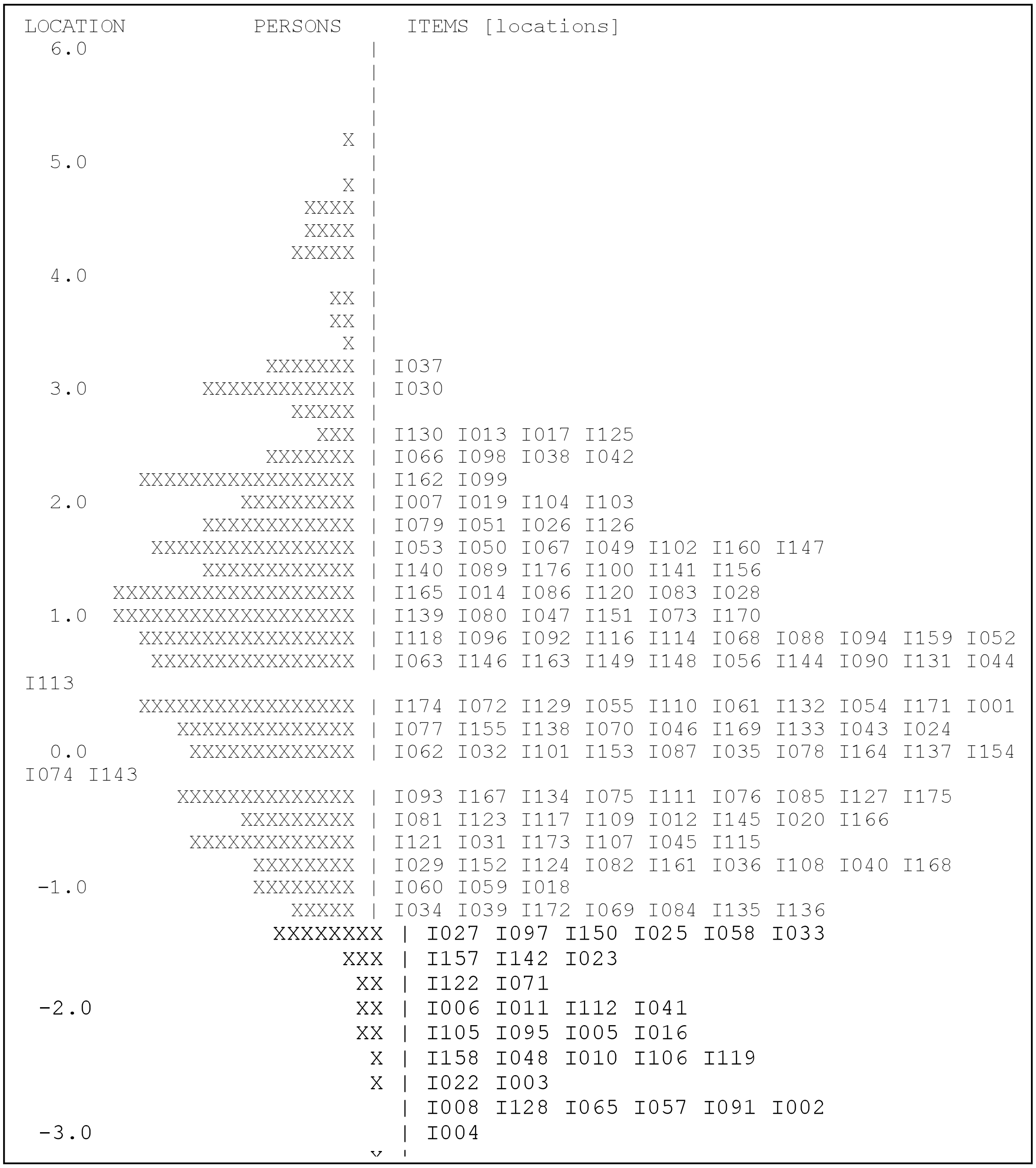
Rasch analysis of these trialling data was undertaken using Rasch Unidimensional Measurement Models (RUMM) computer software. RUMM was used to generate scale scores for items and student measures for the construction of the Numeracy Achievement Scale (NAS). All items were calibrated concurrently, with 61 items linking the 14 forms used. This placed all items and students on a single scale using the same metric, the "logit". Item and person estimates were calculated to ascertain "fit" measures. Item maps were produced by the analysis to show the distribution of all items and students along the "numeracy" variable. The numeracy construct was further refined by producing item maps by strand: Number, Space and Measurement.

Parallel analysis was conducted using QUEST (The Interactive Test Analysis System) software (Adams & Khoo, 1999). This was then compared with the analysis using RUMM that was then corroborated by QUEST analysis. An analysis of variance (ANOVA) was conducted on student ability measures to assess variability between Years (grades) and gender. These analyses were used to discard items not functioning well, and to determine the final item bank for the assessments planned for 2002-2003.

The use of descriptors for each item in the NAI was located along the initial scale as a means of examining more closely the types of knowledge and skills shown by the students. These descriptions provided essential information for making decisions about including additional items and/or eliminating items to produce one assessment schedule per Year level for the Numeracy Assessment Instrument and the construction of the Numeracy Achievement Scale in 2002. Following Rasch analysis of the 242 tasks, 66 tasks that did not conform to the uni-dimensional construct of *numeracy* for this project were discarded. The NAS was determined by the Rasch item difficulties from the remaining 176 tasks.

Assessment of students using the NAI commenced in April 2002. A new group of schools and students was selected for the main study. Once all students had been assessed, the data were analysed using RUMM software to produce estimates of student ability. This analysis was intended to define the final NAS, according to the final item pool. The scale would be defined by the locations of the items (i.e. the item difficulties). These item locations could be applied in any future analysis of student assessment data using the NAI to produce estimates of student ability.

An item map (see Figure 11) shows how student ability scores and item locations are placed on the scale (NAS). The right-hand side of Figure 1 shows a map of 176 item locations. Item 1037 (3.2 logits) is the hardest item and Item 1009 (-5.0 logits) the easiest. The left-hand side of Figure 1 shows a map of 2832 student locations. Each X represents 9 students located on the map. The scale extends from approximately –5.0 logits to 5.2 logits, representing students' ability from Kindergarten to Year 6. When students who respond correctly (or incorrectly) to all items presented to them, scores are interpolated from other data for these students.



```
LOCATION          PERSONS      ITEMS [locations]
  6.0                        |
                             |
                             |
                             |
                           X |
  5.0                        |
                           X |
                        XXXX |
                        XXXX |
                       XXXXX |
  4.0                        |
                          XX |
                          XX |
                           X |
                     XXXXXXX | I037
  3.0           XXXXXXXXXXXX | I030
                       XXXXX |
                         XXX | I130 I013 I017 I125
                     XXXXXXX | I066 I098 I038 I042
           XXXXXXXXXXXXXXXXX | I162 I099
  2.0              XXXXXXXXX | I007 I019 I104 I103
                XXXXXXXXXXXX | I079 I051 I026 I126
            XXXXXXXXXXXXXXXX | I053 I050 I067 I049 I102 I160 I147
                XXXXXXXXXXXX | I140 I089 I176 I100 I141 I156
         XXXXXXXXXXXXXXXXXXX | I165 I014 I086 I120 I083 I028
  1.0    XXXXXXXXXXXXXXXXXXX | I139 I080 I047 I151 I073 I170
           XXXXXXXXXXXXXXXXX | I118 I096 I092 I116 I114 I068 I088 I094 I159 I052
            XXXXXXXXXXXXXXXX | I063 I146 I163 I149 I148 I056 I144 I090 I131 I044
I113
           XXXXXXXXXXXXXXXXX | I174 I072 I129 I055 I110 I061 I132 I054 I171 I001
              XXXXXXXXXXXXXX | I077 I155 I138 I070 I046 I169 I133 I043 I024
  0.0          XXXXXXXXXXXXX | I062 I032 I101 I153 I087 I035 I078 I164 I137 I154
I074 I143
              XXXXXXXXXXXXXX | I093 I167 I134 I075 I111 I076 I085 I127 I175
                   XXXXXXXXX | I081 I123 I117 I109 I012 I145 I020 I166
               XXXXXXXXXXXXX | I121 I031 I173 I107 I045 I115
                    XXXXXXXX | I029 I152 I124 I082 I161 I036 I108 I040 I168
 -1.0               XXXXXXXX | I060 I059 I018
                       XXXXX | I034 I039 I172 I069 I084 I135 I136
                    XXXXXXXX | I027 I097 I150 I025 I058 I033
                         XXX | I157 I142 I023
                          XX | I122 I071
 -2.0                     XX | I006 I011 I112 I041
                          XX | I105 I095 I005 I016
                           X | I158 I048 I010 I106 I119
                           X | I022 I003
                             | I008 I128 I065 I057 I091 I002
 -3.0                        | I004
                           X |
```

Figure 11. Item Map from initial analysis.

This initial item map suggested that the items and students were well matched.

To better understand the NAS, student ability was considered according to students' Year levels. The results of this are presented in Figure 12 below. Each marker on the graph represents an individual student's ability measured in logits. Whilst the scores in each grade level are spread fairly well, there appeared to be much more overlap between the student abilities within adjacent years, as well as across the whole sample. The lowest scoring students were in Years 2 and 4, but this could be due to student effort for such assessments. However, there were many students in Years 1 and 2 who had ability scores well above the majority of year 5 and 6 students. This was not consistent with face validity or common sense.

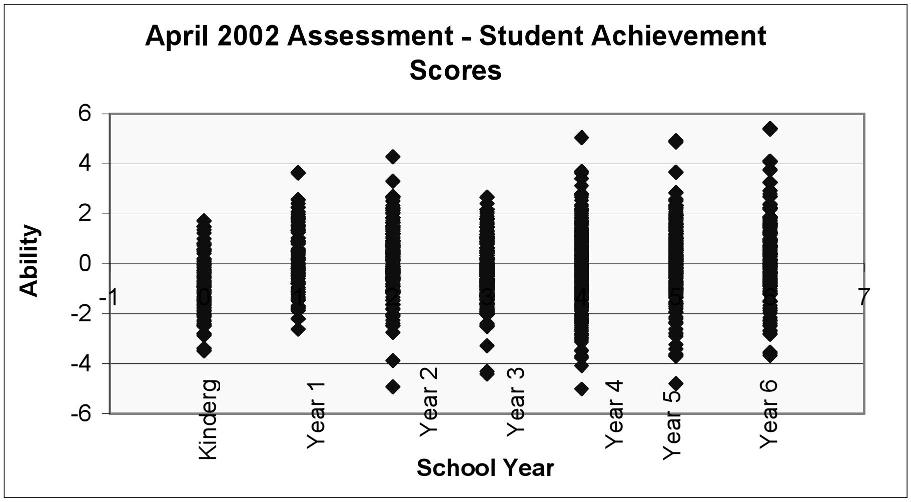


Figure 12. Student ability by Year level – initial analysis.

An analysis of the item locations on the scale by Year level (see Figure 13) revealed similar problems–The most difficult items were from the Year 2 schedule, and there was much more overlap of item locations across the increasing grades.

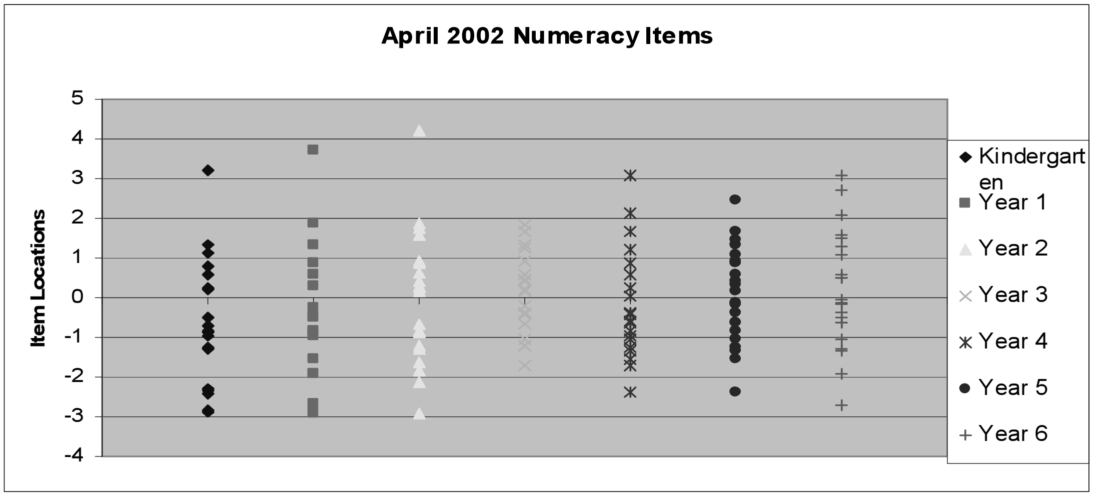


Figure 13. Item locations by year level.

Data entry and coding was checked and it was concluded that there had been no errors with original data and analyses. To investigate this issue further, items were analysed for differential item functioning (DIF) according to Year level. DIF is evident for an item when two groups of students, matched for ability, perform differently on an item. In essence, two students with similar ability have different probabilities of responding correctly to an item.

Differential Item Functioning (DIF)

One of the outcomes of Rasch analysis is Item Characteristic Curves (ICCs) (see Figure 4). The horizontal axis represents the scale of student abilities; the vertical axis represents the probability of a correct response. The ICC shows the probability of (or the expected score for) a student with a certain ability responding correctly to an item. RUMM analysis can extend these ICC graphs to assist with DIF analysis (Looveer, 2004).

Figure 14 shows the theoretical ICC for this item as well as separate graphs for students from Kindergarten and from Year 1 undertaking the same item. The two graphs are close along the ability range for these two groups of students. This item would be considered to display little or no DIF. Thus this item is fair for both groups of students as it operates in the same way for all students. For example, in Figure 14, for a student of ability around 1.8 logits (along the horizontal scale), the probability for responding correctly is around 0.5, irrespective of which year group they are in.

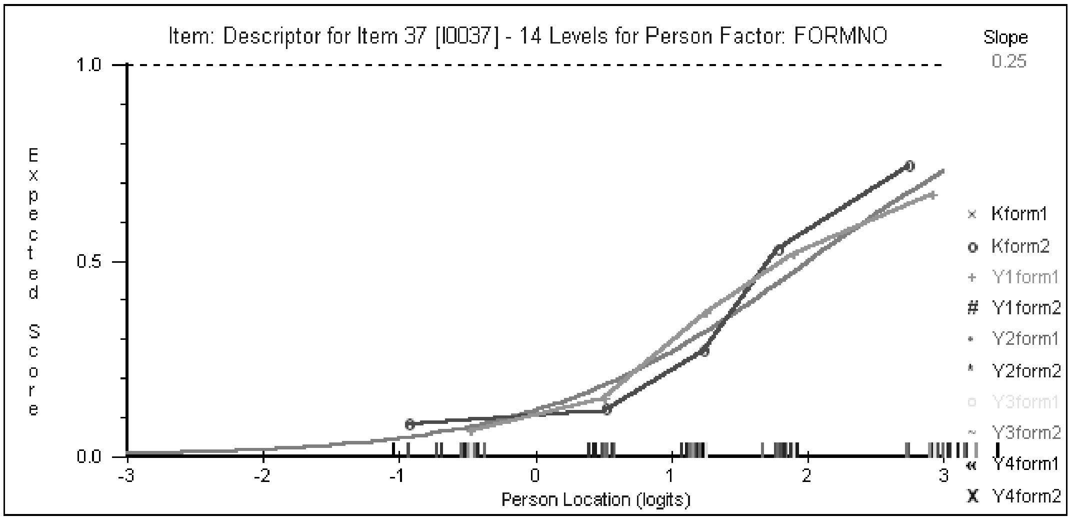


Figure 14. Item Characteristic Curve for Item 37, showing no DIF.

Item 37: A boy had some stickers. His friend gave him 5 more. Now he has 9 stickers. How many stickers did the boy start with? (Screen stickers with a card. Put 5 under the screen, then show 9 by removing card.)

It would be expected that most students would have been exposed to this item and the requisite skills. Hence, it would be expected that differences in performance on this item were due mainly to differences in ability.

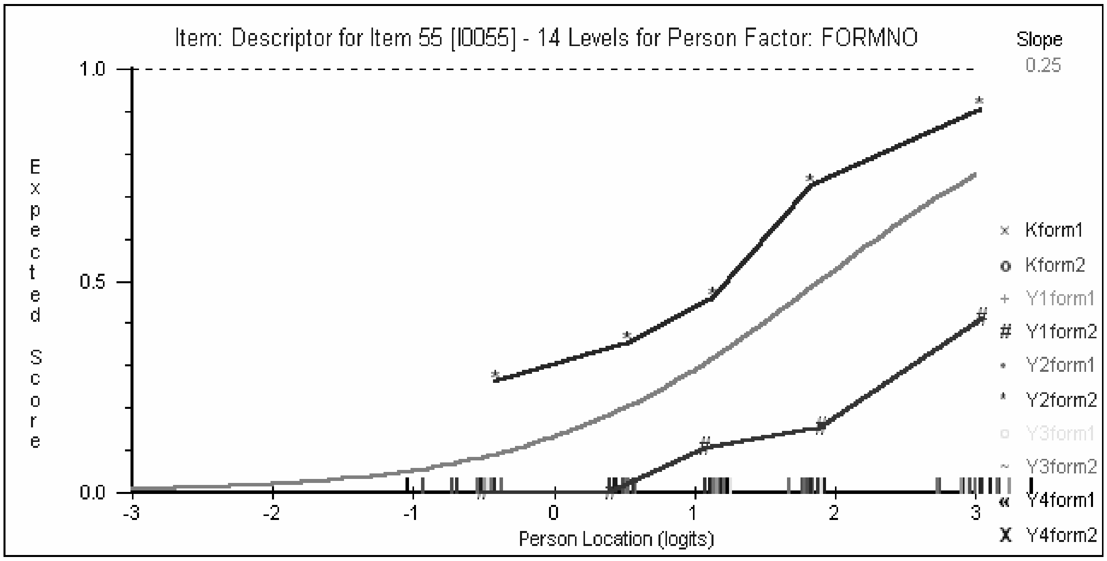


Figure 15. Item Characteristic Curve for Item 55, displaying significant DIF.

Item 55: Show the card: 2, 12, 22, ,

(a) What comes next in this pattern?

(b) Can you describe what is happening in the pattern?

In contrast, Figure 15 shows the ICCs for students from Year 1 and Year 2 for an item that displays large DIF across the ability range. For example, a student from Year 1 with an ability of around 2 logits would have a probability of around 0.1 for responding correctly to this item, whereas a student from Year 2 with similar ability would have a probability of around 0.75 for responding correctly. Thus this item functioned differently for the two groups of students, and was not equally fair for two students with similar abilities.

This item tested a student's ability to count by tens. It may be that students in Year 2 had been introduced to this skill and had practice, whereas students in Year 1 had not; this would be the cause of the differential performance on this item, rather than a difference in ability.

It was inferred that an item which displays significant DIF (as per the ANOVA produced in RUMM) was not suitable to be used as a link item across different test forms, because it performs differently for different groups of students. Nothing was apparent in the research literature regarding this type of problem. However, Linacre advised (personal communication, 2002) that since the purpose of link items is to tie together two different tests, they themselves must not change their characteristics between the tests. This is in reference to item locations on the underlying latent variable, i.e., construct stability.

Within-year DIF (i e DIF between groups in the same year-level) should not occur because the item should be the same for all students at the same instructional level. Where there is DIF across different year-levels, this is evidence of learning effects; in this case the item has become a new, probably easier item. It has limited value as a link item because it has changed its characteristics. In this case, the item really needs to enter into the analysis as two separate items, "before instruction" for one year group, and "after instruction" for the other year group.

Hence it was decided to revise the data structure by removing links for items displaying large DIF. This meant reanalysing the data to create a new Rasch-based scale.

Revision of the NAS

Due to the perceived problems with linking items, data collected in October 2002 was analysed using as links only those items that did not display large DIF. Consequently, as this analysis was intended to produce a Numeracy scale that overcame the original problems of too little difference between differing year-levels, it was agreed that this strategy, if successful, would define the NAS. Hence the definitive scale was constructed from the October 2002 assessment data.

The data from April 2002 was subsequently rescaled using the item locations from the final scale. Link items were as defined from the October 2002 analysis. Therefore some items that appeared in the assessment schedules for two adjacent year groups were treated as separate items for each group.

These results were then inspected more closely. Figure 16 below shows student abilities by year levels for the April 2002 assessment, according to the final scale. Compared to the original analysis (see Figure 12), the results displayed greater differences across year groups, displaying increasing ability for different cohorts. This is consistent with developmental progression of numeracy skills from Kindergarten to Year 6 and indicates strong face validity of the scale.

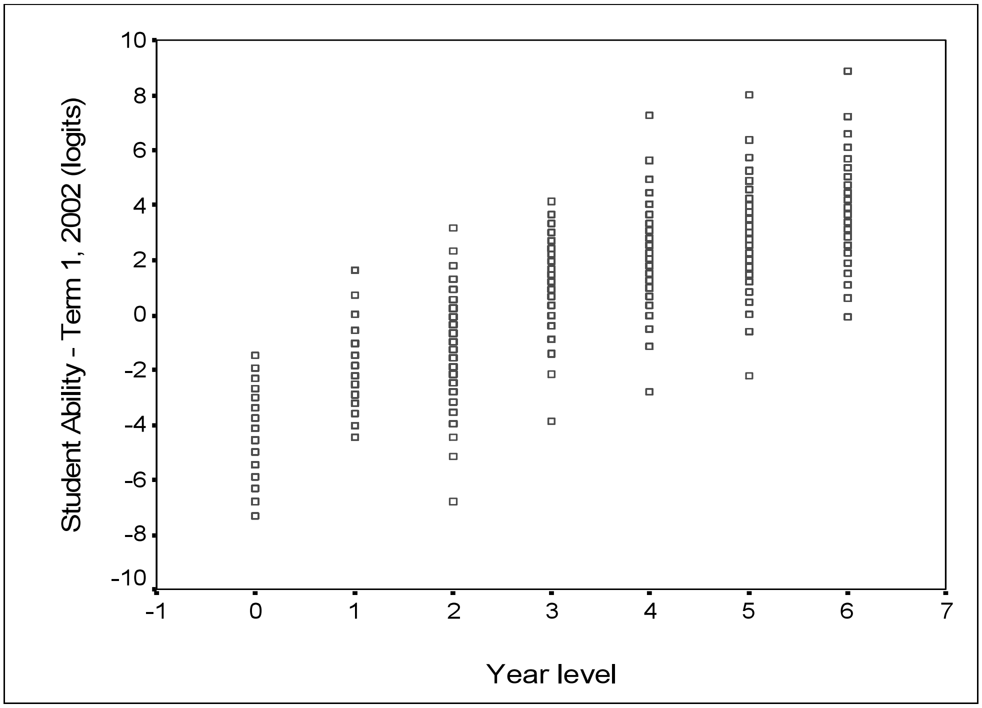


Figure 16. Student ability by year level–final scale.

The item locations were also assessed according to the year levels (Figure 17). This indicated stronger face validity, ion that there was a more hierarchical distribution of item locations according to the targeted year levels.

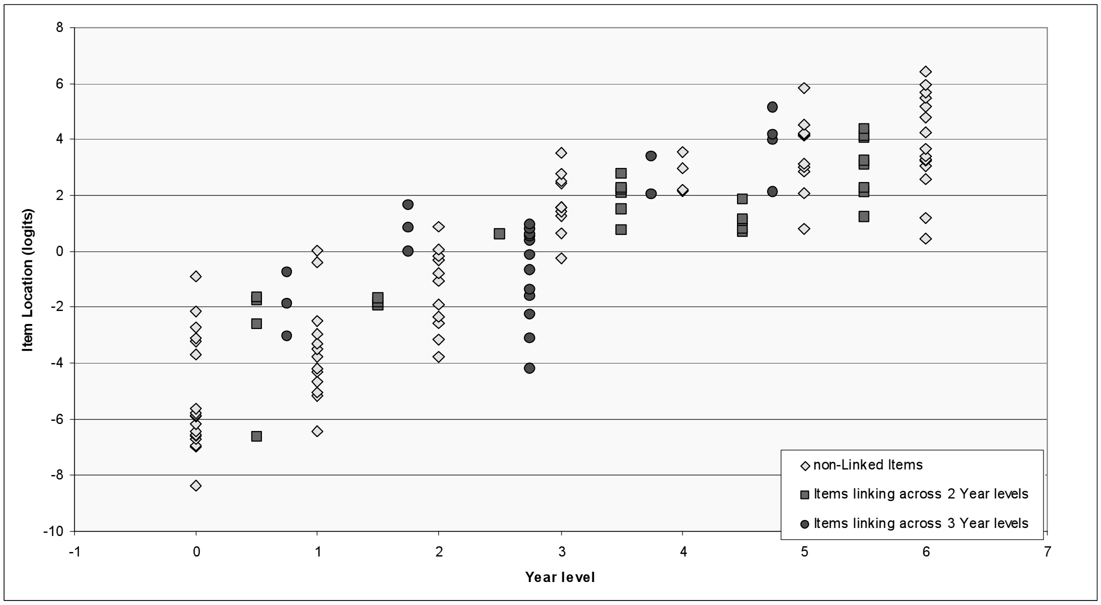


Figure 17. Item locations by year level – final scale.

The Final NAS

Figure 20 shows the item map for the final NAS, generated for all students (n=1705). The scale shows each item coded with a letter and a numeral. The right-hand side of Figure 18 shows a map of 139 item locations for the Term 3 analysis of data. Item I0138 (6.4 logits) was the hardest item and Item I0010 (-8.4 logits) was the easiest. This occurred after the equating of items and after the easiest item was eliminated. The left-hand side of Figure 18 shows a map of 1705 student locations. Each X represents four student locations on the map. The scale extends from approximately –6.0 logits to 9.0 logits, representing students' ability from Kindergarten to Year 6.

The item map shows that the difficulty of items did not sufficiently challenge approximately 64 students measuring from 6.6 logits to 9.4 logits on the scale above the hardest item. At the lower end of the map there were 9 items below that of the least able student (-6.2 logits).

The most difficult Number item (item number I0138) required students to find a percentage of an amount. This task was included in the Year 6 Term 3 assessment, and required students to find 10% of $157. The easiest item overall was a Number item (item number I0010), which required students to identify the number 10. This was the easiest item after four items were removed for construction of the scale. These four items were eliminated since all students answered them correctly.

The most difficult Space item (item number I0116) was included in the Year 5 Term 3 assessment. Students were shown a pie graph, of which half was coloured red, a quarter was coloured black, an eighth was coloured yellow and another eighth was coloured pink. Students were required to indicate the percentage that was coloured black. The easiest Space item (item number I0003) was included in the Kindergarten Term 3 assessment. The task required the student to place an object (a teddy) in front of a given object.

The two most difficult Measurement items (item numbers I0120 and I0137) were included in the Year 5 and Year 6 Term 3 assessments respectively. The hardest Year 5 task required students to determine the number of 2 cm by 2 cm squares which would be needed to cover a rectangle 8 cm wide and 10 cm long. The most common error to this question was 40 squares, rather than the correct response of 20 squares. The hardest Year 6 Measurement task required students to determine how far a person would walk in one minute if the person could walk 2400 m in 30 minutes (at the same speed). Many students had difficulty dividing 2400 by 30. The easiest Measurement item (item number I0013) was included in the Kindergarten Term 3 assessment. The task required students to order 5 sticks from shortest to longest.

Overall, Number (N), Space (S), Measurement (M), and Working Mathematically (WM) items are spread across the scale, with the easiest Space item located at −6.6 logits. The easiest items, not placed on this scale, were Number items

```
LOCATION          PERSONS     ITEMS [locations]
                  XXXXXXXX |
                           | I0138
                       XXX |
  6.0                  XXX |
                       XXX | I0120 I0137
            XXXXXXXXXXXXXX | I0136
                           | I0132
                  XXXXXXXX |
  5.0           XXXXXXXXXX | I0106 I0133
                      XXXX | I0134
                 XXXXXXXXX |
             XXXXXXXXXXXXX | I0117
                    XXXXXX | I0135 I0123
  4.0       XXXXXXXXXXXXXX | I0122 I0115 I0124 I0105 I0112 I0116
           XXXXXXXXXXXXXXX | I0102
                   XXXXXXX | I0130
          XXXXXXXXXXXXXXXX | I0092 I0104
           XXXXXXXXXXXXXXX | I0139 I0111 I0131 I0129 I0090
  3.0    XXXXXXXXXXXXXXXXX | I0114 I0127 I0110 I0113
                XXXXXXXXXX | I0119 I0099
        XXXXXXXXXXXXXXXXXX | I0082 I0083
         XXXXXXXXXXXXXXXXX | I0087 I0088 I0128
             XXXXXXXXXXXXX | I0103 I0086 I0121
  2.0        XXXXXXXXXXXXX | I0080 I0089 I0108 I0109 I0101 I0098 I0091
             XXXXXXXXXXXXX | I0100
         XXXXXXXXXXXXXXXXX | I0048
                 XXXXXXXXX | I0081 I0093 I0084 I0085 I0079
               XXXXXXXXXXX | I0107 I0078
  1.0        XXXXXXXXXXXXX | I0097 I0094 I0126
                  XXXXXXXX | I0118 I0050 I0073 I0071
                   XXXXXXX | I0058 I0076 I0096 I0077 I0070 I0095
               XXXXXXXXXXX | I0125 I0067 I0066
                     XXXXX | I0065
  0.0           XXXXXXXXXX | I0046 I0057
                    XXXXXX | I0072 I0069 I0049
                   XXXXXXX | I0059 I0075
                      XXXX | I0047
                   XXXXXXX | I0060 I0025 I0074
 -1.0                    X | I0018
                   XXXXXXX | I0056
                    XXXXXX | I0064
                       XXX | I0068
                     XXXXX | I0019 I0045 I0026
 -2.0              XXXXXXX | I0044 I0055 I0027 I0032
                      XXXX | I0024
                      XXXX | I0051 I0063
                       XXX | I0053 I0043
                        XX | I0022 I0020
 -3.0                 XXXX | I0033
                           | I0054 I0021 I0061 I0023
                       XXX | I0030 I0017
                           | I0031
                        XX | I0052 I0042 I0013
 -4.0                      |
                       XXX |
                           | I0029 I0062 I0040
                        XX |
                         X | I0039
 -5.0                      |
                         X | I0028 I0041
                           |
                           |
                         X | I0007 I0015
 -6.0                      | I0016 I0002
                         X | I0004
                           |
                           | I0011 I0005 I0001 I0038
                           | I0014 I0009
 -7.0                      | I0003 I0006
                           |
                           |
                           |
 -8.0                      |
                           |
                           | I0010
```

Each X represents 4 students.
The hardest item was I0138 located at 6.4 logits.
The easiest item was I0010 located at –8.4 logits.

Figure 18. Item Map–final scale.

The order of difficulty of items developed from the Term 3 data set matches, to a large extent, the expected order of syllabus objectives and staged outcomes. As mentioned earlier, the stages of schooling are Early Stage 1 (Kindergarten), Stage 1 (Years 1 – 2), Stage 2 (Years 3 – 4), and Stage 3 (Year 5 – 6). An analysis of the outcomes coded as Early Stage 1 (ES1), Stage 1 (WM1, N1, S1, M1), Stage 2 (WM2, N2, S2, M2) and Stage 3 (WM3, N3, S3, M3) and item difficulty matched on the NAS shows some disparity between the data and the expected level of difficulty. In some cases, Stage 2 outcomes were placed at an easier position on the scale to Stage 1 items. These items were related to numeral identification and division into equal parts. Some items categorised as Stage 1 outcomes were positioned in the order of difficulty within Stage 2 outcomes. These items proved more difficult than curriculum outcomes indicated. For example, item I0019 required Kindergarten students to count the number of corners on a box (Space 1.1); however students found this item very difficult.

The description of each item can also be located along the scale as a means of examining more closely the types of knowledge and skills shown by the students. These were matched to curriculum expectations.

Implementation of the Numeracy Achievement Scale: numeracy growth analysis 2002 and 2003

Numeracy growth in Case Study, Trialling, and Reference Schools 2002

It was not expected that large gains in numeracy growth would be made in a short period of time (April–October, 2002). However, substantial numeracy growth (0.76 logits) was shown across all school groups (Case Study Schools, Trialling Schools and Reference Schools); but there were no significant differences found in mean numeracy growth among the three groups (0.75 logits; 0.74 logits; 0.79 logits). There were, however, significant differences found in mean numeracy growth at particular year levels (see Table 1): An analysis of variance (MANOVA) between Trialling and Reference groups indicated a grade level effect for numeracy growth (0.000).

Table 1 Numeracy Growth 2002 by year level and school group

	Group			
Year level	Case Study Schools	Trialling Schools	Reference Schools	Mean for year level
Kindergarten	1.51 (±0.82)	0.45 (±0.67)	0.80 (±0.85)	1.10 (±0.91)
Year 1	1.13 (±1.27)	0.78 (±1.12)	1.13 (±1.05)	1.06 (±1.18)
Year 2	0.58 (±0.67)	1.22 (±0.69)	1.19 (±0.93)	0.95 (±0.84)
Year 3	0.46 (±0.83)	0.72 (±0.85)	0.98 (±1.03)	0.65 (±0.90)
Year 4	0.69 (±0.87)	0.66 (±0.99)	0.43 (±0.88)	0.62 (±0.90)
Year 5	0.84 (±0.85)	0.56 (±1.13)	0.68 (±0.98)	0.74 (±0.96)
Year 6	0.61 (±0.88)	0.88 (±0.96)	0.62 (±1.04)	0.68 (±0.95)
Overall mean	0.75 (±0.92)	0.74 (±0.97)	0.79 (±1.00)	**0.76** (±0.95)

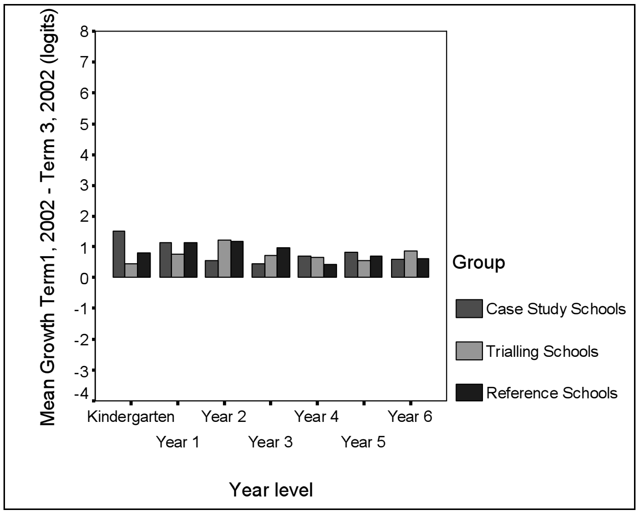


Figure 19. Numeracy Growth by year level and school group

Figure 19 shows that overall, in 2002 there was more growth shown at Kindergarten, Year 1 and Year 2 than for the upper year levels. In contrast, Year 6 students for all school groups had comparatively low growth scores (0.68 logits), and this was similar to the low growth scores for students in Year 3 and Year 4. Some of these disparate values could be explained by the composition and number of the students involved in the samples

Trialling school students showed significant numeracy growth at Year 4 and Year 6, and much less than expected growth at Year 5. Ten of the 19 cohorts from Trialling schools demonstrated greater than expected growth when compared with their counterparts in Reference schools. Gains in numeracy growth of students from Trialling schools were much greater than initially anticipated because they had been performing below the state average in numeracy for a number of years. In comparison, Reference schools were selected because BST data indicated that students were performing at or slightly above the state average.

Even though extensive professional development support was provided to the Trialling schools in 2002, substantial growth was not seen until 2003. Further, even though teachers within a school implemented the same strategies, improvements in numeracy achievement were not equally demonstrated in the individual classes that were assessed. This was not surprising since a significant amount of time and professional development would be required before strategies can be implemented consistently and effectively by teachers across a school.

Numeracy growth in Trialling Schools: 2002-2003

The mean numeracy growth for 276 matched Trialling school students from 2002 to 2003 was very strong overall (2.10 logits); in 2003 it was greater than expected (0.83

logits), and larger than the mean numeracy growth for the same students in 2002 (0.77 logits). Although Case Study schools and Reference schools did not participate in 2003, the numeracy growth of the Trialling school students was higher than the mean growth (0.76 logits) shown by all school groups (Case study, Reference, and Trialling schools).

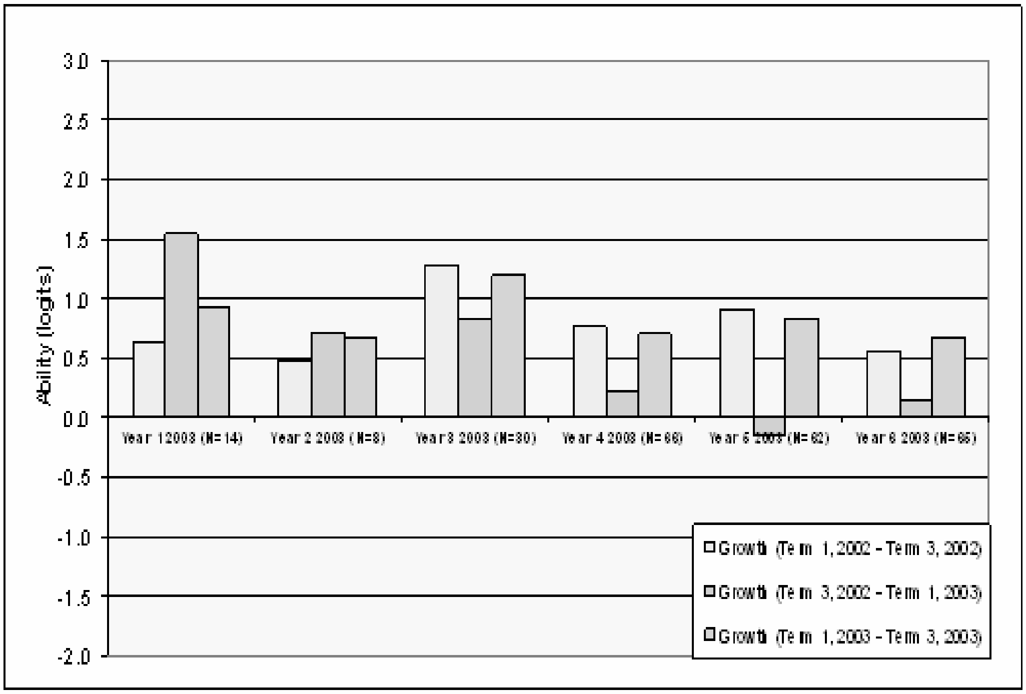


Figure 20. Mean growth (Trialling schools) by cohort by assessment point

Figure 20 shows mean numeracy growth of Trialling school students from 2002 to 2003 highlighting variations in growth between Term 3, 2002 and Term 1 2003. In 2003 students showed more growth at Year 3, Year 4, Year 5 and Year 6 than their counterparts in Reference schools in 2002. The Year 3 cohort showed the greatest growth in the Term 1 to Term 3, 2003 period, and the largest growth by cohort over the entire 18-month period. Further analysis showed that numeracy growth for Year 3 boys in the 2003 sample was much higher than for boys and girls overall. This growth was much greater than the mean growth for all Trialling school boys in 2002.

Limitations of the Research

The very parameters of the project precluded any substantive longitudinal design – it was established initially as a two-year project and extended for monitoring of the Trialling schools. There was limited time (seven months) to implement the selective strategies in the ten Trialling school sites in 2002. The monitoring of Trialling school students was limited to a matched sample of 276 students.

Development of assessment instruments that measure student achievement across a wide range of abilities and ages have limitations: the NAI did not significantly challenge a small minority of extremely capable students; it was developed to report numeracy achievement in general terms for all students from Kindergarten to Year 6. It was not possible to report achievement for each of the strands (Number, Space, Measurement 7 Data) for a given cohort within one Year level. This was because there were a limited number of items from each strand that could be administered within the limits of the NAS.

The sample of 20 Case Study schools participating in 2002 was identified by representatives from each sector as having outstanding numeracy practices in place. Unlike the 25 schools identified in 2001, these 20 Case Study schools were not identified primarily on the basis of student achievement data. The numeracy achievement of these students was found to be similar to that of students from Trialling schools and Reference schools. Thus, the perception that Case Study schools had "an outstanding program" was not supported by student achievement data in the project. Further, the 45 Case Study schools constituted only a small sample of those schools in NSW that were ‘making a difference’ in achieving outstanding numeracy outcomes.

The nature of the research process and dynamic school contexts meant that only a sample of the strategies identified as being outstanding could be adapted, implemented and monitored at the ten Trialling schools. While the focus on the particular strategies being implemented at each Trialling school was maintained, it was not possible to exclude the impact of other factors in achieving project outcomes, such as the influence of individual 'quality' teachers and their pedagogical and content knowledge.

Implications

One significant research implication of this project would be to establish a five-year longitudinal study that used student and teacher identifiers to examine the extent to which such improved numeracy outcomes can be sustained and enhanced over a longer period. The NAI was developed within the context of the NSW mathematics K-6 curriculum as a unique numeracy construct based on student interview data and this supports and extends the use of other instruments. Further research may ascertain the extent to which the NAI can be adapted to the mathematics K-6 curriculum for other Australian States and Territories. Further research could determine the extent to which the findings drawn from this sample of schools could be generalised across educational systems and specific school types within and across Australia.

References

Adams, R. J. & Khoo, S.T. (1996). *Quest – Interactive item analysis system*. Version 2.1. Melbourne: Australian Council for Educational Research.

Australian Association of Mathematics Teachers, (1997). *Numeracy = everyone's business (Report of the Numeracy Education Strategy Development Conference*). Adelaide:

Australian Association for Mathematics Teachers Inc./ Department of Education, Training and Youth Affairs.

Askew, M., Brown, M., Rhodes, V., Wiliam, D., Johnson, D. (1997). *Effective teachers of numeracy in primary schools: Teachers' beliefs, practices and pupils' learning.* Paper presented at the British Educational Research Association Annual Conference, September 11-14, 1997, University of New York.

Battista, M.C. (1999). Spatial structuring in geometric reasoning. *Teaching Children Mathematics*, November, 171-177.

Board of Studies NSW (2000). *Mathematics K-6 Outcomes and Indicators* (draft). Sydney:Board of Studies NSW.

Brown, M. (2000). What kinds of teaching and what other factors accelerate primary pupils' progress in acquiring numeracy? *Improving Numeracy Learning: What does the research tell us?* ACER Research Conference, 2000.

Callingham, R. & McIntosh, A. (2001). A developmental scale of mental computation. In J. Bobis, B. Perry, & M. Mitchelmore (Eds.), *Numeracy and beyond. Proceedings of the 24th annual conference of the Mathematics Education Research Group of Australasia*, (pp. 130-138). Sydney: Mathematics Education Research Group of Australasia.

Clarke, D., Sullivan, P., Cheesman, J., & Clarke, B. (2000). The Early Numeracy Research Project: Developing a framework for describing early numeracy learning. In J. Bana & A. Chapman (Eds.), Mathematics education beyond 2000 (*Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia*

Inc., pp. 180-187) Perth: Mathematics Education Research Group of Australasia Inc.

Commonwealth of Australia (2004). *What's making the difference? Achieving outstanding numeracy outcomes in NSW Primary Schools.* Canberra: Department of Education, Science and Training.

Department of Education, Training and Youth Affairs (2000). *Numeracy, a priority for all: Challenges for Australian schools.* Canberra: Department of Education, Training and Youth Affairs.

Horne, M. & Rowley, G. (2001). Measuring growth in early numeracy: Creation of interval scales top monitor development. In M. van den Heuvel-Panhuizen (Ed.). Proceedings of the 25th Annual Conference of the International Group for the Psychology of Mathematics Education, (Vol. 3, pp. 161-167). Utretcht: The Netherlands: Freudenthal Institute.

Lokan, J., Doig, B., & Underwood, C. (2000). *Numeracy assessment and associated issues.* Canberra: Australian Association of Mathematics Teachers Inc. Department of Education Training and Youth Affairs

Looveer, J. (2003). Using modern psychometric theory to identify differential item functioning in polytomously-scored constructed response items. Unpublished doctoral thesis. University of NSW: New South Wales.

McIntosh, A. & Dole, S. (2000). Early arithmetical learning and teaching. In K. Owens & J. Mousley (Eds.) *Research in mathematics education in Australia 1996–1999.* Sydney: Mathematics Education Research Group of Australasia Inc.

Ministry of Education, Victoria (2001). The Middle Years Numeracy Research Project. Melbourne: Ministry of Education, Victoria.

Ministry of Education, Victoria (2002). *The Early Years Numeracy Research Project*. Melbourne: Ministry of Education, Victoria.

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

NSW Department of Education and Training, (1999). *Counting On: A professional development package*. Sydney: NSW Department of Education and Training, Curriculum Support Directorate.

NSW Department of Education and Training, (2000). *Count me in too: A professional development package*. Sydney: NSW Department of Education and Training, Curriculum Support Directorate.

NSW Department of Education & Training, (2001a). *Basic Skills Testing: Literacy and Numeracy*. Sydney: NSW Department of Education & Training, Assessment and Reporting Directorate.

NSW Department of Education & Training, (2001b). *Secondary numeracy assessment program (SNAP)*. Sydney: NSW Department of Education & Training, Assessment and Reporting Directorate.

NSW Department of Education & Training, (2001c). *English Language and Literacy Assessment (ELLA)*. Sydney: NSW Department of Education & Training, Assessment and Reporting Directorate.

NSW Department of Education and Training, (2001d). *Count me into measurement: A professional development package* Sydney: NSW Department of Education and Training, Curriculum Support Directorate.

Outhred, L. & Mitchelmore, M.C. (2000). Young children's intuitive understanding of area measurement. *Journal for Research in Mathematics Education, 31*, 144-168.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.Chicago: University of Chicago Press.*

Siemon, D. & Griffin, P. (2000). Researching numeracy in the middle years of schooling. In J. Bana & A. Chapman (Eds), *Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia.* (pp. 542-550) Perth: Mathematics Education Research Group of Australasia Inc.

Willis (1998). Which numeracy? *Unicorn*, 24 (2), 32-42.

Wright, R. & Gould, P. (2002). Mapping overlapping waves of strategies used with arithmetical problems. In A, Cockburn & E. Nardi (Eds.), *Proceedings of the 26th annual conference of the International Group for the Psychology of Mathematics Education*, (Vol 1, pp.197-202). University of East Anglia, Norwich.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, R., Mulligan, J. T., Bobis, J. & Stewart, R. (1996). Research on early number learning. In B. Atweh, K. Owens & P. Sullivan (Eds.), *Research in mathematics education in Australasia 1992-1995* (pp.281-311). Sydney: Mathematics Education Research Group of Australasia.

[1] This project, funded under the *Commonwealth Government's Numeracy Research and Development Initiative*, began in January 2001 and was completed in January 2004 through the NSW Department of Education and Training Strategic Research Directorate. The first author was the academic consultant, the second author was the NSW DET psychometrician and the third author was the Project Manager.