SCAFFOLDING NUMERACY IN THE MIDDLE YEARS LINKAGE PROJECT 2003-2006

RESEARCH APPROACH AND RATIONALE FOR THE USE OF RASCH ANALYSES[1]

This paper describes the research approach in more detail, in particular, the decisions and assumptions made in relation to each research question, the development of the instruments, and the rationale for the use of Rasch analyses including the use of summary statistics and effect sizes for the evaluation of student progress.

The *Scaffolding Numeracy in the Middle Years Linkage Project 2003-2006* was designed to address six research questions. These were:

- To what extent can we accurately identify key points in the development of multiplicative thinking and rational number beyond the early years?
- To what extent can we gather evidence about each student's achievements with respect to these key points to inform the development of a coherent learning and assessment framework?
- To what extent can authentic assessment tasks be developed and used to assess student performance against the framework?
- To what extent does working with the tasks and the knowledge they provide about student understanding assist teachers to improve student numeracy performance at this level?
- What strategies and/or teaching approaches are effective in scaffolding multiplicative thinking and rational number understanding in the middle years? and
- What are the key features of classroom culture and discourse needed to scaffold students' numeracy-related learning at this level?

Accurate identification of key points

The first step in identifying key points in the development of multiplicative thinking and rational number beyond the early years was to develop a Draft Learning and Assessment Framework for Multiplicative Thinking (LAF) or hypothetical learning trajectory (Simon, 1995). This was done on the basis of data from the *Middle Years Numeracy Research Project* (Siemon & Virgona, 2001) and relevant research literature. The draft framework was used to inform the design of authentic (rich) assessment tasks that could be used to evaluate multiplicative thinking in Years 4 to 8, and to test the validity of the key points and assumptions inherent in the draft framework.

Gathering evidence about each student's achievements

A range of specifically-targeted, authentic (rich) tasks were developed to assess the key points in the draft framework. These were broadly accessible, and provided opportunities for students to respond at different levels of sophistication in terms of their solution strategies and reasoning.

[1] Prepared by Dr John Izard, Adjunct Professor, RMIT University and Professor Dianne Siemon, RMIT University

Development of authentic (rich) assessment tasks

Given the debate about the use of the term *authentic* more generally and its subsequent use in the project to describe school-based learning projects, it was decided to use the term *rich* assessment tasks to describe the type of tasks intended here. Rich assessment tasks may be relatively simple or quite complex. In order to provide the evidence required in this study, it was agreed that the rich tasks needed to include multiple items to more accurately evaluate the extent of an individual's achievement. Further, it was agreed that for each item within a task, it may be necessary to identify different levels of achievement to accommodate the richness of the evidence. As some items within a task might be scored on a correct/incorrect basis (ie, assigned a score of 1 or 0) and others on a partial credit basis (ie, assigned scores of 0, 1, 2, or 3 or more on the basis of the quality of the student response), it was necessary to develop rich descriptions for the qualitatively different levels of responses to more closely identify an individual's overall achievement level. This was done by developing scoring rubrics for each item within each task.

The rich assessment tasks and the scoring rubrics that were developed and used for the project can be found in the *Assessment Materials* and the *Support Materials* (under Additional Assessment Tasks) sections of the CD-ROM.

The status of an individual student (meaning where they stand on the continuum of intended achievement) can be described in terms of their success on a sequence of tasks representing examples of achievement at key points on the continuum. Item Response Modelling (RASCH Modelling) investigates the extent to which performance on a range of tasks provides stable estimates of achievement. For example, are the tasks consistent with each other, do they measure similar concepts and applications, and do they remain in a similar order of difficulty over different testing occasions. Any tasks which did not satisfy these conditions were rejected.

A pool of tasks were developed and used over the course of the project to reduce the possibility of 'teaching to the task' and/or students recognising tasks used in the initial assessment. From this pool of tasks two alternative assessment task booklets have been provided on the CD-ROM which will enable teachers to locate student achievement in relation to the *Learning and Assessment Framework for Multiplicative Thinking (LAF)*. Both tests can be used for pre or post assessment purposes provided a different test is used each time.

Assessing student performance against the framework

To distribute student achievement against the LAF, the difficulty level of the rich assessment tasks need to vary. If an easy task is given, scores on the items within that task will be high while difficult tasks will tend to have lower scores. All tasks need to be calibrated against one another. When items are chosen for an initial achievement test they generally differ from the items chosen for a subsequent achievement test. If items from both the initial and the subsequent tests have not been given to a common group of students to calibrate their relative difficulty, it cannot be deduced whether the scores vary because the tests differ, because learning has occurred, or a combination of these. To take account of variation in test difficulty, raw scores need to be interpreted in terms of the scaled positions (levels of cognitive development) on the continuum (LAF). For the assessment options provided in the CD-ROM a *Raw Score Translator* has been developed for each option (see *Assessment Materials* on the CD-ROM)

Assisting teachers to improve student numeracy performance

The development of the *Learning and Assessment Framework for Multiplicative Thinking* enabled teachers to locate students on this continuum after assessment had taken place. When a student's location was identified accurately and validly, research school teachers were supported to intervene in targeted ways with a view to shifting the student to a higher level of multiplicative thinking. This was done by consolidating students' understanding at their current level and introducing and developing concepts and ideas at the next level. Advice to support these interventions was provided with the Draft LAF. A revised and elaborated version of this appears with the final version of the LAF which is included on the CD-ROM.

Effective strategies and/or teaching approaches

There are two ways to evaluate the effectiveness of teaching strategies and/or approaches. That is, by a longitudinal comparison and/or a cross-sectional comparison. In this project, the longitudinal comparison was afforded by a comparison of achievement on the initial and final assessment over a two year period after intervention had taken place. That is, individual students were identified and their progress followed over time. This measure relies on that the assumptions that changes in student achievement occur due to learning (rather than other factors such as changes in classroom composition or teacher experience), that students had a fair chance of showing their skills and knowledge in the assessment, and that coaching for the assessment did not occur.

For the cross-sectional comparison, scores for students in Research Schools were contrasted with scores from students in Reference Schools (matched with respect to location and approximate socio-economic status). The assumptions are similar to those in the longitudinal comparison. In addition, it is assumed that that the cohorts come from the same schools and staffing was consistent over the comparison period. Since individual students in the Reference Schools were not identified, their progress over time cannot be followed.

Development of the Instruments

The *Learning and Assessment Framework for Multiplicative Thinking (LAF)* on which the instruments were based and the associated teaching implications are included in other sections of the CD-ROM.

The pre- and post-tests were chosen from rich assessment tasks to provide evidence of multiplicative thinking relating students progressing through Years 4 to 8 in Victoria and Tasmania. There were two extended tasks [Butterfly House (bth) with separate parts (items) a – i and Tables and Chairs (tch) with separate parts a – m]. The other tasks were intended to supplement the extended tasks with each task having a smaller number of parts [Packing Pots (pkp) parts a – d; Pizza Party (pzp) parts a – c; Missing Numbers (msn) parts a – b; Canteen Capers (cnc) parts a – b; Adventure Camp (adc) parts a – b; Filling the Buses (ftb) parts a – b; Fencing the Freeway (ftf) parts a – d; and Swimming Sports (sws) parts a – b]. Not all tasks were to be administered to all students. Table A shows the tasks administered for each version of the test used in the initial assessment phase of the study.

Table A: Allocation of Tasks to Test Versions

	Version 1	Version 2	Version 3	Version 4
Butterfly House (bth) btha, bthb, bthc, bthd, bthe, bthf, bthg, bthh, & bthi	•		•	
Tables and Chairs (tch) tcha, tchb, tchc, tchd, tche, tchf, tchg, tchh, tchi, tchj, tchk, tchl, & tchm		•		•
Packing Pots (pkp) pkpa, pkpb, pkpc, & pkpd	•	•	•	•
Pizza Party (pzp) pzpa, pzpb, & pzpc	•	•	•	•
Missing Numbers (msn) msna & msnb	•			
Canteen Capers (cnc) cnca & cncb	•	•		
Adventure Camp (adc) adca & adcb	•	•	•	
Filling the Buses (ftb) ftba & ftbb		•	•	•
Fencing the Freeway (ftf) ftfa, ftfb, ftfc, & ftfd			•	•
Swimming Sports (sws) swsa & swsb				•

Rationale for the Analyses

Each of the pre-test and post-test analyses required a person-by-item matrix of data. The four initial test versions had common items so that other tasks could be calibrated against these common tasks. This was necessary to take account of differences in score due to variation in task difficulty, and to allow comparable scaled scores to be obtained for students who attempted different tasks. It was also essential to allow comparisons of progress made in the time of the study.

In each analysis, the first run checked the technical properties of the test. Which checked that the test was internally consistent and that the items distinguish between the higher achievers and the lower achievers. It is important to check that the items are able to separate the students, and in turn, that the students are separated by the items. Further it is important that the test has a range of difficulty to ensure that the test tasks can assess improvements over time (as shown by an improvement in the scaled student mean). If there is a ceiling effect then such a test will be ineffective for showing progress: those with perfect scores would be likely to achieve perfect scores, again, after further learning but the “evidence” from the testing would imply that there had been no change. Students with perfect scores are likely to be beyond the range of the test and therefore no increase in achievement would be detected. Subsequently analyses calibrated the tasks against each other and located components of each task on the learning continuum.

Evaluation of Student Progress

Without valid student assessment practices the actual achievements are never compared in a legitimate way with the intentions (Izard, 2002a). Valid assessments must document the level of achievement prior to a particular stage of learning and a later assessment must document a higher level of achievement to provide *evidence of the changes* in the students. Note that changes are not always positive. In some cases students fail to make progress, or obtain lower scores on the required skills after the "teaching" of those skills. Before it can be shown that progress has been achieved in a teaching program, the current achievement status of each student needs to be indicated and the subsequent assessments have to include tasks representative of the skills which were intended to be taught.

In this project, data were collected on calibrated rich assessment tasks to identify where students 'were at' in relation to the knowledge, skills and strategies embodied in the LAF. In particular, the testing identified what knowledge, skills and strategies had been established, and what knowledge, skills and strategies were not yet within reach. This evidence (collected with comparable tests at two points of time) described the changes in students between these two times.

One way of reporting the effects of learning is to report the difference between means (on two or more occasions) in standard deviation units (obtained by taking the square root of the pooled variance – see Cohen, 1969, 1977, 1988, and Glass, McGaw, and Smith, 1981). When the magnitudes of improvements are expressed as effect sizes in standard deviation units Cohen's descriptors can be used as a common language to describe these magnitudes. Table B shows these descriptors together with ranges assigned by the writer. Effect sizes are described as "very small", "small", "medium" or "large".

Another way is to report how much overlap there is between the pre-test and the post-test results (see Cohen, 1977, pp. 20-27). For example, for two normal populations with equal variability and equally numerous, an effect size of 0 indicates 100% overlap or 0% non-overlap. An effect size of 0.2 indicates 14.7% non-overlap (the component of the combined distribution not shared by the two populations). The corresponding non-overlap values for effect sizes of 0.5 and 0.8 are 33% and 47.4%.

Table B: Descriptors for magnitudes of effect sizes (after Cohen, 1969, p.23) and assigned ranges (Table 1 from Izard, 2004)

Effect Size Magnitude	Cohen's Descriptor and Cohen's Example	Assigned Range
< 0.2	Very small*	0.00 to 0.14
0.2	Small - difference between the heights of 15 and 16 year old girls in the US	0.15 to 0.44
0.5	Medium ('large enough to be visible to the naked eye') difference between the heights of 14 and 18 year old girls	0.45 to 0.74
0.8	Large ('grossly perceptible and therefore large') difference between the heights of 13 and 18 year old girls or the difference in IQ between holders of the Ph.D. degree and 'typical college freshmen'	0.75 or more

- Note that "very small" is a descriptor devised by Izard (2004) for magnitudes less than "small".