Numeracy Assessment: How Reliable are Teachers' Judgments?

Gill Thomas
Maths Technology Ltd
<gill@nzmaths.co.nz>

Andrew Tagg
Maths Technology Ltd
<andrew@nzmaths.co.nz>

Jenny Ward
Maths Technology Ltd
<jenny@nzmaths.co.nz>

Research reports on the Numeracy Development Projects (NDP) have consistently shown that student performance in numeracy improves as a result of participation. The evidence for these reports has been based largely on student achievement data collected by classroom teachers. This paper reports on an investigation into the consistency of teacher judgments of students' strategy stages on the Number Framework. The findings show a high level of agreement between the judgments of classroom teachers and those of independent researchers.

Background

The New Zealand Numeracy Development Projects (NDP) are a large-scale Government initiative in mathematics education aimed at improving students' mathematics ability through the professional development of teachers. Each project has three main components: the Number Framework that describes a progression in strategy stages that students use to solve problems and the key knowledge that they will require to progress; a diagnostic interview that enables teachers to assess students' capabilities according to the Number Framework; and a professional development programme.

Since their inception, the NDP have been informed by research that has examined both the experiences of the participants and the achievement of students within the projects (Ministry of Education, 2005c). Positioned within this body of research is the NDP Longitudinal Study, which has examined the impact of the projects in schools that have been involved over a number of years. One component of the Longitudinal Study in 2005 focused on the teachers' use of numeracy assessment information and the accuracy of their strategy stage assessments. This paper reports on those findings.

The primary and most compelling reason for teachers to use diagnostic assessment information is to improve teaching and learning. After reviewing literature on the consistency of teacher judgments, Bobis (1997) noted:

> Such findings emphasise the necessity of determining the reliability of teacher ratings ... particularly given the pivotal role such ratings play in determining instructional decisions for individual and groups of children to help them advance thorough the stages and levels of the Learning Framework. (Bobis, 1997, p. 4)

The NDP encourage teachers to use the information they obtain about their students from the diagnostic interview to group students for instruction and help determine learning experiences.

> NumPA [the Numeracy Project Assessment diagnostic tool] provides a wealth of diagnostic assessment information about students. There needs to be an appropriate link between this data and the learning experiences you provide for your students through the classroom programme. (Ministry of Education, 2005b, p. 2)

By the end of 2005, approximately 460 000 students had participated in the NDP (Parsons, 2005). Data for the majority of these students has been collected on the national database. This data has been used to describe student achievement within the project, but the reliability of the teacher judgments on which this large data set is based has never been fully examined. The consistency of patterns of student performance and progress over several years has been used as the primary source of evidence of the consistency of the teacher judgments.

> For the third year running, the majority of students improved during their participation in the Advanced Numeracy Project. The patterns of achievement within year groups in 2003 confirm previous years' results. (Higgins, 2004, p. 11)

Reliability is an issue in all educational assessments, but the degree to which it can be considered a problem is related to the function and use of the assessment (Nystrom, 2004). One of the aims of the NDP Longitudinal Study is to collect numeracy data from students in the years following their school's participation in the NDP to help establish benchmarks or expectations for achievement and progress. If schools are using this data to compare the performance of their students with national norms, it is important that the data is reliable.

> ... if the performance of students from different classes, schools or regions are to be compared for any reason in the future, it is imperative that the inter-rater reliability be evaluated. (Bobis, 1997, p. 4)

A study with similarities to the current investigation was undertaken in conjunction with the Count Me in Too project (CMIT) in New South Wales (Bobis, 1997). The key features of CMIT include the Learning Framework in Early Number and a performance-based assessment instrument, the Schedule for Early Number Assessment (SENA). In the Bobis (1997) study, 16 teachers viewed a video recording of five students and were asked to rate the students' performance using the SENA. Quantitative analysis found a high degree of inter-rater reliability between teachers' ratings, with a high correlation between individual teachers' ratings and the mean rating of the whole group.

The present investigation differs from the Bobis (1997) study in that it focused on the observation of teachers assessing their own students. This enabled a more naturalistic examination of teachers' assessment judgments. In addition, the current study also involved analysing the responses of a large number of teachers to written assessment scenarios.

Method

The Longitudinal Study began in 2002 with the participation of 20 schools. Each year, new schools are randomly selected from a list of schools that have recently completed the NDP training. The list is stratified by decile to ensure that the sample in the Longitudinal Study closely approximates the national sample and has similar numbers of students in years 1 to 8. In 2005, a total of 26 schools participated in the Longitudinal Study, including 12 of the original 20 schools. Table 1 summarises the composition of schools participating in the 2005 Longitudinal Study, hereafter referred to as the 2005 longitudinal schools. The low-decile band includes decile 1–3 schools, the medium-decile band includes decile 4–7 schools, and the high-decile band includes decile 8–10 schools.

Table 1
Composition of Schools in the 2005 Longitudinal Study

	Decile Group			
School Type	Low	Medium	High	Total
Years 1–6	4	5	5	14
Years 1–8	4	3	1	8
Years 7–8	1	2	1	4
Total	9	10	7	26

Two methods of data collection were used in the current study: observations and questionnaires. The observations of teachers' assessment interviews took place in two locations, with one southern and one northern city selected for this purpose. A random selection of the 2005 longitudinal schools within these two cities was used to determine the sample of schools for the observations. Letters of invitation were sent to the principals of selected schools describing the purpose of the research, what would be required of participating teachers, and the observation process. All five schools accepted. Up to 10 teachers within each school were invited to participate in the assessment observations. Where the school had fewer than 10 teachers, all were observed. Where there were more than 10 teachers, the lead teacher invited 10 of the teachers to participate in the assessment observations.

In total, 37 teachers were observed assessing the strategy stages of their students. Each teacher individually interviewed one or two students, and a total of 70 students and 156 teacher judgments across the three operational domains (additive, multiplicative, and proportional) were observed. Two researchers, both experienced in the assessment of student strategy stages, undertook the observations.

The teachers were instructed to select a diagnostic assessment technique to assess their students' strategy stage. While the teachers were encouraged to use whatever techniques they employed in their usual classroom practice, NumPA, the Global Strategy Stage (GloSS) assessment, or the teachers' own diagnostic questions were identified as possible methods in the letter of invitation. The teachers were told that the researchers might ask the students extra questions to clarify their strategy stages and that there would be a brief follow-up discussion with the teacher to clarify how their judgments for each student were made. The teacher and students were withdrawn from the classroom for the assessment observations.

Questionnaires were sent to all teachers in the 2005 longitudinal schools. Of the 400 questionnaires distributed, 230 responses were received. It is not possible to calculate an accurate response rate because the number of questionnaires distributed was based on an estimated class size of 23 students, ensuring more than sufficient questionnaires were provided to each school. The questionnaire included assessment scenarios that asked the teachers to identify the students' strategy stage based on the information given. Also included were questions focused on student achievement and the assessment of numeracy.

Findings[1]

The Picture of Numeracy Assessment in Longitudinal Study Classrooms

The large majority of teachers (83%) in the 2005 longitudinal schools have completed the full NDP professional development programme. Of the 17% of teachers who have not completed the full professional development programme, 10% have received no training at all while the others have attended a variety of professional development workshops or covered the numeracy project in their pre-service training. The majority of teachers (80%) had undertaken the NDP professional development programme in their current school.

Almost all teachers (94%) in the 2005 longitudinal schools track the strategy stages of students in their class. Eighty-four percent of teachers reported using NumPA, with over half the teachers (58%) using it once or twice a year to group students. Eleven percent of teachers use NumPA to assess new students when they enter the classroom:

> At the beginning of school year to identify stage and assign maths group in class. (Teacher)
>
> Beginning of year to recheck levels and usually at beginning of Term 4 for school reports. (Teacher)

The large majority of teachers (84%) reported the use of informal ongoing numeracy assessment during class time, with approximately two-thirds of teachers (69%) using informal methods of assessment at least once a week or on an informal basis as required to check student progress or confirm their instructional groupings.

> I regularly observe during group time to check the appropriateness of the group level and the placements in the group. (Teacher)
>
> I do one-to-one checking and conferencing, especially with targeted children. (Teacher)

One-third of the teachers reported use of the GloSS assessment, primarily to check students' strategy stages throughout the year:

> Reporting to parents – a way to show gains. (Teacher)
>
> To show progress through the strands. (Teacher)

The main use reported by teachers for the numeracy assessment information was classroom grouping. A small number of teachers noted that they used the assessment information to report to parents and/or the principal.

> At the beginning of the year, I interview each child and use this for grouping. (Teacher)
>
> I use the children's responses to questions that are part of my teaching to make changes to groups. Where I am not sure, I will target these questions to particular children. I try to make sure that I check all the children in this informal way every couple of weeks. (Teacher)
>
> I make up my own "test" to determine what levels children are at. I use the responses to this written test as information for regrouping if necessary. (Teacher)

Teachers reported a high level of confidence in the assessment judgments they are making, with the large majority of teachers (84%) identifying that they are fairly confident or very confident that they know the strategy stage each student in their class is operating at.

[1] Where questionnaire responses do not total 100%, this is due to teachers leaving questions unanswered.

Time Spent on Numeracy Assessment

Time is an important commodity for any teacher, and the amount of time taken to complete an assessment has important implications for classroom practice. With the large majority of teachers using a full NumPA assessment at least once a year and nearly a quarter of teachers (24%) conducting NumPA assessments on all students twice a year, the amount of time taken for each assessment is worthy of consideration.

In the observations of the teacher assessment interviews, the average time taken for teachers to assess the strategy stages of a student was 13 minutes. These observations involved the assessment of students' strategy stages and did not include assessment of the knowledge domains. Adding the knowledge domains to the assessment would conservatively increase the time taken to 20 minutes. This equates to over 8 hours of assessment time for a class of 25 students, which in turn creates resourcing issues. Teacher comments support this assertion.

> Testing requires at least two full relief days per teacher. It's very costly. (Teacher)
>
> At times I use the diagnostic test to help teachers with stage placement, as there are time issues when it comes to the teachers testing all the children in their class. (Teaching principal)
>
> We are very fortunate to be given as much release time as necessary to interview each child at the beginning of the year. I believe this is the key to appropriate and effective grouping and instruction. (Teacher)

The observation of the teacher assessment interviews indicated considerable variation between teachers in the time taken to assess students' strategy stages. A comparison of teachers who made assessment judgments in all three domains shows an interesting pattern. Table 2 compares the accuracy of teacher judgments made in all three domains with the time taken to make decisions. The table shows that the teachers taking the most time to make assessment decisions were least likely to be in agreement with the researchers.

Table 2
Time Taken to Assess Three Operational Domains

	Less than 10 minutes	10–20 minutes	More than 20 minutes
Number of judgments	36	45	42
Number in agreement with researchers	30	38	28
Precentage in agreement with researchers	83	84	67

This finding demonstrates that at least some of the teachers were able to make assessment judgments swiftly and accurately, as recommended in the NDP teacher materials.

> As you become more familiar with the items and how to evaluate students' responses, you will become much quicker at administering the NumPA. You will get better at assigning stages for each area of strategy and knowledge from the assessment using the least possible number of questions. (Ministry of Education, 2005a, p. 1)

Reliability of Teacher Judgments in Assessment Interviews

A total of 156 teacher judgments were observed across the additive, multiplicative, and proportional domains. Table 3 summarises these judgments and compares them to those made by the two researchers. The two researchers had 100% agreement in their judgments.

Table 3
Agreement with Teacher Judgments

	Domain			Total
	Additive	Multiplicative	Proportional	
Number of judgments	70	45	41	156
Number of judgments in agreement	62	34	31	127
Percentage of judgments in agreement	89	76	76	81

The majority of teacher judgments (81%) agreed with those made by the researchers. The judgment decisions made in the additive domain showed a higher level of agreement with the researchers (89%) than those made in the multiplicative and proportional domains (76%).

Table 4 shows that in approximately two-thirds of the judgment differences observed, teachers rated students' strategy stages lower than the rating of the researchers. Anecdotally, teachers explained some of these decisions on the basis of consolidating students' understanding at an existing level, describing the student as "not ready" to move up to the next instructional group. In this way, teachers' assessment judgments were closely aligned to their classroom programmes and the instructional groupings of students. These assessment decisions differ from the NumPA instructions that direct teachers to judge students at the highest strategy stage demonstrated within each operational domain:

> enter the highest stage the student demonstrates within each operational domain. (Ministry of Education, 2005a, p. 3)

Table 4
Differences between Teacher and Researcher Judgments

Domain	Additive	Multiplicative	Proportional	Total
Number of differences	8	11	10	29
Number of differences where teachers rated lower	7	6	7	20 (69%)
Number of differences where teachers rated higher	1	5	3	9 (31%)

Most of the judgment differences between the teachers and the researchers occurred in stages 4 and above. Table 5 below shows where these differences occurred in all three operational domains. The shaded regions indicate where the teacher ratings are higher than the researcher ratings.

Table 5
Stage Ratings for Judgment Differences

	Researchers	Teachers	Number of incidences
Additive	Stage 5	Stage 4	6
	Stage 4	Stage 3	1
	Stage 6	Stage 7	1
Multiplicative	Stage 7	Stage 6	1
	Stage 6	Stage 5	3
	Stage 5	Stage 4	1
	Stage 4	Stages 2–3	1
	Stage 6	Stage 7	1
	Stage 4	Stage 5	4
Proportional	Stage 7	Stage 6	1
	Stage 6	Stage 5	3
	Stage 5	Stages 2–4	3
	Stage 4	Stage 5	3

The highest number of disagreements (6) occurred in the additive domain, where teachers rated a student stage 4 while the researchers assigned a rating of stage 5. When asked to clarify their decisions, the teachers said that the students predominantly used counting strategies and that they felt that using one simple partitioning strategy wasn't sufficient to rate them at stage 5 (early additive).

Of the teachers observed, 27 (72%) had taken part in the full NDP professional development programme. Three had received NDP professional development within the pick-ups programme[2], while the remaining seven had received within-school training. Table 6 reports on the variations in judgments between the groups of teachers with different training. The small sample sizes for those teachers without numeracy project training make direct comparison difficult, but there appear to be no significant differences in judgments between these groups of teachers.

Table 6
Reliability of Teacher Judgments and NDP Training

Training received	No. of teachers	No. of judgments	No. of agreements	% agreement
NDP training	27	114	93	82
Within school	7	30	25	83
Pick ups	3	12	9	75
Total	37	156	127	81

[2] The pick-ups programme refers to the national NDP workshop programme that teachers new to previously trained schools are invited to attend.

Reliability of Teacher Judgments Using Written Scenarios

Assessment scenarios were included in the questionnaire given to all teachers in the 2005 longitudinal schools. The scenarios were based on examples outlined in the diagnostic interview provided to teachers (Ministry of Education, 2005a). The teachers were instructed to rate the student at the stage of the Number Framework at which they appeared to be operating, based on the information given. Both the number and name for the stages were given with each scenario. A space was provided for the teachers to either comment on their choice of strategy stage or to suggest a question they might ask to further clarify the student's strategy stage.

An expert panel comprising three regional co-ordinators, a national co-ordinator, and three researchers discussed the scenarios and came to a consensus on which ratings could be considered reasonable, based on the limited evidence available. The panel also identified the types of questions that could be used in each scenario to confirm or question strategy judgments. Teacher ratings are compared with the decisions of the expert panel.

Of the teachers who responded to the scenario illustrated in Figure 1, 198 either made a singular stage rating or left the scenario unrated but indicated a further question was required. The numbers reported in Table 7 are based on these responses. There were 15 teachers who chose to identify a range of stages, and these responses have been excluded. Eleven of these 15 teachers identified the student as using a counting strategy (stages 0–3).

Teacher: Please hold out your hands for me. Here are four counters. Here are another three counters. How many counters have you got altogether?

Student: Four and three.

Figure 1. Scenario One

Table 7
Teacher Judgment of Student Strategy Stage for Scenario One

	Strategy Stage						
	0	1	2	3	4	5–7	No rating
Percentage of teachers without question	17	20	12	2	1	2	
Percentage of teachers indicating question	10	12	10	3	2		9

Shaded regions indicate those responses considered appropriate by the expert panel, giving a total of 78% agreement for scenario one. The panel believed that on the evidence given in scenario one, a rating of zero or one would be reasonable. Questions considered appropriate could clarify the student's counting ability by asking them to either count out a specified number of counters or to count how many counters altogether. The questions identified by teachers reflected both these themes:

> "Can you put them in my hand and point to them as you count?" "Can you carry on counting?" (count with child) Also repeat "Let's see how many we have altogether." (Teacher, left student unrated on scenario one)
>
> Can't determine – ask follow-up question: "How many is that altogether?" From that, you can determine what level they are at. (Teacher, left student unrated on scenario one)

The expert panel also believed it was reasonable to make no judgment on the student's strategy stage if a follow-up question was asked to clarify that the student needed to give the total number of counters in both hands. The responses of a further 9% of teachers can be considered in agreement on this basis. The panel also believed a rating of stage 2 could be considered in agreement if it was indicated that there had been a follow-up question asking about the total number of counters present. This accounts for another 10% of teachers.

Teachers commented that they felt most comfortable rating the scenarios that were in the same range as the strategy stages of the students in their classrooms. It is interesting to note that of the eight teachers who incorrectly rated scenario one at stage 4 or above, seven were middle-school (years 4–6) or senior-school (years 7–8) teachers.

> This [rating scenarios] is quite difficult. Teachers tend to focus on the stages they deal with daily! (Teacher)
>
> As a new entrant teacher, I would not be familiar with all the scenarios – however, I feel confident with testing stages at my level. (Teacher)

Of the teachers who responded to scenario two (Figure 2), 183 made a singular stage rating. The numbers reported in Table 8 are based on these responses. There were 20 teachers who chose to identify a range of stages, and these responses have been excluded. Nineteen of these 20 teachers identified the student as using an advanced strategy (stages 6–8).

Teacher: Ivan has 2.4 kilograms of mince. Each pattie takes 0.15 kilograms of mince. How many patties can Ivan make?

Student: 10 patties would be 1.5 kilograms, so 15 would be 2.25 kilograms, and one more would make 2.4 kilograms exactly. So that's 16 patties.

Figure 2. Scenario Two

Table 8
Teacher Judgment of Student Strategy Stage for Scenario Two

	Strategy Stage					
	3	4	5	6	7	8
Percentage of teachers without question	1	2	3	20	41	19
Percentage of teachers indicating question				2	9	3

Seventy-two percent of the teacher judgments were in agreement with the expert panel rating of stages 7 and 8 as appropriate. Follow-up questions considered appropriate were those of a similar nature to the original question but using different numbers to check whether the student was able to use another strategy. Teacher questions reflected this.

> Ask for another way of working this out. If they can use another strategy, mark student as level 8. (Teacher, rated student stages 7 to 8 on scenario two)

It is interesting to note that of the 27 teachers who did not answer this question, 18 were junior school teachers (years 0–3) and so would be less familiar with the advanced stages of the Number Framework. Also of note is the fact that teachers in the 2005 longitudinal schools may not have been aware of the addition in 2005 of stage 8 to the multiplicative domain of the Framework. Prior to 2005, stage 7 was the highest stage in the multiplicative domain.

Scenario three (Figure 3) received single-stage rating responses from 201 teachers. These results are summarised in Table 9. Of the 17 teachers who chose to identify a range of stages, 14 rated students at stages that included stage 5.

Teacher: There are nine counters under this card and eight counters under this card. How many counters are there altogether?

Student: 17.

Teacher: How did you work that out?

Student: I know that nine plus nine is 18, and it is one less, so that's 17.

Figure 3. Scenario Three

Table 9

Teacher Judgment of Student Strategy Stage for Scenario Three

	Strategy Stage				
	3	4	5	6	7
Percentage of teachers without question	3	14	64	5	1
Percentage of teachers indicating question	1	4	8		

A stage rating of 5 was considered appropriate by the expert panel, and 72% of the teachers were in agreement with this. The panel also believed it would be reasonable for further questions to probe the student's strategy repertoire, looking for evidence of a strategy other than doubles. The questions given by teachers were in accordance with this.

> "Can you think of another way to work this out without using your doubles?" (Teacher, rated student stage 5 on scenario three)
>
> I would ask another similar question with larger numbers and also ask if there was another way the student could solve this problem. (Teacher, rated student stage 5 on scenario three)

There were 194 teachers who gave a single stage rating in response to scenario four (Figure 4). The results presented in Table 10 are based on these responses. Of the 18 teachers who chose to identify more than one strategy stage, 11 of these responses identified students as being at a range of stages that included stage 4.

Teacher: Here is a forest of trees. There are five trees in each row, and there are eight rows. How many trees are there in the forest altogether?

Student: Um, 5, 10, 15, 20, 25, 30, 35, 40.

Teacher: If I planted 15 more trees, how many rows of five would I have then altogether?

Student: 5, 10, 15 [counting on fingers]. That's three.

Figure 4. Scenario Four

Table 10
Teacher Judgment of Student Strategy Stage for Scenario Four

	Strategy Stage				
	2–3	4	5	6	7
Percentage of teachers without question	5	62	16	3	
Percentage of teachers indicating question		9	3	1	1

The expert panel believed a rating of stage 4 was reasonable, and 71% of the teacher judgments were in agreement. The panel considered it reasonable for further questions to clarify that the question is asking for a total number of rows of trees. Teachers who rated the student at stage 4 and indicated a further question was required identified questions with this purpose.

> "So how many rows would I have altogether with the new rows and the rows I already had?" (Teacher, rated student stage 4 on scenario four)

Scenario five (Figure 5) received 184 responses from teachers making a singular stage rating. The numbers reported in Table 11 indicate the judgments made by these teachers. A further 18 teachers who identified a range of stages were excluded from the analysis. Sixteen of these 18 teachers identified students at stages that included stages 7 or 8. Of the 27 teachers who did not respond to this scenario, 25 taught years 1–4.

Teacher: There are 21 boys and 14 girls in Ana's class. What percentage of Ana's class are boys?

Student: Well if you add them together that's 35, so you can multiply them by three and that's pretty close to the percent. Twenty-one times three is 63, so it must be a little bit less than 63%.

Figure 5. Scenario Five

Table 11
Teacher Judgment of Student Strategy Stage for Scenario Five

	Strategy Stage				
	4	5	6	7	8
Percentage of teachers without question	1	3	18	42	23
Percentage of teachers indicating question		1	1	7	4

The expert panel believed a rating of either stage 7 or stage 8 was appropriate for this scenario, and 76% of teachers were in agreement. The panel agreed that effective further questions would focus on asking the student to calculate an exact answer to the problem. Teachers' follow-up questions reflected this.

> "An excellent answer. But now work out exactly the percent of boys in the class." (Teacher, rated student stage 8 on scenario six)

> "That's a good estimation. Can you tell me the exact percentage? Is there another way to work this out?" I would look at other strategies of working this out. (Teacher, unrated student on scenario six)

Table 12 summarises the levels of agreements across the five written scenarios, giving an average of 74%.

Table 12
Teacher Agreement with Expert Panel

	Agreement with expert panel
Scenario one	78%
Scenario two	72%
Scenario three	72%
Scenario four	72%
Scenario five	76%
Average agreement	74%

Concluding Comment

Most teachers in longitudinal classrooms track the strategy stages of students in their class, using a variety of approaches to determine students' strategy stages. In general, teachers report a high level of confidence in the assessment judgments they are making.

There was a high level of agreement between the teacher assessment judgments observed and those made by the researchers. The judgment decisions made in the additive domain showed a higher level of agreement than those made in the multiplicative and proportional domains. In approximately two-thirds of the judgment differences observed, teachers rated students' strategy stages lower than the rating of the researchers, explaining their decisions in terms of consolidating students' understanding at an existing level and aligning their judgments with the instructional groupings of students. The amount of time taken by teachers to make assessment decisions varied. Some teachers were able to make swift and accurate decisions, while those that took the most time tended to be less accurate.

Teacher ratings in the written scenarios were judged to be slightly less reliable than those that were observed, but agreement with the expert panel was still high. The slightly lower rating may be attributed to the limited information available to teachers in the written scenarios.

References

Bobis J. (1997). *Count Me In Too 1997 report.* Retrieved 5 December 2005 from www.curriculumsupport.nsw.edu.au/maths/countmein/pdf/97_report.pdf

Higgins, J. (2004). *An evaluation of the Advanced Numeracy Project 2003:* Exploring issues in mathematics education. Wellington: Ministry of Education.

Ministry of Education (2005a). *Book 2: The diagnostic interview.* Wellington: Ministry of Education.

Ministry of Education (2005b). *Book 3: Getting started.* Wellington: Ministry of Education.

Ministry of Education (2005c). *The numeracy story continued: What is the evidence telling us?* Wellington: Ministry of Education.

Nystrom, P. (2004). Reliability of educational assessments: The case of classification accuracy. *Scandinavian Journal of Educational Research, 48,* 427–440.

Parsons, R. (2005, March). *Numeracy Development Project: Scope and scale.* Paper presented at the Numeracy Development Project Reference Group, Wellington.