Overall Summary of Teacher Feedback from the Calibrations and Trials of the asTTle Reading, Writing, and Mathematics Assessments

Technical Report 33, Project asTTle, University of Auckland, 2002

Lyn Lavery & Gavin T L Brown
University of Auckland

This report shows how evaluative feedback from teachers was used to improve the quality of assessment materials in mathematics, reading, and writing. Data were collected from the trial and standardisation of asTTle test papers conducted between October 2000 and June 2002. Feedback was used to improve further test forms, test items, and the asTTle software. Teachers identified as their most serious criticism of the asTTle test papers the mismatch in test paper difficulty with the ability of all students in their classes. The final asTTle software will allow teachers to customise test difficulty for the ability of their own students.

Table of Contents

Formative evaluation is "designed, done, and intended to support the process of improvement" (Scriven, 1991, p. 20) and is normally carried out "by the staff of the originating institution" (Scriven, 1991, p. 22). Formative evaluation provides improvement-oriented feedback in response to three specific questions, which are (a) Where is the process or product now?, (b) What is the target for the process or product?, and (c) What steps are needed to reach the target? (Clarke, Timperley, & Hattie, 2003). It is critical in the development of a national teacher-controlled assessment tool that evaluative comment is obtained from the intended users. Research on teachers' conceptions of assessment has shown that teachers are positive about obtaining improvement oriented evaluation of their teaching through student teaching but that they also consider externally mandated assessments as measuring only surface dimensions of learning (Brown, 2002). Thus, this report describes the formative evaluation conducted during the development of standardised assessment items in literacy and numeracy and how the feedback was used to improve the quality of the assessment materials to ensure maximum impact on teachers' use and acceptance of the new assessment tools.

The asTTle Project

The Assessment Tools for Teaching and Learning (asTTle) project delivers a computer based set of tools for classroom, teacher-controlled assessment of student progress in literacy and numeracy at Levels 2 – 4 of the New Zealand curriculum in both English and Te Reo Maori. Specifically, this includes Reading, Writing, Mathematics, Panui, Tuhituhi, and Pangarau. Once released by the Ministry of Education to schools, asTTle is designed to be used as a classroom assessment by teachers.

The asTTle tools provide teachers with the ability to track progress and achievement of

individual students or groups/subgroups of students. Teachers design an asTTle test by selecting the curriculum areas and levels of difficulty that they wish to assess. These selections are maximised by the asTTle tool to create a 40-minute pencil and paper test consisting of a mixture of open- and closed-response items. Once student responses and scores are entered into the asTTle tool teachers may select a range of reports that allow them to interpret student performance by reference to nationally representative norms, curriculum levels, and curriculum achievement objectives. Specifically, asTTle answers questions related to (a) how well are students doing compared to similar students, (b) how well are students doing on important achievement objectives, (c) how well students are doing compared to curriculum achievement levels, and (d) what are some teaching resources that would assist in improving students' performance.

Methodology

New Zealand teachers participated in a variety of workshops around the country run by the asTTle team to write and review assessment items for reading and writing. Their task was to write and review assessment materials appropriate to the interests, achievement objectives, and ages of students in Years 5-7 learning in curriculum levels 2-4 consistent with the principles of the appropriate curriculum document and consistent with good classroom practice. Once reviewed, items were assembled into test forms estimated to require 40 minutes for the majority of students. Two different strategies were used in assembling test forms. In reading and writing, each form had an approximately equal distribution of items or tasks across the Curriculum Levels 2 to 4 and were administered to students at any of the target year levels. In contrast, the mathematics test forms, although containing materials from all three levels, were designed to have more of a certain level for use in each of the target year levels. The Year 5 papers had more Level 2 items, the Year 6 papers had more Level 3 items, while the Year 7 papers had more Level 4 items. Note that the test forms as administered were balanced in quite a different fashion to the asTTle computer tool which allows teachers to custom select the difficulty desired regardless of the year or age of students. For example, a teacher using asTTle can create a test with no or few hard items for a younger or less able group of students and vice versa.

Trials and calibrations for asTTle assessments across the reading, writing, and mathematics domains in both languages were conducted on behalf of asTTle by classes of students in New Zealand schools. Each trial and calibration was administered by teachers with their own class of students. Between October 2000 and June 2002 data were collected for English reading, writing, and mathematics a total of ten times (i.e., six for literacy and four for numeracy). Trials were conducted on relatively small samples to identify dimensions of the items and test forms that might need modification. Calibrations were conducted after the trials with large nationally representative samples for the purpose of establishing New Zealand student performance norms. Table 1 shows that nearly 50,000 students completed the trials and calibrations for all three English subjects.

Table 1
Trial and Calibration Student Sample by Subject

Subject	Sample Size
Reading	24,513
Writing	10,377
Mathematics	13,397
Total	48,287

In Maori, there are about 2,500 students in Maori-medium instruction in each of the target years. In order to prevent over usage of this small number of students, trials and calibrations were combined so that data were collected three times between November 2001 and June 2002. Note that this report is based on the feedback and teacher evaluative comments to the English reading, writing, and mathematics assessments only.

Each teacher who administered the tests was asked to complete a form that structured feedback around key item and tool development characteristics. The key characteristics for which evaluative feedback was sought revolved around the quality of the items, specifically their appropriateness in terms of interest or engagement, difficulty, length of time needed to complete, and around the quality of the instructions supplied with the test forms. In addition, teachers were asked for their own general comments, and were asked to summarise the nature of students' experience and evaluation of the materials. It was intended to use the teacher feedback to modify items for use in the asTTle tool by adjusting the language, content, or nature of items, the length of time allowed for completing items, and for improving administration instructions.

Results

The total number of teachers possible to participate in the feedback was calculated as one teacher per 30 students. On that basis, approximately 820 responses related to reading, 350 responses related to writing, and 450 responses to mathematics could have been expected. From the most frequently asked questions it is possible to identify that at least 459 reading teachers, 483 writing teachers, and 333 mathematics teachers participated. This translates to about 56% of reading teachers, 138% of writing teachers, and 74% of mathematics. The high number of writing responses suggests that some teachers completed more than one form.

Responses were in the nature of comments to prepared questions. The comments were generally coded using a "Yes", "No", "Both yes and no", or "No answer". The category "Yes" indicates a favourable or positive response to the question, a "No" an unfavourable or negative response to the question, and "Both yes and no" indicates a response that contains both positive and negative comments. "No answer" includes comments that were incapable of meaningful interpretation.

Data for this report come from previously reported results (Zwiegelaar, 2000; Langstaff, 2000; Irving, 2001; Parker, 2001, Parker & Brown, 2002; Zwiegelaar & Brown, 2002; Schouten & Brown, 2002; Lavery & Brown, 2002) of teacher feedback to the asTTle test forms and items. The data, although not always collected in the same manner across instances, have been converted to a common format for this report. For example, the student response and general comments questions in both the reading and writing domains had to be altered in order to be combined. As some responses had been coded per teacher and others had been reported by total number of comments, any responses reported in the latter format were converted to percentages and then calculated into frequency counts based on the number of teacher responses. Furthermore, as not all questions were asked in each trial or calibration, questions are summarized by topic.

Content Appropriateness

This section asked whether the content was appropriate for students' age and ability in the teacher's class. Overall, teachers were positive regarding this, particularly in the reading and writing domains (Table 2). While many teachers simply gave a 'yes' response to this question, others were more descriptive. For example, in reference to the mathematics assessments, one teacher wrote "Yes, the range allowed for <u>all</u> students to be able to answer some questions. Content was familiar to them and their understanding".

Table 2
Appropriateness of content

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	195 (59)	41 (12)	88 (26)	9 (3)	333 (100)
Reading	327 (71)	51 (11)	64 (14)	17 (4)	459 (100)
Writing	350 (72)	52 (11)	72 (15)	9 (2)	483 (100)

Across all three domains, when a negative or mixed response was indicated, this generally reflected the mixed ability levels in the class or the fact that a small proportion of the students (especially those with a language other than English at home or with low ability) experienced difficulties. This criticism is an artefact of the method used in assembling test forms for calibration. The flexibility of asTTle test creation means that teachers will be able to assemble tests that are customised for these special cases.

Content Interest and Engagement

Teachers were very positive overall about the interest and engagement of the content in the assessment items and tasks (Table 3). Positive comments included, "All the students found it interesting and wanted to keep doing it" (mathematics assessments), "Yes, children showed a keen interest. Good variety, visually good and the use of different genre" (reading assessments), and "They enjoyed it a lot and found it challenging, especially deciding what to write about" (writing assessments).

The responses in the 'both' category tended to reflect the inappropriate difficulty spread in the test form rather than a criticism of the actual content. While most of the class were interested and engaged, usually a small proportion of others were not because of the mismatch of difficulty to ability. For example, "The children who have ability in maths enjoyed it. But the less able children totally switched off" and "Children engaged at the beginning and started to lose interest when difficulty increased" (reading assessments).

Negative comments made up a low proportion overall and were often qualified as being due to external environmental factors, particularly in the first mathematics calibration; for example "We have just done a lot of tests so they were not focused".

The feedback identified to the asTTle team the benefit of using classroom teachers as the original source of material and tasks. Just as importantly, it showed the positive impact of the use of desk top publishing for illustration and layout of materials – the fact that generous amounts of white space were used contributed to the children's positive interest and engagement.

Table 3
Content interesting and engaging

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	256 (77)	28 (8)	39 (12)	10 (3)	333 (100)
Reading	368 (80)	36 (8)	38 (8)	17 (4)	459 (100)
Writing	321 (66)	69 (14)	81 (17)	12 (3)	483 (100)

Teacher Instructions

In most of the calibrations, teachers were asked whether the teacher administration instructions were easy to use and adequate (Table 4). Responses were positive across all three domains with the majority of teachers simply responding by indicating "Yes". While a few negative comments were made, these were often because they were suggestions for improving the instructional materials.

Table 4
Teacher Instructions

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	216 (78)	25 (9)	29 (11)	6 (2)	276 (100)
Reading	152 (86)	5 (3)	19 (10)	1 (1)	177 (100)
Writing	144 (80)	18 (10)	16 (9)	1 (1)	179 (100)

This feedback was incrementally implemented across time and has resulted in a concise set of administration guidelines that are delivered to teachers every time asTTle creates a test.

Time Allocation

Teachers in the two initial mathematics trials and the second calibration of writing assessments were asked whether the amount of time allowed to complete the test form was appropriate. Table 5 shows that almost three-

quarters of teachers involved with the mathematics assessment did not feel the time allocation was appropriate. Their comments reflected the fact that the amount of work required in the time allowed was too much. It was apparent that the developers had underestimated the time needed to compete tasks.

Table 5
Time Allowance

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	8 (13)	42 (74)	6 (11)	1 (2)	57 (100)
Writing	31 (59)	16 (31)	3 (6)	2 (4)	52 (100)

In contrast, approximately two-thirds of teachers involved with the writing assessments felt the time allowed was acceptable, while a further third did not. Of those that did not feel the time frame was adequate, respondents were reasonably evenly split between those that felt it was too long and those who felt it was too short.

Based on the evaluative comments it was decided to allow 20% more time for open-ended items over multiple-choice selected-response items. However, the actual amount of time needed for a student to complete assessment items is dependent on the interaction of the student's ability, interest, and motivation and the item's difficulty and content. The asTTle tool will design a test centred around the majority of students completing in 40 minutes, but it will always be up to the professional judgement of teachers to accept and administer tests with appropriate difficulty for the ability of their own students.

Level of Difficulty

In most trials and calibrations, teachers were also asked whether the difficulty level was appropriate for all students. Two thirds of the teachers believed the literacy tests had appropriate difficulty while only one-third of mathematics teachers agreed (Table 6). While the overall percentage of teachers giving positive responses was much lower for this particular question, it is important to note that this reflects the fact that it is difficult to design a single test paper that is suitable for all students. The difficulty level was either too easy or too hard depending on the teacher's class. This is well summed up in the following teacher's response from the mathematics assessments: "There is an implication that all students on a particular year are at the same level – this is not the case. Therefore – no, it was not appropriate for all students".

Table 6
Level of difficulty

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	100 (36)	71 (26)	91 (33)	14 (5)	276 (100)
Reading	306 (67)	82 (18)	45 (10)	26 (5)	459 (100)
Writing	307 (64)	87 (18)	71 (15)	18 (3)	483 (100)

Despite the difficulties in designing an assessment instrument suitable for all students, of the teachers responding positively, some commented "It allowed the less able to participate but also challenged the more able" (reading assessments), and "Yes, I think there was a good range of questions which provided a variety of skills" (writing assessments).

Once again, comments in the 'no' and 'both' categories often reflected mixed abilities in the classroom, for example, "Yes, for the mid-range. Does not allow for students above/below average".

Again it is worth reiterating that in the released asTTle software, teachers will be able to customise assessments for the ability of not only whole classes but also for individual students. The large difference between teachers' reaction to the difficulty of the numeracy and literacy assessments merits further research.

Student Response

Teachers tended to comment favourably when they were asked to report the overall response of their students to the asTTle test papers (Table 7). Overall, approximately half the teachers indicated that their classes responded positively to the assessment. For example, "Children commented that it was 'cool' – the diagrams were interesting and kept them focused." (mathematics assessments), and "They enjoyed it and found it easy because they could choose the topic for instructions [a writing task in which students had to write instructions on how to do something] themselves, without being asked to write about something or answer questions about which they have no experience" (writing assessments).

For the mathematics and writing assessments, almost one third of teachers' responses were classed in the "both' category, usually reflecting a mixed response from their class. Across all subjects, negative comments often reflected length or difficulty, for example, "The students' response to the paper was that it was too long and boring. When asked why it was boring the answer was they didn't know how to do things, questions were hard and difficult to understand" (mathematics assessments).

Again, with the actual asTTle tool, teachers will be able to ensure that the difficulty of the assessment is appropriately set for students.

Table 7
Student responses to paper

Subject	Number of Responses (Percent)				
	Yes	No	Both	No answer	Total
Mathematics	165 (50)	55 (16)	96 (29)	17 (5)	333 (100)
Reading	285 (62)	113 (25)	52 (11)	9 (2)	459 (100)
Writing	257 (53)	96 (20)	120 (25)	10 (2)	483 (100)

General Comments

When asked if there were any general comments they would like to make, teachers made a variety of responses (Table 8). These were reasonably evenly mixed between positive, negative, both, and suggestions for improvement. However, a slightly larger proportion of teachers made positive comments regarding the reading assessments while a slightly larger proportion made negative comments regarding the writing assessments.

Table 8
General comments

Comment	Number of Responses (Percent)		
	Mathematics	Reading	Writing
Positive	44 (14)	91 (38)	83 (19)
Negative	45 (15)	41 (17)	156 (36)
Both/ Neutral	38 (13)	22 (9)	27 (6)
Suggestions for Improvement	50 (17)	21 (9)	10 (2)
No answer/ Uninterpretable	123 (41)	66 (27)	160 (37)
Total	300 (100)	241 (100)	436 (100)

Concluding Comments

The asTTle test items and materials in reading, writing, and mathematics were well received by both teachers and students in terms of content, interest and engagement, and for the instructions supplied to teachers. The test forms were less positively rated in terms of difficulty and time required, with the mathematics assessments being somewhat less positively regarded than the literacy ones.

The evaluative feedback was used by the development team to adjust the length of time allotted to completing each item, to adjust teacher instructions, and most importantly in the design of the asTTle test creation process. The overriding negative feedback was about inappropriate difficulty in the test forms which were not custom designed to the students of each teacher's class. However, with the asTTle software teachers will be able to design and administer custom designed tests for the ability of classes and students and thus get around the fault of one size not fitting all.

The evaluative feedback collected by the asTTle development team has been used to improve the quality of the asTTle materials and software. Furthermore, the feedback

indicates clearly that users can have confidence in asTTle's assessment material to engage and motivate students to show their true performance provided teachers have administered tests of an appropriate difficulty level.

References

Brown, G. T. L. (2002). *Teachers' Conceptions of Assessment.* Unpublished doctoral dissertation, University of Auckland, Auckland, NZ.

Clarke, S., Timperley, H., & Hattie, J. (2003). *Unlocking formative assessment.* Auckland, NZ: Hodder Moa Beckett.

Irving, E. (2001). *Evaluation of the teacher feedback from the first calibration of reading and writing assessments.* (Technical Report 7). Auckland, NZ: University of Auckland, Project asTTle.

Langstaff, J. (2000). *Evaluation of the teacher feedback from the second trials of reading and writing assessments.* (Technical Report 3). Auckland, NZ: University of Auckland, Project asTTle.

Lavery, L., & Brown, G. T. L. (2002). *Report on teacher feedback from the third calibration of writing assessments* (Technical Report 32). Auckland, NZ: University of Auckland, Project asTTle.

Lavery, L., & Brown, G. T. L. (2002). *Summary of the Teacher Feedback from Trials of English Mathematics Assessments* (Technical Report 24). Auckland, NZ: University of Auckland, Project asTTle.

Parker, A. M. (2001). *Evaluation of the teacher feedback from the link calibration of reading and writing assessments.* (Technical Report 8). Auckland, NZ: University of Auckland, Project asTTle.

Parker, A. M., & Brown, G. T. L. (2002). *Evaluation of the teacher feedback from calibration 2 of English literacy assessments.* (Technical Report 19). Auckland, NZ: University of Auckland, Project asTTle.

Schouten, J., & Brown, G. T. L. (2002). *Report on teacher feedback from the first calibration of mathematics assessments* (Technical Report 31). Auckland, NZ: University of Auckland, Project asTTle.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation & education: At quarter century* (Vol. Part II, pp. 19–64). Chicago: NSSE.

Zwiegelaar, J. (2000). *Evaluation of the teacher feedback from the first trials of reading and writing assessments.* (Technical Report 2). Auckland, NZ: University of Auckland, Project asTTle.

Zwiegelaar, J. B., & Brown, G. T. L. (2002). *Teacher evaluation of the reading level 4 assessments: Summary* (Technical Report 30). Auckland, NZ: University of Auckland, Project asTTle.