Formative Evaluation of an Educational Assessment Technology Innovation: Developers' Insights into Assessment Tools for Teaching and Learning (asTTle)

John A. Hattie, Gavin T. L. Brown, Lorrae Ward,
S. Earl Irving, and Peter J. Keegan

The University of Auckland, Auckland, New Zealand

Notes

The asTTle[1] Project is a New Zealand Ministry of Education funded research and development program at the University of Auckland. The authors would like to acknowledge the feedback of thousands of teachers and students who contributed to improving the asTTle test materials, reports, professional development (PD) efforts, and software, and thank the many staff who helped develop and implement asTTle Versions 1 to 4.

[1] Pronounced "astill"

Abstract

Formative evaluations conducted by the development team, focusing on the consequences of choices to be made by the developers, during the development of asTTle, a new ICT[2]-mediated educational assessment resource, are reported. The evaluations focused on the validity, accuracy, added value, training, and utility standards identified as important in the deployment of educational technology. As a result of in-house utilization of the evaluations, it is argued that users can have confidence that asTTle has demonstrated validity through alignment with the curriculum and classroom practice, that the reporting systems add value to teachers' work and that accuracy of understanding is enhanced by professional development. Insights into effective PD have been obtained, though further improvements in the training component of the system are required. The utility of the software has been constantly increased, but there are still unresolved issues that are currently being responded to. The gradual implementation mechanism has been successfully used to obtain user buy-in as well as identification and response to identified needs. Users and funding agencies can have confidence that asTTle has been designed, developed, and implemented in a fashion consistent with maximizing benefits to the end-user. The software has been developed in such a way that the validity, added value, accuracy, training and utility standards for educational technology have been met, at least in part. The data reported here demonstrate again that formative evaluations which maximize end-user benefits can add significant value to an educational technology innovation.

[2] Information & Communications Technology

The introduction of computers into schools has been rapid and extensive, but the impact on teacher instructional practice or classroom activities has been generally limited (Cuban, 2001; Hedges, Konstantopoulos, & Thoreson, 2003). The major obstacle to successful implementation is not so much the provision of hardware but teachers' personal views of how learning occurs, what teachers believe curriculum objectives are, and school environmental factors (such as infrastructure and teacher competency) (Compton & Jones, 1998; Lewis, 1999; Means, Haertel, & Moses, 2003; Parks, Huot, Hamers, & H.-Lemonnier, 2003). These factors produce great variation in the quality and quantity of how any innovation is deployed. Watson (2001) argued that lack of clarity in purpose for educational technology at the national curriculum level contributed to uneven uptake.

Institutional barriers to adoption of computer-assisted assessment technologies have included such aspects as personal time cost, unrealistic expectations, inherent conservatism, and lack of technical and pedagogic support or training (Bottino, Forcheri, & Molfino, 1998; Conole & Warburton, 2005; North, Strain, & Abbott, 2000). Fichman's (1992) review of information technology diffusion identified strongest results when the IT was relatively simple to adopt and when an individual, rather than an institution, was responsible for the adoption. In a like manner, the design of the technology's interface and usability also impact on effectiveness of educational technology (Bourges-Waldegg, Moreno, & Rojano, 2000). Finally, Baker and Herman (2003) identified any automation of assessment capacity as an intervention trying to transform the nature and functions of the teaching and learning process and thus one that would be difficult to evaluate or determine effectiveness.

The evaluation studies reported here constitute a description of some of the "necessary preconditions needed for the innovation to produce the desired learning outcomes" (Lesgold, 2003, p. 65). Such contextualized evaluations examine the context of implementation and the way the innovation unfolds within a complex organization. The use made of these contextual evaluations after identifying whether the implementation of the innovation provided to the intended users the quality of experience that it was intended that they would receive; was to subsequently use that information to further refine and improve the product. Explicit attention to maximizing end-user beneficial consequences permeated these studies (Sen, 2000). Often, improvement-oriented evaluations of educational technology are ineffective because the data come too late, systems are too inflexible to be revised in the light of new data, there is a limit to the resources needed to make the changes, or it is unclear who makes the changes (Baker & Herman, 2003). These evaluation studies were conducted throughout the development process, and the overall approach to evaluation taken in this report is consistent with improvement (Posavac & Carey, 1997), formative (Scriven, 1991), consequence (Sen, 2000), discrepancy-analysis (Provus, 1973), utilization (Patton, 1978), and product (Stufflebeam, 1983) focused evaluation models.

Context

The innovation that is the subject of this report is the New Zealand Assessment Tools for Teaching and Learning (asTTle) educational software (Hattie, Brown, & Keegan, 2003). This software automates assessment capacity not by delivering interactive on-screen test questions, tasks, or portfolios as suggested by several educational technologists (Carroll, 2001; Riel, 2001), but by integrating computer-aided test assembly with computer analysis of test performance into a teacher-

controlled, curriculum-based system that graphically reports student achievement against norms, objectives, and standards. Thus, in Conole and Warburton's (2005) terms, asTTle is a combination of computer-based and computer-assisted assessment with networking capability. The currently deployed system is a necessary step, according to Raikes and Harding (2003), in moving from an environment in which only paper-based tests are possible to the full promise of technology—that is innovative, multi-media product and performance assessments delivered through computer-adaptive testing, automatic IRT-based scoring, and rich, interactive reporting and communication. The asTTle system is designed to provide teachers and administrators with feedback as to where children are, relative to curriculum and teaching objectives, and to indicate possible teaching and learning resources for the next steps (Hattie, 1999). The heart of the asTTle technology is a reporting engine which allows teachers to see what student performance on the assessment tasks means. The emphasis is thus on the consequences of testing and the validity of the testing is a function of the accuracy and usefulness of the various reports provided to the teachers, students, and school leaders.

The asTTle tools provide teachers and school leaders with the ability to analyze 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. The asTTle software 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. The application then uses a sophisticated linear programming heuristic 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 students are doing compared to similar students, (b) how well students are doing on important achievement objectives, (c) how well students are doing compared to curriculum achievement levels, and (d) what teaching resources would assist in improving students' performance.

Often technological innovations fail because of external environmental factors such as limited availability of equipment, support by management, skilled use by teachers, and appropriate timing and integration with curriculum (Baker & Herman, 2003). asTTle was designed to address these concerns by functioning on the typical kind of computers known to be in schools, by being adopted on a voluntary basis, by being supported with in-person and on-line professional development, and by being fully integrated with the national curriculum statements. Four versions of the asTTle software have been developed and released to New Zealand schools with each version increasing the range of content and features (Brown & Hattie, 2005).

asTTle Version 4 (V4) permits assessment of student performance in reading, writing, mathematics and the Mäori equivalents of pānui, tuhituhi, and pāngarau across Levels 2-6 of the New Zealand curriculum. Released by the Ministry of Education for schools to use on a voluntary basis, asTTle can be used for classroom decisions by teachers and for school-wide planning and reporting decisions by administrators and governors. The tool contains over 4,000 assessment questions and nearly 100 writing or tuhituhi prompts as well as

performance data from a nationally-representative sample of over 92,000 students. The technology functions on both Apple Macintosh and PC computers and the multi-user versions support Windows, Mac, Linux, and Novell servers. The application can be used as a single-user or networked version and interoperability with school student or information management systems is supported through manual and automatic data transfer systems. Professional development resources are made available on the installed computer and on the asTTle web site (http://www.asttle.org.nz); resources include a multi-chapter manual, context-sensitive help, FAQs, technical reports, and Macromedia Breeze presentations with voice-over and notes. Further assistance to schools is provided by a Ministry of Education free-phone technology help desk, and by Ministry funded assessment-focused in-service, professional development providers.

Evaluation Methods

A guiding principle throughout the development has been attention to evaluation – particularly by the end-users (the teachers, school leaders, and students). The themes of usability, interoperability, training, value, and accuracy have also been guiding principles, along with a rigorous measurement theory underlying the development of the items, scoring, and test creation. This article outlines the multiplicity of evaluation methods relating to teacher and student evaluation of the test materials, the accuracy and added value of the reports, the reactions and effects of the professional development, and the evaluation of the utility of the software. Unlike many evaluations of tests, the current study is as concerned with the consequences of the tests (Messick, 1989), and specifically with how teachers are using, interpreting, and modifying their thinking and teaching as a function of using the asTTle tool.

Methods

In order to evaluate the product, standards against which user experience could be evaluated were developed. In these evaluation studies, the developers focused on the first three of Lesgold's (2003) standards for mature software. Specifically, the ability to inter-operate with other software in the environment, the provision of training, and the interface's features (i.e., is it easily mastered, understood, and used?). Baker (2005) suggested that efficiency and quality are required of technology in order to add educational value. The validity, accuracy, and utility of a technology contribute to determining its quality. The evaluations conducted by the asTTle development team were designed to determine the following aspects of the system:

Validity. The items and materials used as assessments had to have integrity within both the curriculum and teacher classroom realities.

Utility. The software had to be easy to use especially as its use was voluntary—if the technology adopted made it difficult to use, then it would be unlikely that the product would be of much value This included the requirement for compatibility or interoperability with other school technology systems and for the system to be resilient as it was developed (Baker, 2005).

Added Value. The system had to create value for the teachers; if asTTle did not ease workload or improve the quality of teacher decision-making then, no matter how easy it might be to use, it would be of lesser value, and less likely to be used.

Accuracy. The developers, having designed and implemented creative reporting mechanisms, needed to know whether users were interpreting and using the educational reports correctly.

Training. The professional development processes and resources employed to enable users in implementing the product well had to be effective.

In order to determine whether the asTTle software met these objectives, multiple studies using multiple methods were conducted. Questionnaire surveys were used to elicit from teachers their evaluations of the asTTle test questions, tasks, and instructions. Similarly, students were surveyed about their opinions of the test materials at the end of a trial test. Focus groups and survey questionnaires were used to determine the accuracy of understanding teachers had of the asTTle reports. Telephone interviews, questionnaire surveys, and field visits were used to examine the effectiveness of the professional development resources and processes and the software itself. Thus, this article reports a series of management-oriented evaluations conducted by the development team for the purpose of improving the quality of the product and the supporting training materials and processes supplied with the product.

Results

Results for this series of evaluations are reported in four sections. Section 1 summarizes teacher and student feedback as to the qualities of the asTTle test materials. Section 2 summarizes findings about teacher understanding of the asTTle reports. Section 3 reports findings about the professional training and support processes, while Section 4 reports results about school uses of the asTTle software.

Validity—Teachers' & Students' Evaluations of the asTTle Test Materials

Teacher Evaluations

New Zealand teachers participated in workshops to write and review assessment items for all subjects, and many administered the asTTle trial tests with their own classes during the standardization calibrations that took place between 2000 and 2004. In the norming sample, over 90,000 students from over 700 schools participated, involving over a thousand teachers. The asTTle test forms were designed to take 40 minutes to complete and were intended to be appropriately difficult for the majority of students to whom they were assigned.

Between half and three-quarters of all teachers who had administered the tests completed a survey questionnaire about 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 to summarize the nature of students' experience and evaluation of the materials. This feedback was used to modify items for use in the asTTle tool by adjusting the language, display format, content, or nature of items, the length of time allowed for completing items, and for improving administration instructions.

Responses were in the nature of responses to prepared questions. The comments were generally coded using a "Yes", "No", "Both yes and no", or "No answer". The category "Yes" indicates a favorable or positive response to the question, a "No" an unfavorable 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 could not be meaningfully interpreted.

Approximately two-thirds of responding teachers agreed that the content was appropriate across all subjects ($M = 67\%$, $SD = 8\%$), with only 10% ($SD = 4\%$) indicating that it was not appropriate. Some teachers 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). The negative responses were often associated with comments that some students (e.g., English language or low progress learners) experienced difficulties.

Nearly two-thirds of teachers ($M = 61\%$, $SD = 12\%$) agreed that test forms were of a suitable difficulty level for the whole class, with 16% ($SD = 9\%$) indicating that they were either too hard or too easy. It is important to note that this lower positive rating came about because it was difficult to design a single test paper that was suitable for all students in any one class or group. This deficiency would be resolved in the implementation of asTTle, where teachers could custom-fit the difficulty of a test to groups of their own students.

Nearly three quarters of teachers agreed that the content of the tests was interesting and engaging ($M = 72\%$, $SD = 12\%$), with only 7% ($SD = 6\%$) indicating that they were not so. 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 genres" (reading assessments), and "They enjoyed it a lot and found it challenging, especially

deciding what to write about" (writing assessments). This feedback gave confidence to the interpretation that the materials fit well with classroom realities.

Just over four out of five teachers ($M = 82\%$, $SD = 1\%$) indicated that the administration instructions were easy to follow and adequate. Requests for clarification in terms of language used in the writing rubrics and layout of the writing rubrics were incrementally addressed, though it should be noted that a full on-screen, web-style interface is probably necessary to adequately meet the need for real-time, on-the-spot, customized help. About half of the teachers ($M = 52\%$, $SD = 11\%$) reported that their students had had a favorable experience completing the tests, with only about a fifth ($M = 18\%$, $SD = 8\%$) noting a negative experience. On the positive side, teachers indicated, 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). The negative comments suggested that it was not so much the tests themselves that were being negatively evaluated, as factors that, under operational conditions, would not exist if the teacher exercised personal control of the asTTle test creation and administration systems: "some students struggled with the paper when they had to write an argument", "students were comfortable with some aspects, but lacked the narrative writing ability", "the test is probably too long for the age group" "not enough space for story planning or the story", "time allowance for planning/writing was too long", "some really liked the test. A few didn't like it—[they] don't like any test". These criticisms are common-place for externally

controlled and imposed testing systems which is how the trial tests were experienced by some teachers and students. But, once operational, teachers had the power to select only content that suited their teaching programs and had the power to modify administration procedures to ensure that timing and difficulty were appropriate. This feedback gave the developers some confidence that the teachers perceived the students as reacting positively to the materials.

Student Evaluations

Additionally, in four subjects a sample of students completed a set of ten items in which they rated the quality of the items from their own perspectives. By subject the following approximate numbers of students provided data: mathematics 1000, reading 1400, pānui 1800, and tuhituhi 1300. Students responded to the items by indicating the degree to which they agreed with the statement with a positively packed rating scale (1 = Strongly disagree to 6 = Strongly agree, with two points expressing disagreement and four points for agreement) (see Brown, 2004a for details). Maximum likelihood factor analysis with direct oblimin rotation was conducted with the 10 items for each subject. The resulting scale scores were calculated and used to infer student evaluation of the asTTle materials.

Mathematics. A three factor solution (i.e., Student Enjoyment, Layout of the Test, and Student Confidence to do the Questions) was found. Students expressed only slight agreement ($M = 2.87$, $SD = 1.30$) with the factor related to student enjoyment in doing the questions and test and gave almost the identical level of agreement ($M = 2.90$, $SD = 1.01$) to their confidence in doing the assessment items. The students barely enjoyed doing the tests and were also only just confident that they could do the questions. In contrast, students moderately agreed with the layout of

the paper and the use of white space ($M = 4.30$, $SD = 1.24$), supporting asTTle's layout design of pages. The negative response indicated that despite best efforts of the asTTle developers, students perceived that the items were challenging and thus not really enjoyable. Under operational conditions, teachers would be able to adjust the difficulty and challenge of an asTTle test to meet the students' lack of confidence. It should also be noted that mathematics is the only area in the asTTle testing where the norm population exhibited an inverse relationship between confidence and achievement, a result echoed by this finding. In other words, New Zealand (NZ) students who are good at mathematics still lack confidence in their own abilities and thus were daunted by the challenge embedded in the trial tests. It is expected that in classroom operation, teachers will be able to address this psychology before test administration.

Reading. A three factor solution (i.e., Student Enjoyment of the Material and Its Layout, Difficulty of the Materials, and Test-Likeness of the Experience) was found. Students expressed between slight and moderate agreement ($M = 3.53$, $SD = 1.08$) that they had enjoyed the tests and their appearance, and gave a similar rating ($M = 3.43$, $SD = .68$) to the idea that the tests were harder than those they had already done in class. They gave a slightly stronger rating ($M = 4.30$, $SD = 1.24$) to the idea that the asTTle test felt like a test on which they exerted their best effort. Unlike mathematics, these students were somewhat more positive about their ability to do the items. The reading students' overall opinion of the test fell between the mathematics students' ratings about the layout and design of the test and their enjoyment in doing the tests—a somewhat similar result as the reading factor contained both those concepts.

Pānui. A two factor solution (i.e., Test Difficulty, and Student Enjoyment of the Test and Its Layout) was found. Almost identical scores at the level of moderate agreement were found ($M = 3.99$, $SD = 1.33$ and $M = 3.97$, $SD = 1.13$ respectively). In other words, the Māori students found the tests about as enjoyable, but somewhat harder than the English-medium reading students found their tests.

Tuhituhi. A three factor solution (i.e., Enjoyable Test, Beneficial Use of Space, Confidence in Doing Well) was found. The students gave moderate levels of agreement to all three factors ($M = 4.20$, $SD = 1.13$; $M = 4.24$, $SD = 1.38$; $M = 3.95$, $SD = 1.25$ respectively). These values are fundamentally the same as the pānui and reading values and confirm that the students perceived enjoyment in doing the asTTle test, liked the use of white space on the test forms, and thought that they could do well.

These studies confirmed what the teachers had reported earlier—students were generally favorably disposed towards doing the tests, liked the use of white space, and were reasonably confident of doing well. The exception to this pattern is the mathematics students who were considerably less confident in their ability to perform well on the test and who did not enjoy the testing process. The developers took that negative message to be a consequence of a centrally designed test being applied randomly and seemingly arbitrarily on students and that operational implementation of asTTle would result in teachers validating each test for their own students prior to administration.

Concluding Comments

The asTTle test items and materials in all six subjects were well received by both teachers and students in terms of content, interest, enjoyment, and engagement, and for the instructions supplied to teachers. Especially in mathematics, teachers and students were less positive about student ability to succeed at the tests. This concern ought to be removed with actual use of the asTTle software which provides for teacher customization of test difficulty. 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 evaluative feedback collected by the asTTle development team indicated that users, students, parents, and policy-makers can have confidence that the asTTle assessment materials engage and motivate students to show their true performance (provided teachers have administered tests at an appropriate difficulty level) and that the materials have validity and integrity with the curriculum.

Accuracy and Added Value—Understanding asTTle Reports

One of the more significant innovations embedded in the asTTle assessment system is the graphical reporting of performance information to teachers. Several studies were conducted including interviews (both face to face and online), iterative, interactive focus groups, review of the research literature, and testing of user accuracy of understanding. Evaluation studies were conducted during the initial design phase, during the pilot release of asTTle in the primary school sector, and in the pilot release of asTTle in the secondary school system. Studies were also conducted during the preliminary design and prototyping of the electronic, on-screen, internet assessment system currently being developed by the University of Auckland.

Initial Design Phase

An initial study (Meagher-Lundberg, 2000) found that teachers wanted to be able to compare student performance on a range of school and student demographic variables for a multitude of planning and reporting purposes. The desire for teachers to be able to make comparisons to similar schools resulted in a study of how best to aggregate school information such that "league table" comparisons could not be possible (Hattie, 2002). That study produced a set of clusters by which school variables were aggregated to produce a comparison variable entitled "Schools Like Mine."

A review of the literature (Brown, 2001) identified a number of problems around communicating assessment information to teachers and administrators such that they could accurately understand the data being displayed. Based on a commitment to developing a high-quality user-interface (Spolsky, 2001), two focus group studies were conducted in which various mock-ups of reports were tested for preference, clarity, and accuracy of understanding (Meagher-Lundberg, 2001a, 2001b). These studies lead to the refinement of the interface such that teachers could concentrate on the content being communicated rather than on the technologies or environments by which the information was displayed. They also indicated a strong interest in the type of information being displayed and suggested that teachers were likely to understand the information.

As a result of these design processes six major reports were designed for the asTTle software. Each report was designed to answer a different major educational question. The Console Report and its Comparison Groups selection system permitted interpretation of performance compared to similar students in the asTTle

norming sample. The Individual Learning Pathways Report permitted diagnostic description of student strengths and weaknesses at the curriculum objective level such that both teacher and student could identify what the next teaching or learning step was. The Group Learning Pathways Report aggregated the diagnostic information such that priorities for teaching and learning could be determined across a group, cohort, or whole school population. The Curriculum Levels Report aggregated student performance by curriculum level sub-groupings determined by a reputable standard setting procedure (initially examinee- and item-centered methods were used but phased out in favor of the bookmark method) and which permitted naming of students in each sub-level group so that appropriate educational materials and activities could be assigned to students at similar levels of progress. The Tabular Report provided a classic workbook-style listing of scores for high level curriculum functions and processes suitable for export to a student management system or other data analytic tools. The What Next Report provided a linkage from current performance levels to a web site that provided an indexed catalogue of teaching and classroom resources for each curriculum sub-level and each curriculum category used in the asTTle test creation system. Together these reports met many of the interpretive requirements identified in the previous evaluative studies.

asTTle V1 Primary School Pilot Phase

Upon release of asTTle Version 1 to a pilot sample of 110 New Zealand primary schools in mid-2002, a systematic evaluation of Year 5 to 7 teachers' understanding of the asTTle reports was conducted (Ward, Hattie, & Brown, 2003). To that end, a comprehension test, partially inspired by Hambleton and Slater (1997) and Linn and Dunbar (1992), of asTTle reports was created and

administered as part of a survey questionnaire delivered to participants. A series of three questionnaires were presented in a matrix sampling pattern (each teacher received two of the three questionnaires) that asked about the interpretation of the asTTle reports. For example, a Console Report (based on fictitious data) was presented and teachers asked "In which learning area did this class of students get the highest score?" or an Individual Learning Pathways Report was presented and teachers asked: What is the best way to understand the difference between Julie's asTTle Literacy Scale Score and the NZ reference group Year 5 mean?" Altogether there were 35 questions answered by 193 participants. The estimate of reliability of these items was .93, indicating that we can meaningfully interpret these items and the total score.

The overall mean for the Console Reports (.67 on a 0 = incorrect to 1 = correct scale) and What Next (.78) indicated a higher level of correct interpretations, but the means were lower for Individual Learning Pathways (ILP) (.51) and for Curriculum Levels (.57). Five of the items for ILPs were the lowest, particularly relating to writing, and it was clear that more professional development for the interpretation of the ILPs for writing was needed. A major part of the misinterpretation was that too many teachers were interpreting the concepts and items in each of the cells (Gaps, Strengths, and Achieved) relating to the NZ or class norms and not relative to the student's individual average ability. The concept that they should have understood is that relative to this particular student's ability, here are the concepts and items that he/she should have got correct (i.e., items relatively lower than this student's average proficiency in writing or reading) but did not (gaps), etc. Teachers, instead, were incorrectly interpreting the concepts as

relative to the overall class or NZ norm group. Modifications were subsequently made to the reports to make these issues more clear.

There were no statistically significant differences in the interpretation score between teachers and principals; $F(1, 172) = .02$, $p = .892$), nor between those who did and did not attend professional development; $F(1, 163) = 2.41$, $p = .122$). The major correlates (r ranged between .17 and .36) of interpreting the asTTle reports correctly were eight positive attitudes towards asTTle and its reports, including ease of use. Those teachers with a conception of assessment related to "assessment is powerful for improving teaching" had higher interpretation scores ($r = .34$), whereas those who had a conception of assessment as related to school accountability had the lowest interpretation scores ($r = -.21$). As argued by Brown (2004b) professional development needs to attend to the conceptions of assessment held by teachers before introducing asTTle, because those who see assessment as being about school accountability rather than improvement of teaching are less likely to accurately interpret, or attend to the information in educational assessment reports. Furthermore, those who most correctly interpreted asTTle made more positive (.30) and less negative (-.16) comments. It seems that those with the most negative comments made them because they had the least understanding of the reports. This information was used to improve the quality and quantity of professional development resources supplied to asTTle users by the Ministry of Education. It was also used to indicate what should be in the professional development—clearly teachers needed assistance in accurately understanding and thus using asTTle correctly.

Evaluation of Changes in the asTTle V3 Secondary School Pilot Phase

In preparing for the pilot release of asTTle Version 3 for use in secondary schools, a number of focus groups were conducted (Irving & Higginson, 2003). One of the confusions reported through the primary school release of asTTle was about the use of the ellipse on the Console Report to indicate the group mean plus or minus one standard error of measurement. Teachers reported understanding that the ellipse represented the performance of their whole group—to avoid this confusion focus group activities were conducted to determine whether the ellipse or a modified box-and-whisker group would deliver more accurate interpretations. Accuracy of understanding was similar for both formats with practice and group discussion both contributing to more accurate interpretations regardless of format. However, all but one participant indicated a preference for the box-and-whisker graph for the richness of information to assist both planning and reporting. The groups also provided useful feedback to the developers as to color schemes and layout options for the box-and-whisker reports that were subsequently implemented in the asTTle V3 report system.

During the pilot release of asTTle V3 in 55 secondary schools, multiple methods were used to elicit from participants their opinions and attitudes towards the asTTle reporting system (Hattie, Brown, Irving. et al., 2004). Feedback was obtained through an online discussion group, a mail-out questionnaire survey, and focus groups. Teachers were generally positive about the asTTle reports indicating that the amount of detail on reports was good and that the reports were very relevant to their needs. A third of teachers reported that the ability to report

performance to parents, colleagues, or students was of significant benefit. As was intended from the asTTle design, about half the teachers reported significant help from the formative and diagnostic reporting functions at both aggregated and disaggregated levels of reporting. In addition, a third of participants found benefit from the supposed-summative interpretations of aggregated data by seeing students' curriculum level performance using the Tabular Report, the Curriculum Levels Report and the Console Report. Teachers reacted positively to the reporting changes on the Console Report introduced in asTTle V3. The use of the box-and-whisker plots was seen as a good step forward in displaying not just the centre but also the distribution of scores within a group. The display of the norm score as a colored field within the gauges instead of as a number below the gauge was also seen as an enhancement. Teachers also appreciated being able to gauge student performance against norms for Years 4 through 12, instead of the restriction to Years 5 to 7 in previous versions.

In addition to the deliberately elicited information from these studies, the developers monitored requests for clarification and assistance that came in directly to the developers or indirectly through the Ministry of Education ICT Help Desk or through Ministry funded professional development providers. A significant number of queries concerning the correct or accurate interpretation of the reports were present. Comprehension problems that were identified included: how asTTle scores are calculated; what the scores are and what they mean; how curriculum levels are determined and what they mean; what the SOLO taxonomy is and what the depth of thinking categories mean; how the asTTle norms were derived and communicated; and what the 'Schools Like Mine' clusters are and how they were derived. It was especially noted that teachers found the self-referencing nature of

the Individual and Group Learning Pathways Reports relatively novel and these require extra attention in professional development and in asTTle documentation. Despite the availability of most of the information sought by asTTle V3 users through the online PDF manuals, users were finding it difficult to locate this information. It is suggested that adoption of a hyperlinked document solution might provide greater flexibility and speed in directly linking users to pages that explain facets of the reports without having to search long documents. Fundamentally, however, enhanced report interpretation needs to be foregrounded in professional development and asTTle documentation.

During the pilot period, feedback was collected as to secondary school teacher opinions about reporting requirements not currently met in asTTle V3. Two new types of reporting were identified; that is longitudinal reporting (i.e., seeing how individuals or cohorts progress across time), and comparison of performance with dissimilar or unlike students. Additionally, cross-sectional analyses of multiple reports of the same subject were requested. These requests have been evaluated for feasibility and their relative priority for development will be negotiated with funding agencies.

Concluding Comments

As a consequence of these studies, a series of online Macromedia Breeze presentations have been developed that explains the various asTTle reports, and a new chapter has been added to the asTTle Manual to ensure the correct interpretation of asTTle reports. The evaluations described here suggest that the scale of the problem related to correct understanding of the asTTle reports is large. Nevertheless, the evaluations also point to positive aspects and future directions.

The teachers are receiving and are appreciative of the rich information available through asTTle that assists with the planning and reporting of instruction with students in both primary and secondary schools. Professional development and commitment to assessment for improvement both contribute to more accurate understanding of the reports. Measuring teacher competence to use asTTle through assessing their comprehension of the asTTle reports is a relatively straightforward means of determining that competence. User based design and development procedures are essential in producing creative and communicative documents. Thus, these evaluations provide confidence that the reports are appropriately designed and are capable of adding value to teachers' educational practice, but raise significant concern over the ability of teachers to implement them properly—this is a matter of urgency both in terms of professional development and ICT development.

Training—Professional Development

This section provides a description of the professional development (PD) resources and services made available to users, a brief narrative history of the evolution of the PD, and findings from the various evaluations pertinent to the issue of PD. Provision of in-service PD to and within schools is becoming increasingly difficult—the cost of teacher substitutes, the absence of staff from normal duties, timetabling problems, and so on, make traditional classroom style PD difficult (North et al., 2000). In terms of educational technology, PD that does not make use of a school's own data has been shown to be relatively ineffective (North et al., 2000). Indeed, it has been recommended that educational technology PD move quickly to helping teachers make decisions in light of their classroom practice and responsibilities (Holland, 2001). The evaluations described here echo the transition

from classroom-style PD that focuses on skills to mechanisms that permit teachers to learn how to make educational use of asTTle within the context of their own institutions.

PD Resources

The Ministry of Education provided support to teachers for the use of asTTle by contracting for educational professional development and information and communication technology (ICT) support. Specifically, these included the asTTle website, contracted PD providers, and a technology helpdesk.

asTTle Website. A major mechanism for providing support to asTTle users for both technology and educational issues was the asTTle website (www.asttle.org.nz). This site was initially designed by the asTTle development team and deployed on the Ministry's education portal—Te Kete Ipurangi (TKI), a service to the Ministry provided by an independent commercial organization. The resources for maintenance, updating, and improving the asTTle website were provided by a separate direct contract between TKI and the Ministry. All updates to the asTTle website, initiated by the asTTle development team, went through the TKI quality assurance and budgetary mechanisms agreed between the Ministry and TKI, independent of the asTTle development team.

Professional Development. PD for teachers' use of the asTTle software was contracted by the Ministry of Education to a body of independent professional development contractors with experience in assisting schools with the use of assessment for the improvement of learning. This group of contractors, known as Assess to Learn (AtoL), reported directly to the Ministry's curriculum division and supplied educational support for asTTle and a wide range of other assessment tools

and resources, including the National Exemplars and the Assessment Resource Banks. During the deployment of asTTle Versions 1 to 3, AtoL consisted of nine independent organizations situated in New Zealand's six main centers (Auckland—3 groups, Hamilton, Palmerston North, Wellington—2 groups, Christchurch, and Dunedin). From the release of asTTle Version 4 in 2005, a group in Auckland and a group in Wellington were no longer contracted to perform this service, resulting in seven groups in the six centers. Schools contracted directly with the AtoL providers in their region to obtain training or support for the use of asTTle. Many other PD experiences were also lined up to support the messages re formative assessment: such as in literacy workshops, conferences of subject/teacher associations, and messages in official government gazettes. Indeed, for a while many other initiatives were aligned with the core messages from asTTle.

Technology Help Desk. The Ministry of Education's ICT Help Desk provided to schools, through a toll-free phone number and email access, a national, front-line technology support mechanism. The Help Desk was the first port of call for all school-based technology problems, including networking, operation of desk-top software, issues with operating systems, student management systems, and asTTle. For asTTle educational issues the Help Desk referred users to the AtoL agencies and for asTTle technology issues that were more complex, the Help Desk escalated the issue to the asTTle development team. These escalated issues became content for frequently asked questions (FAQs) documents posted on the asTTle website.

Thus, all training and communication resources to support the use and understanding of the asTTle software were fundamentally outside the control of the asTTle development team, being contracted directly to the Ministry of Education.

Furthermore, there were few mechanisms in place to ensure standards of service delivery around the country or to ensure sufficient supply of service to meet demands during the initial deployment and diffusion phases of the software.

History of asTTle PD

Some 12 months prior to the deployment of asTTle V1 in mid-2002, the Minister of Education announced the launch of the asTTle software and system. At that time, the asTTle website was deployed. The look and content of the website were designed by the asTTle development team and significant editing and quality assurance processes were carried out by TKI at the time. The website at that time provided descriptions of how asTTle would work and its intended purposes and rationale. The site was relatively non-interactive, in that it had no facility to provide updates or news or generate feedback to the developers.

At the time asTTle V1 was prepared for release in mid-2002 to a pilot group of 110 primary schools, the asTTle development team provided a one-day 'train the trainers' experience for all AtoL facilitators. The presentation materials and training data sets were provided to the facilitators at that time. The same training data set was posted on the asTTle website for use by schools, teachers and facilitators, so that inspection of the reports could take place without having to install school, class, and student information in the software. Greater interactivity on the asTTle website was added by making available the What Next materials for Levels 2 to 4 of reading and writing. Additionally, the asTTle technical reports were made available on the website for download; it was expected that these reports would provide useful information to users. An evaluation of user experiences and effects of the PD was conducted as part of the pilot evaluation

(Ward et al., 2003) and details are reported below. At the same time, the asTTle development team provided a series of briefings to the ICT Help Desk on technical issues surrounding the functionality of asTTle.

Two-thirds of the 110 schools volunteering to pilot asTTle V1 received at least one day of PD conducted by the Assessment to Learn (AToL) facilitators. The schools were well spread across New Zealand, and represented all types of schools from small to large, rural to urban, low to high socio-economic status, and primary and intermediate. Most participants at the PD days provided responses about their experiences or were requested to email their reactions to the asTTle team. The AToL facilitators themselves provided further feedback. Questionnaires were returned by 176 teachers, constituting a response rate of around 60%of all schools.

With each asTTle version release (V2 in 2003, V3 in 2004, V4 in 2005) briefings by the development team to the AtoL, a meeting of formative assessment experts, and ICT Help Desk providers on new features were given. At the release of V3, a PowerPoint slide show created by an AtoL group and validated by the development team about accurate interpretation of the asTTle reports was posted on the asTTle website and made available for download.

With the release of asTTle V4, which contained significant ICT changes, the AtoL groups were again given a single day 'train the trainers' experience, except this was delivered in the six main centers by two key personnel from the asTTle development team. The content for the training day was based partly on the need to explain and advise changes in the asTTle content and reporting systems (e.g., addition of Level 5 and 6 reading and writing materials, additional multi-test reporting) and on requests for focused assistance customized to the preferences of

each region. In addition to the direct instruction, the AtoL facilitators were provided with copies of the presentations used for adaptation by the facilitators in their own training programs. There were also technical days where one key person from each region came together to learn about the more educative uses of asTTle.

Another innovation, introduced with asTTle V4, in supporting the PD of teachers was the development and deployment of Macromedia Breeze presentations accessed through the asTTle website. These presentations provided interactive, self-access instruction on both technological and educational issues related to the use of asTTle. The intention was that, thanks to the presence of controllable voice-over and notes, schools could use these materials for their own internal staff development without having to wait for one of the AtoL facilitators to be available. Also, the technical presentations gave school teachers and administrators the power to support and direct their ICT technicians without having to be technically knowledgeable themselves.

Evaluation of asTTle V1 Primary School Pilot PD

During the deployment of asTTle V1, a series of telephone interviews (random selection of schools each week for 16 weeks) were conducted. Additional information was obtained from a multi-form survey questionnaire that examined teacher knowledge of the asTTle reports, attitudes towards ICT, assessment, PD, and asTTle, and experiences related to asTTle and asTTle PD. The three cross-linked questionnaire forms, in addition to open-ended questions, allowed participants to indicate direction and strength of opinion using a six-point positively packed response format (Brown, 2004a). Positive-packing means that there are more degrees of positive response than negative response and is useful

when it is expected that participants are positively inclined towards a phenomenon. Where feasible, factor-analytic procedures, using maximum likelihood estimation with oblimin rotation, were implemented. A structural equation path model of the questionnaire data was developed in light of Fishbein and Ajzen's (1975) Theory of Reasoned Action, and Ajzen and Madden's (1986) Theory of Planned Behavior. Together these theories identify behavior as influenced intent to adopt and use mitigated by attitudes, subjective norms, and perceived behavioral control.

All participants, regardless of whether they had received PD or not, were asked to state which of a number of given methods of PD had helped or would help them use asTTle. The most frequently used type of assistance was being self-taught through trial and error (82%) and 76% of these found this method the most helpful. When added to the totals for use of the asTTle Manual and the asTTle help function, then participants appear to have found teaching themselves asTTle to be the most beneficial method of learning how to use it. Two-thirds of participants stated that they had obtained informal assistance from colleagues with 64% finding this helpful. Those who were not offered professional development claimed to prefer to learn at their own pace, teach themselves, make use of internal rather than external professional development provision, and reading manuals. Clearly those who did not attend professional development found ways to support their own self-learning.

Additional insight as to preferred methods of PD was found when preferences for learning how to use any ICT application were investigated. Two factors were found: Short Practice-Oriented Training (Cronbach's alpha = .60), and Self-Controlled, Peer Supported, Informal Learning (Cronbach's alpha = .79). The participants showed a preference for practice-oriented training sessions and

disliked lectures or reading manuals. They also mostly agreed that several short sessions of training with time in between to practice was optimal, moderately agreed that professional development was more beneficial if there had been time to try things out beforehand, and moderately agreed that it was better to have someone from outside the school run the training sessions. Thus, the preference was for externally run, hands on, short sessions, with some pre-preparation exercise. The impact of these stylistic preferences was seen only in self-controlled learning ($\Lambda = .761$, Mult. $F = 3.77$, $df = 5.60$, $p = .005$); those who did not attend professional development favored self-controlled learning (all effect-sizes > .50).

About 70% of participants who received any PD considered it mostly to very adequate (using a six-point, positively packed scale). The most adequate aspect of professional development was in the creation of the tests while the least adequate was in interpreting the test marks. There were no statistically significant differences between principals and teachers when asked about the adequacy of the professional development received ($\Lambda = .920$, Mult. $F = 1.19$, $df = 7,96$, $p = .315$). Those participants who had not received any professional development indicated that they believed PD in seven areas to lie between moderately and mostly useful ($M = 4.6$, $SD = 1.3$).

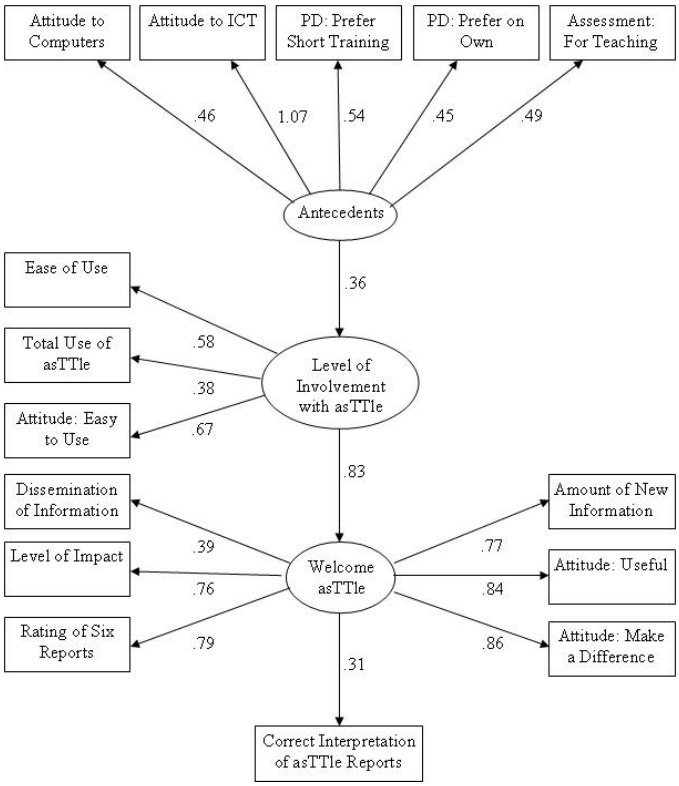
Those who received PD reported that they took longer to familiarize themselves with asTTle, to enter student data for their class, and to mark one writing script, than those who did not. Likewise, those who attended PD indicated that they obtained more information from the six reports than those who did not attend PD. Indeed, those who attended PD saw more value in four of the reports (i.e., the Individual Learning Pathways, Curriculum Levels, Tabular, and Group Learning Pathways Reports) than those who did not receive PD. Attending PD increased the

perceived value of most of the reports. Those who attended PD gave asTTle higher agreement for its Usefulness, particularly when using asTTle for talking to students about their own work, reporting to parents, and planning ahead for the whole class. Attending PD increased familiarity with asTTle, enhanced the spread of information to other persons (students, senior management, board, and parents), gave more confidence in creating reading tests and entering scores, and increased the use of Console, Group Learning Pathways, Tabular Reports, and using the reports for future planning.

A structural equation model relating teacher attitudes to their degree of involvement with asTTle during the pilot and subsequently to the level of accuracy they exhibited in understanding the asTTle reports was reported (Figure 1). The major messages from this analysis were that PD needed to be oriented mostly towards encouraging a positive attitude towards ICT as being valuable in teaching and learning. By positively influencing the teachers' attitudes towards ICT there was a higher likelihood of being actively involved with asTTle. Higher levels of involvement with the asTTle pilot predicted a more positive attitude towards welcoming and making use of asTTle; that is, being positive about the level of impact asTTle would make, positively rating the quality of the asTTle reports, and a positive belief that the results from asTTle would make a difference. This attitude of asTTle being useful to subsequent teaching predicted correctness of interpretation of the asTTle reports. In other words, teachers who welcomed asTTle for its usefulness to their teaching practice more accurately understood the reports.

Figure 1.

Structural Model of Professional Development Impact on Correct Interpretation of asTTle Reports



Evaluation of asTTle V3 Secondary School Pilot PD

During the first five months of asTTle V3's availability, a multi-method approach to obtaining secondary school users' experiences and opinions was used in the 55 pilot schools (Hattie et al, 2004). Data were summarized from self-initiated comments to the asTTle development team and from responses to focus groups, a survey questionnaire, an online discussion group, telephone interviews, and field visits.

Teachers indicated that there was a significant need for more PD, especially at school level, along with the more multi-school general courses that had been offered. Further, the teachers urged that the PD go beyond the operating-the-software 'driving licence' phase and focus on teachers using the data to make inferences about their own students. Additionally, the secondary teachers indicated that the PD offered did not adequately focus on secondary school departmental contexts and policy constraints. There were still beliefs that asTTle was a summative and accountability mechanism, and thus further PD was offered at school levels to show how asTTle reports could be used for both purposes. A noticeable difficulty with PD courses was the lack of real school-based data on which to base the learning. This meant that the interpretations teachers made in training were not fully informed by a complete knowledge of the students, and the curriculum content being tested. Thus, further development of contextualized professional development resources and processes still needed to be implemented so that teachers could be confident that they understood asTTle reports correctly. An identified need from this evaluation was a change to the training of trainers process and development of appropriate materials.

Teachers indicated that the asTTle V3 Manual and the other supporting help functions did not fully provide all the support that they might need. The asTTle Manual contains a large volume of relevant information and is structured in a linear fashion; it is not supported by navigation tools like indices or hyperlinks nor does it have any moving images or audio. Also, the Manual does not account for every possible variation of circumstance and thus users were required to apply the general concepts and principles to their own situations. Some users found the ICT delivered documentation unsatisfactory and they requested the inclusion of a print

booklet beyond the current CD slick to assist in these matters. A range of options for putting documentary power in the hands of users were identified by this evaluation. The Ministry, in conjunction with the asTTle developers, implemented the following revisions in time for the release of asTTle V4 in early 2005:

- a restructuring of the help documentation to improve the usefulness of the Manual with enhanced table of contents and hyper-linking,
- placing the Manual, technical reports, and frequently asked questions documents on the asTTle website for download, and
- creating online self-access tutorials that had extra notes and voice-over explanations and which also permitted user control and flow through the information.

The point of these improvements to communication mechanisms was to permit self-access learning and training experiences in a context where the provision of relevant face-to-face PD could not be guaranteed.

Concluding Comments

The two systematic, internal evaluations of asTTle deployment in primary (V1) and secondary (V3) school contexts have identified a number of principles of effective PD. PD is powerful when it:

- attends to teachers' perceptions of the usefulness of assessment or educational technology innovations with the goal of inculcating positive attitudes towards ICT,

- devotes attention to the usefulness and correct interpretation of the asTTle reports in order to make a positive difference to the quality of classroom teaching and learning,

- responsively supports self-access or self-learning with brief, face-to-face, hands-on practice training sessions, and

- is contextualized in the user's school environment as a kind of action-research examination of the school's own data, students, and teaching.

- development of a "slick" that was part of the CD that outlined the major purposes and uses of the tool. Much effort, focus groups, and design effort went into this slick as for many teachers it was there first major encounter with the application.

These recommendations have since been implemented in the asTTle PD system as they have informed the design of continuous improvements. The Ministry has since funded a whole new level of workshops – one in each region of the country specicalising in asTTle – these PD providers go into schools and sit with teachers and work through their issues. The evaluations indicated that there was noticeable benefit from participating in the PD in terms of accuracy of interpretation and thus, it is argued that the PD adds value and accuracy in teachers' practice.

Utility—asTTle Software

It is important that any educational technology or software be easy to use and produce a beneficial user experience. asTTle is an educational resource that happens to use technology, not a technology resource looking for a use in

education. Assessing real learning is a real educational application that desperately needs tools to relieve teacher workload and to improve teacher effectiveness. The touchstone of the asTTle project has not been "are we using the newest and best ICT?", but rather "when teachers use asTTle do they focus on the technology or the education?" Indeed, the guiding principle was that if teachers talked about the technology, we had failed. Fundamentally, the technology has to become invisible to users (i.e., have high degrees of utility) so that attention is spent on achieving the purpose of the software—in this case, improving the quality of school-based assessment such that improved teaching and learning takes place. Special attention to utility requirements has been a key development and evaluation concern for asTTle. As discussed under Section 2, evaluations were conducted of the asTTle reporting mechanisms during the development process. This section reports evaluations of the asTTle software primarily during its pilot deployments in primary and secondary schools.

The asTTle ICT Development

The asTTle system requires ICT in order to achieve several fundamental features: linear programming is used in customized test creation, item response theory is used to generate sophisticated score calculations, the internet is used to access an indexed catalogue of teaching and learning resources, and database systems are required to permit data sharing and interoperability. Improving the quality of teaching and learning requires accurate score calculation, rapid test creation, and widespread dissemination of information. None of these could be achieved without powerful desktop microcomputing. Only through ICT can accurate and rapid estimates of a student's strengths and needs be attained, freeing the teacher to concentrate on the important decisions and actions he or she will take based on that

information. The transformation of scores into meaningful displays requires computer technology that captures expert processes—again freeing teachers from the drudgery of detailed test or item analysis. Thus, ICT was necessary but not the real purpose of asTTle—this is a case of an educational innovation that happens to use ICT, rather than an ICT innovation in education.

The design, selection, and deployment of ICT was done in a manner consistent with the infrastructure status and development plans of the country (Brown & Hattie, 2005). asTTle V1 had to create 40-minute, paper-based tests and report on student performance using the kinds of computers already in schools. Hence, it functioned on stand-alone computers only such as Mac Classic (OS 8.6 and 9.2) and Windows 95. Based on the positive response of teachers to that resource and evaluations which documented various calls for improvements, asTTle V4 has extended ICT functionality over a wider range of ICT systems and technologies.

asTTle V4 now gives users the option of a shared database system (Multi-user asTTle) that operates on various thick-client server types (i.e., Mac OSX Panther and Tiger; Windows 2000, 2003, and NT; Linux Redhat 9 and Enterprise; and Novell Netware 6.5). Thus, users now can more readily share tests, students, and reports across a network. In addition, asTTle V4 client application operates on laptops and desktops using Mac OSX Panther and Tiger, and Windows 2000 and 2003, and, furthermore, in stand-alone mode still functions on Windows 98. This gives users choice and control over technology systems, rather than being constrained to one system. Furthermore, communication with student or school management or information systems (SMS) is possible; indeed, accreditation as a Ministry of Education approved SMS provider requires interoperability with

asTTle V4. This has meant that the deployment has attained appropriate levels of utility and interoperability.

Evaluation of asTTle V1 Primary School Pilot ICT

Participants ($n = 165$) were asked to indicate their attitudes towards the use of computers. Overall, they moderately agreed with using ICT and there was much support for the notion that computers were useful in teaching and learning. They tended to be somewhat negative towards their ability to cope with the more technical aspects of computers. Participants indicated that all the ICT-dependent tasks (i.e., test creation, data entry, and report generation) involved in using asTTle software were, on average, between moderately and very easy to use. Just over half (52%) found asTTle very easy to use, just under half (44%) found it slightly or moderately easy to use, and only 5% found asTTle hard to use. Three significant technology-related issues were identified—the amount of time needed to do assessment, data entry and sharing, and paper cost and consumption.

The use of asTTle required significant time, with most time spent on the marking of scripts. On average, participants spent between 6 and 30 minutes on individual asTTle tasks. For a class of 30 students, the time taken to create a reading test, mark the scripts, enter the scores, create and interpret the reports would be about five hours. Nevertheless, compared to creating one's own 40-minute test and administering, scoring, and creating reports, it could be argued that this is a significant saving of time.

Data entry was also a major concern with 35 comments related to the time taken to enter data. Score entry required two key strokes per question (i.e., one to enter the score and a second to move to the next cell) in order to ensure accuracy of score

entry. Furthermore, data entry for multiple-choice questions could only be done through alpha keys (to distinguish these from open-ended question). Additionally, many comments focused on difficulties around data sharing and transfer. For example, teachers wanted to be able to easily transfer data between computers either at school or between home and school and to share data on a network. For many teachers, it was time-consuming to create classes and student records since there was no easy linkage to school student management systems. Despite the fact that data sharing was possible in asTTle, the feedback lead to changes in V3 and V4 such that data entry became single key, the asTTle data were networked, and greater interoperability with school management systems were introduced.

Participants also identified the amount of paper required and the cost of photocopying required by asTTle tests to be a significant barrier. Nevertheless, this was part of the original technology brief and will be addressed with the design and deployment of on-screen test administration.

The first issue was treated by the developers as primarily a training and communication problem, the second was one which could be addressed technically, and the third had to be ignored until the infrastructure and will existed to enable onscreen or online testing. Nevertheless, the users liked the flexibility and control asTTle gave them over assessment and fundamentally found the software usable.

Evaluation of asTTle V3 Secondary School Pilot ICT

In evaluating asTTle V3 in secondary schools a large number of ICT problems and obstacles were identified. These issues revolved around interoperability, data sharing, paper sharing, and speed of operation.

Interoperability: A significant number of cases concerning broken communication between asTTle and school management systems were attributed to poor coding by the various SMS providers. This triggered the development and deployment of a protocol for seamless synchronization between asTTle and any SMS in asTTle V4. Limitations were also identified in SMS providers giving inaccurate advice to schools about interoperability between asTTle and SMS. Consequently, the Ministry of Education has made interoperability with asTTle a requirement for accreditation of any SMS for use in schools. Interoperability with the Apple MacIntosh operating systems caused significant problems. In order to be compatible with Apple directions, support for Mac Classic operating systems was dropped. Despite free school access to Mac OSX through a Ministry of Education agreement, several complaints were received. More importantly, the frequency of OS upgrades from the Apple company threatened user confidence in asTTle; for example, a change in OSX 10.2 after asTTle V2 was released and a change in OSX 10.4 after asTTle V4 was released broke asTTle's interoperability with Acrobat Reader in those environments. Another application upon which asTTle is dependent is Acrobat Reader. Versions 6 and 7 of that application have introduced minor faults in handling refresh of asTTle reports that did not exist in Version 5. Again, this fault is attributed to a deficiency in the asTTle software by many users. Despite patches or workarounds being developed and released there is a constant threat to the viability of the application due to third party programming. A fully web-enabled e-asTTle would ensure true platform and OS independence.

Data Sharing and Entry: Frustration with the slowness of data entry reported in the V1 evaluation was once again evident and the V4 version implemented one key entry mechanisms for each question. Despite having been told at the start of the

pilot study that asTTle V3 was not network compatible, participants criticized asTTle for deficiencies in data sharing, controlling, and storing. Schools wanted to be able to install and operate asTTle across a network so that back-up and security of data could be implemented. This was implemented and deployed in asTTle V4 multi-user, local-area-network, data sharing systems. A few cases of database corruption had been reported during asTTle V2 caused by users attempting to launch multiple instances of the asTTle software. asTTle V3 detected and prevented multiple instances of the application launching and supplied a utility application that corrected the corruption. Thus, this development overcame a problem caused by unintended and unexpected user behavior. To further reduce user impatience with the application launch process, asTTle V4 provides a splash screen to show that the application is starting. These actions demonstrated developer responsiveness and increased utility of the software. Some schools discovered that their student identity numbering systems were not unique as they tried to share asTTle data in a common repository for multi-school comparisons. Procedures for ensuring unique student identification systems common across all SMS providers have become one ICT development strand for the Ministry of Education as a result of this type of experience.

Paper: Constant comments were received about the amount and cost of paper involved in reproducing the paper-and-pencil based asTTle tests. This criticism came despite feedback from students that they enjoyed the appearance of the asTTle tests (see section earlier on Validity). Nevertheless, until users are provided with an option for a completely digital presentation of assessments, complaints about this aspect of asTTle will continue This issue is being addressed as part of

the next generation development of the asTTle software which will include onscreen testing.

Speed of Operation: Slow and lengthy installation cases were attributed to low-end specification machines or the action of anti-virus software monitoring the transfer of files. Lack of user installation privileges also contributed to installation problems. A partial response was to raise the standards for asTTle hardware and greater efficiency was found with newer anti-virus applications. Installation rights and privileges remain a significant obstacle that will be reduced with a fully online version of asTTle where all processing would take place on the server. Speed of operation was sometimes negatively affected, not just by client machines, but also by school networks that were so unstable that internal communication was halted. The asTTle V3 pilot identified that processing of data to create one type of report was taking an inordinate amount of time. Testing by the asTTle team verified the problem and revisions to the code were implemented so that the report generation time was reduced some 80 to 90% and deployed in asTTle V4. Thus, real user experience was used to define the nature of development priorities and maintain user confidence that asTTle was usable within school ICT environments.

Concluding Comments

These two evaluations demonstrated that the gradual development and improvement approach adopted by asTTle developers has resulted in a software application that is usable. The development team identified and responded to issues found in the field by users and were able to implement significant technological enhancements currently available to users. The ICT capabilities of asTTle V4 are well matched to the school-based ICT infrastructure and user requirements.

However, the requests for new technical features and options never stop. Users have clearly indicated a requirement for onscreen testing, easier data entry, greater interoperability, greater robustness in the face of third party changes, and simpler deployment. As the school-based infrastructure develops and as the political will to invest in these kinds of enhancements is exercised, it is expected that the enhancements and extensions will be delivered. In this incremental fashion, funders and users can be assured of on-time and on-budget delivery of features and functionality that do operate within ICT contexts in schools and which deliver utility to the process of educational assessment. The asTTle development team are currently beginning work, under contract to the Ministry of Education, on an electronic assessment system that would deliver many of these requirements.

Conclusion

The studies reported here may be taken as case studies of consequential evaluation (Sen, 2000). The developers took responsibility for the consequences of their design decisions in the politically charged environment of national testing, and sought out evidence that permitted maximization of end-user benefits, even when it was not clear what the optimal decision might be. The gradual implementation process meant that decisions could be and were made but then could be monitored for improved information as to optimal decision making. Attention was paid to consequences associated with how asTTle was developed, as well as to the culmination outcomes of having a state-funded, national testing program. The publication of the evaluation studies themselves and the transparency of the processes by which the package was developed ensured that the developers maximized end-user benefits and the end-users were left in no doubt as to how important consequences for them were for the developers. The very process of

paying attention to and being seen to pay attention to end-user consequences meant the product and process were legitimated. Furthermore, these studies have shown that attention to consequences as the basis for evaluating a development ensured that the asTTle assessment tool was actively constructed to minimize abuse of end-user rights by enforced implementation of an inappropriately designed assessment tool. At the same time, the tool was developed in such a way that teacher obligations to the state and parents concerning appropriate monitoring of and response to student learning could be fulfilled. Thus, consequential-focused evaluation as practiced here meant that an educationally sound assessment tool has been designed for and adopted by teachers such that teaching and learning can be improved while maintaining national standards and improving the flow of accountability information. Attention to consequences in evaluating any innovation is a powerful framework for determining those qualities that maximize end-user benefits.

It is worth noting that the processes used to evaluate and develop asTTle by the development team are consistent with the rules of thumb identified by Baker (2005) for designing and using technology in support of accountability. Specifically, asTTle has stuck to its core business—creating and reporting psychometrically robust assessment of learning. It has not become a student management system, nor a digital learning object warehouse. asTTle began early with focusing on educational outcomes and understanding how best to communicate those in ways that allow teachers to improve the quality of teaching and learning. Technology has been made subservient to educational value and teachers have participated and consulted on the basis of their acknowledged

expertise. Development has been deliberately incremental and has expanded as both infrastructure, finances, and desire has required and permitted.

These evaluations have demonstrated that the asTTle development team have investigated appropriate categories of issues and responded with the explicit intention of utilizing or implementing those innovations that research showed were of maximal benefit to the end-user teacher community. As a consequence of this in-house evaluative work, users can have confidence that asTTle has demonstrated validity with the curriculum and classroom practice, that the reporting systems add value to teachers' work and that accuracy of understanding is enhanced with professional development. The utility of the software has been constantly increased, but there are still unresolved issues that are currently being responded to. The gradual implementation mechanism has been successfully used to obtain user buy-in as well identification and response to identified needs. Users and funding agencies can have confidence that asTTle has been designed, developed, and implemented in a fashion consistent with maximizing utility and benefit. The software has been developed in such a way that the validity, added value, accuracy, training and utility standards for educational technology have been met, at least in part. The data reported here demonstrate again that formative evaluations can add significant value to an educational technology innovation.

References

Ajzen, I., & Madden, T. J. (1986). Predictions of goal-direct behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology*, *22*, 453-474.

Baker, E. L. (2005, July). *Improving accountability models by using technology-enabled knowledge systems* (TEKS) (CSE Report No. 656). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Baker, E. L., & Herman, J. L. (2003). A distributed evaluation model. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: Effective research designs for improving learning* (pp. 95-119). New York, NY: Teachers College Press.

Bottino R. M., Forcheri P. & Molfino M. T. (1998). Technology transfer in schools: From research to innovation. *British Journal of Educational Technology*, *29*(2), 163-172.

Bourges-Waldegg, P., Moreno L., Rojano T. (2000). *The role of usability on the implementation and evaluation of educational technology*. Proceedings of the 33rd Hawaii International Conference on System Sciences, pp. 1-7.

Brown, G. T. L. (2002). *Teachers' conceptions of assessment*. Unpublished doctoral dissertation, The University of Auckland, Auckland, New Zealand.

Brown, G. T. L. (2004a). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports*, *94*, 1015-1024.

Brown, G. T. L. (2004b). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice*, *11*(3), 305-322.

Brown, G. T. L., & Hattie, J. A. C. (2005, September). *School-based assessment and assessment for learning: How can it be implemented in developed, developing and underdeveloped countries?* Paper presented at the APEC East Meets West: An International Colloquium on Educational Assessment, Kuala Lumpur, Malaysia.

Carroll, T. G. (2001). Do today's evaluations meet the needs of tomorrow's networked learning communities? In W. F. Heinecke & L. Blasi (Eds.), *Methods of evaluating educational technology* (pp. 3-15). Greenwich, CT: Information Age Publishing.

Compton, V., & Jones, A. (1998). Reflecting on teacher development in technology education: Implications for future programs. *International Journal of Technology and Design Education*, *8*, 151-166.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J, Research in Learning Technology*, *13*(1), 17-31.

Cuban, L. (2001). *Oversold and underused: Reforming schools through technology, 1980-2000*. Cambridge MA: Harvard University Press.

Fichman, R. G. (1992). *Information technology diffusion: A review of empirical research*. Cambridge, MA: Massachusetts Institute of Technology, Sloan School of Management.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430.) Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Hattie, J. A. (1999, August). *Influences on student learning*. Inaugural Lecture: Professor of Education, The University of Auckland, Auckland, New Zealand.

Hattie, J. A. (2002). *Schools like mine: Cluster analysis of New Zealand schools*. (asTTle Tech. Rep. No. 14). Auckland, New Zealand: The University of Auckland, Project asTTle.

Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning (asTTle). *International Journal of Learning*, *10*, 771-778.

Hedges, L. V., Konstantopoulos, S., & Thoreson, A. (2003). Studies of technology implementation and effects. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: Effective research designs for improving learning* (pp. 187-204). New York, NY: Teachers College Press.

Irving, S. E., & Higginson, R. M. (2003). *Improving asTTle for secondary school use: Teacher and student feedback* (asTTle Tech. Rep. No. 42). Auckland, New Zealand: The University of Auckland/Ministry of Education.

Lesgold, A. (2003). Detecting technology's effects in complex school environments. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: Effective research designs for improving learning* (pp. 38-74). New York, NY: Teachers College Press.

Lewis, A. (1999, June). *Comprehensive Systems for Educational Accounting and Improvement: R&D Results* (CSE Technical Report No. 504). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies, University of California, Los Angeles.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 177–194.

Meagher-Lundberg, P. (2000). *Report on comparison groups/variable for use in analyzing assessment results* (Technical Report No. 1). Auckland, New Zealand: The University of Auckland, Project asTTle.

Meagher-Lundberg, P. (2001a). *Report on output reporting design: Focus group 1* (Technical Report No. 9). Auckland, New Zealand: The University of Auckland, Project asTTle.

Meagher-Lundberg, P. (2001b). *Report output reporting design: Focus group 2* (Technical Report No. 10). Auckland, New Zealand: The University of Auckland, Project asTTle.

Means, B., Haertel, G. D., & Moses, L. (2003). Evaluating the effects of learning technologies. In G. D. Haertel & B. Means (Eds.), *Evaluating educational technology: Effective research designs for improving learning* (pp. 1-13). New York, NY: Teachers College Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 2-103). New York, NY: Macmillan.

North, R. F. J., Strain, D. M., & Abbott, L. (2000). Training teachers in computer-based management information systems. *Journal of Computer Assisted Learning*, *16*, 27-40.

Parks, S., Huot, D., Hamers, J., & H.-Lemonnier, F. (2003). Crossing boundaries: Multimedia technology and pedagogical innovation in a high school class. *Language Learning & Technology*, *7*(1), 28-45.

Patton, M. Q. (1978). *Utilization-focused evaluation*. Beverley Hills, CA: Sage.

Posavac, E. J., & Carey, R. G. (1997). *Program evaluation: Methods and case studies* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Provus, M. M. (1973). Evaluation of ongoing programs in the public school system. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

Raikes, N., & Harding, R. (2003). The horseless carriage stage: Replacing conventional measures. *Assessment in Education: Policy, Principles and Practice*, *10*(3), 267-277.

Riel, M. (2001). Evaluating educational technology: A call for collaborative learning, teaching, research and development. In W. F. Heinecke & L. Blasi (Eds.), *Methods of evaluating educational technology* (pp. 17-40). Greenwich, CT: Information Age Publishing.

Sen, A. (2000). Consequential evaluation and practical reason. *The Journal of Philosophy*, *47*(9), 477-502.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation & education: At quarter century* (pp. 19–64). Chicago, IL: National Society for the Study of Education.

Spolsky, J. (2001). *User interface design for programmers*. Berkeley, CA: APress LP.

Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation*. Boston, MA: Kluwer-Nijhoff.

Ward, L., Hattie, J. A., & Brown, G. T. L. (2003). *The evaluation of asTTle in schools: The power of professional development* (asTTle Tech. Rep. No. 35). Auckland, New Zealand: The University of Auckland/Ministry of Education.

Watson, D.M. (2001). Pedagogy before technology: Re-thinking the relationship between ICT and teaching. *Education and Information Technologies*, *6*(4), 251-266.

Note

Relevant asTTle Technical Reports not cited in this article but available to the interested reader at www.asttle.org.nz include the following:

Brown, G. T. L., Irving, S. E., Hattie, J. A., Sussex, K., & Cutforth, S. (2004, August). *Summary of teacher feedback from the secondary school calibration of asTTle reading and writing assessments for curriculum levels 4 to 6* (asTTle Tech. Rep. No. 49). Auckland, NZ: University of Auckland/Ministry of Education.

Hattie, J. A. C., Brown, G. T. L., Irving, S. E., Keegan, P. J., Sussex, K., Cutforth, S., et al. (2004). *Use of asTTle in secondary schools: Evaluation of the pilot release of asTTle V3* (asTTle Tech. Rep. No. 47). Auckland, New Zealand: The University of Auckland/Ministry of Education.

Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., et al. (2004). *Assessment tools for teaching and learning (asTTle) version 4, 2005: Manual*. Wellington, New Zealand: The University of Auckland/Ministry of Education/Learning Media.

Keegan, P. J., & Pipi, A. (2002). *Summary of the teacher feedback from the calibration of asTTle v2 pänui, pängarau and tuhituhi assessments* (asTTle Tech. Rep. No. 27). Auckland, New Zealand: The University of Auckland/Ministry of Education.

Keegan, P. J., & Pipi, A. (2003). *Summary of the teacher feedback from the calibration of the asTTle V3 pängarau assessments* (asTTle Tech. Rep. No.

44). Auckland, New Zealand: The University of Auckland/Ministry of Education.

Lavery, L., & Brown, G. T. L. (2002). *Overall summary of teacher feedback from the calibrations and trials of the asTTle reading, writing, and mathematics assessments* (asTTle Tech. Rep. No. 33). Auckland, New Zealand: The University of Auckland, Project asTTle.