asTTle—a national testing system for formative assessment: How the national testing policy ended up helping schools and teachers

Gavin T L Brown,
University of Auckland

Stake (2004) suggests that national educational standards, tests, and stipulated national curricula are ‘criterial’ approaches which derive criteria for educational quality ‘somewhere else’ – i.e. out of context. These are imported into an educational setting and used to measure or make judgements about performance, displacing local judgements. While this may be true in top-down imposed models, in this chapter I wish to suggest that it is possible to have one’s criterial cake and still eat one’s experiential, local, and contextualized cake. To do this, I will first describe the Assessment Tools for Teaching and Learning (asTTle) computer-assisted, standardised testing system. Then, I will address how the operationalization of asTTle led to a highly contextualised, criterial testing system that helps teachers and schools. Finally, I will address three controversies that arose in the research and development process which highlight tensions between policy makers, developers, and teachers. These tensions reveal contrasting experiential and criterial bases for evaluating school evaluation tools, policies, and processes.

Notwithstanding the potential negative impact of government-mandated compulsory assessments, the 2000 Labour government of New Zealand saw a need for high-quality national assessments. Specifically, they wanted assessments that could contribute to improving the quality of school assessment decisions, ensuring national standards, and improving the quality of reporting to parents (Gilmore, 1998; Hattie & Peddie, 2003; Ministry of Education 1990, 2002; New Zealand Government, 1999; Philips, 1999; Phillips, McNaughton, & MacDonald, 2002; Robinson, Timperley, Ward, Tuioto, Stevenson, & Mitchell, 2004). Thus, a request for proposals was announced to develop national, externally-referenced tests at years 5 and 7 in literacy and numeracy in English and Māori. A number of bids were prepared, including one by NZCER to which I contributed. The proposal from the University of Auckland, led by Professor John Hattie, was immediately selected as the preferred provider and over a six-month period contracts were negotiated, with first work being done in August 2000.

The asTTle test system

It is within this context of increased Ministry attention to meeting teacher and school resource needs (i.e., tools and professional development) in order to meet the goals of improved learning outcomes for all students and improved accountability for meeting national standards, that the development and deployment of the asTTle test system was situated.

Development History 2000-2005

The initial asTTle contract involved developing, by the end of 2002, a testing system in English and Māori for mathematics/pāngarau, reading/pānui, and writing/tuhituhi for use in Years 5-7 and related to Curriculum Levels 2-4. This meant that asTTle had to report against both norms and standards so that consistent interpretation of performance in the key domains of literacy and numeracy could be achieved. It should be noted that this contract was managed by the Curriculum and Assessment group within the Ministry of Education, rather than the Research Division or Information Technology group. The first testing of items for reading and writing was carried out in late 2000, with three more rounds of item development and trialling carried out by June 2002. The first working demonstration of the asTTle software was given at a series of National Assessment Regional Seminars in early 2002, and a fully functional application (asTTle V1) for reading and writing (English only) was
deployed in a pilot study of 100 primary schools in mid-2002. asTTle V2 with the contracted materials was delivered in December 2002 and deployed to schools in early 2003.

In late 2002, a 2-year extension of the asTTle research and development contract was agreed to incorporate materials for curriculum Levels 5-6 and extend the normative base to include Year 12. My understanding is that the secondary school teacher union (Post Primary Teachers Association—PPTA) was influential in motivating the Ministry to implement this extension, because they argued that high school teachers had to work with students arriving in Year 9 still functioning in curriculum Levels 2-4, despite the curriculum expectation that they would be working at Levels 5 and above. Two rounds of testing were carried out in 2003 and 2004 respectively and asTTle V3 (containing L2-6 mathematics and pāngarau materials) was released to about 50 pilot schools in early 2004. Version 4 (containing Level 2-6 content for literacy and numeracy in both languages with norms for Years 4-12 in both English and Māori medium schooling) was delivered to the Ministry in December 2004 and released to schools in early 2005. asTTle V4 had many technical advances which allowed it to function across school local area networks.

In response to many issues raised by schools, a proposal to develop a fully electronic testing system (i.e., computer administered testing, web-based interface, and possible computer adaptive testing) was developed at the behest of the Ministry of Education. Apparently, there were tensions between the then Minister of Education (Hon. Trevor Mallard) and Ministry officials as to whether e-asTTle would be funded, with the former insisting on it and the latter resisting it. Hence, in late 2004, an extension of the asTTle project was contracted to ensure smooth and reliable delivery of the technology in case technical issues arose—much in the form of an insurance policy. This also allowed time for the Ministry to obtain funding from cabinet for the development of e-asTTle and for preliminary work to be done as to the feasibility of developing such a mechanism. That contract was signed in late 2005.

**Description**

By the end of 2005, New Zealand schools had access to asTTle V4 on a voluntary basis. The system was an electronic test creation and reporting engine and test item bank covering Curriculum Levels 2 to 6 of reading, writing, and mathematics in both English and te reo Māori with norms for Years 4 to 12 based on the performance of over 92,000 students on over 4000 test items and 100 writing prompts (60 English, 40 Māori) (Hattie, Brown, & Keegan, 2003). The asTTle tools provided teachers and school leaders with the ability to analyze achievement of individual students or groups/subgroups of students, gain insights as to strengths and weaknesses, and be pointed to additional teaching and curriculum resources through an online catalogue known as ‘What Next’. The asTTle software allows teachers and/or school leaders to customise standardised 40-minute tests according to their priorities for test difficulty and content regardless of the year or age of students. The software 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 to fit the teacher request. However, teachers have the final say as to whether the test is valid for their purposes or not, through the ability to reject a draft test and change selections multiple times until an appropriate test is created. The system software calculates student achievement, strengths, and weaknesses using an item response theory formula so that regardless of the items on a test, student scores are on a common scale, which ensures interpretation against 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.

The asTTle system had certain key characteristics which were included to reduce the negative impacts of external centrally mandated testing while maximally approaching the legitimate information needs of students, teachers, parents, school leaders, and government officials. These are described more fully in Hattie and Brown (2008), but in summary they include six Cs. asTTle provides Curriculum-aligned measures that are Calibrated to national norms and curriculum standards; it generates enhanced graphical Communication of assessment information to teachers by having computer-assisted analysis of test results; it gives teachers Control over the content, timing, and validity of tests; it is not Compulsory to use the tool; and results are not Centrally reported or captured. What this means is that New Zealand has managed to avoid the worst features of compulsory, national assessment by developing a wide range of assessment tools and resources that assist teachers in improving the quality of teaching and learning. A key feature is that many of those resources are indexed against both national norms and curriculum levels and are used voluntarily by schools for their own self-managed delivery of national curriculum expectations. Note, unlike American calls for new kinds of tests aligned to current theories of learning and pedagogy (National Research Council, 2001), the New Zealand solution continues to use much more traditional-style tests. The real difference is that instead of making assessment compulsory and keeping the test power in the hands of central offices, assessment in New Zealand is voluntary, low-stakes, and in the control of the teacher and school leader. This mix means that information about what needs to be changed goes very quickly to the people who have the power and responsibility to improve teaching and learning outcomes.

Success Factors

A number of factors have been identified as contributing to the success of the asTTle system (Archer & Brown, 2013; Darr, 2003; Hattie & Brown, 2008, 2010; Hattie, Brown, Ward, Irving, & Keegan, 2006). These factors reflect a combination of criterial standards and experiential, contextualised principles. These included:

1. a robust commitment to the centrality of the curriculum statements and subsequent clarification of curriculum objectives and levels;
2. the closing of the curriculum-teaching-assessment triangle by linking student results to high-quality teaching resources indexed to curriculum levels;
3. a robust commitment to explicitly serving the goal of improving outcomes by involving the teacher’s professional judgement;
4. the large-scale involvement of teachers in the development and evaluation of asTTle materials and software;
5. the sensitivity of the development process and asTTle product to the concerns, beliefs, values, and attitudes of teachers and school leaders;
6. a constant focus on communicating to teachers and leaders information they needed to address both improvement and accountability requirements;
7. the policy of making usage voluntary and complementary to existing practices;
8. the use of sophisticated item response theory statistical analysis of item characteristics and student performances as the basis of score creation and reporting of results;
9. the flexibility of the asTTle software to operate on pre-existing school technology systems;
10. the gradual approach used to introduce the assessment methods and educational technology systems;
11. the computerisation of technical item by item analysis of student performance and rapid reporting of reports allowing teachers to concentrate on teaching curriculum to students rather than having to become expert test analysts;
12. the extensive support given to schools by the Ministry of Education through professional development services, a technology telephone help-desk, and online tutorials;
13. the ability to serve both improvement and evaluation goals by provision of multiple reports;
14. the transparency of the development processes and decisions through the dissemination of technical reports on the internet;
15. the avoidance of externally imposed consequences which encouraged more honest detection of poor or unsatisfactory performance;
16. the insistence that the standardised asTTle tests were not the sole or best judgement of student performance combined with the provision of multiple methods of evaluating progress and achievement; and
17. the independence of the research and development team from the Ministry of Education.

Schools that used the asTTle system soon reported significant changes in student learning outcomes. This was seen most robustly in the curriculum domain of writing where large gains in achievement were seen after a two-year teacher professional development project (McDowall, Cameron, Dingle, Gilmore, & MacGibbon, 2007; Parr, Timperley, Reddish, Jesson, & Adams, 2007). Teachers anecdotally report using asTTle effectively to improve student learning outcomes (Archer & Brown, 2013; Brown & Harris, 2009; Brown & Hattie, 2009).

An important success factor in bringing asTTle into the Māori medium school system was the strong and persistent refusal by the research and development team to resort to translation of items (Keegan, Brown, & Hattie, 2013). This commitment was a consequence of strong belief in the importance of a socio-culturally sensitive test system that was by Māori, in Māori, and for Māori.

Challenges to asTTle

It is worth considering a number of challenges that arose during the development of asTTle and how these were addressed. These challenges illustrate the highly political nature of assessment and perhaps reflect more on the politics of government bureaucracies than anything. Further, these challenges identify different experiential and criterial factors that were brought to bear on the development of the asTTle system.

Defining the curriculum

While the developers were aware of the ‘reading wars’ (i.e., debate between top-down and bottom-up models of learning to read), they were not aware that within the New Zealand curriculum there was a ‘math war’. This struggle focused primarily around the purported deficiencies of the conventional approach to teaching students algorithmic methods of arithmetic operations. To exemplify the vertical form approach, lay out the problem of add 27 to 16 vertically with numbers aligned according to place value (Figure 1A). Then apply the conventional algorithm of first adding the values in the ‘ones’ column, carrying the 10 value into the tens column and then add the values in the tens column. This algorithm can be effectively repeated for subtraction and multiplication and with a transformation applied to division also.

The claim made within the mathematics education and curriculum division was that
mechanical application of these algorithms did not ensure students had a deep understanding of number properties and contributed to many mechanical errors. Instead, the Numeracy Project advocated that students should first become competent with a variety of mental strategies for dealing with such number problems. For example, students would be taught to make numbers easier to handle by adding or subtracting to ensure one number has zero in the ones column. Hence, since 27 is close to 30, 3 could be taken from 16 to make 30 and thus the problem becomes a much simpler 30+13 with no complex borrowing (Figure 1B). Note that Figure 1B displays the problem in horizontal form and students would be encouraged to do this mentally rather than utilise paper and pencil.

\[
\begin{array}{c}
16 \\
+ 27 \\
\hline
30 \\
\end{array}
\]

A. Vertical form algorithm

\[
27 + 16 \rightarrow 30 + 13 = 43
\]

B. Horizontal form strategy

Figure 1. Two approaches to an addition problem

To support this approach to mathematics education, the Ministry of Education had promulgated a Numeracy professional development project supported with a diagnostic oral interview assessment (NumPA) (Ministry of Education, 2008). As the asTTle team moved into the development of primary school mathematics test questions, the mathematics curriculum advisor within the Ministry sought to have the asTTle mathematics component stopped altogether because it was not formally aligned with the new experiential-based approach being advocated and already being deployed in the school community. After failing to stop the extension of asTTle into mathematics, the advisor insisted that all items below Level 4 of the curriculum must not appear in the conventional vertical form. Careful inspection of the items revealed only one item that had been trialled in both horizontal and vertical form, since in accordance with the criterial emphasis in developing curriculum assessments of mathematics had focused on the product (i.e., the answer), rather than the process students used to answer questions. This lack of vertical form items was explicitly commented on by Darr (2003) in an early evaluation of the system.

To further address the concerns of the official, a research project was undertaken in which test items were developed that probed student strategy or process usage rather than simply accuracy of responding. A review of the international literature on number knowledge, number strategies, and frameworks for classifying children’s (aged 9-11) learning of number was commissioned (Ell, 2002). The report identified that there is a general progression from concrete thinking tied to physical models and counting methods to abstract thinking using known number facts and relationships, but these progressions were not formal developmental phases, but rather instructional sequences that help teachers to decide ‘where to next’ for a child. Nonetheless, the review also identified considerably more strategies than were implemented in the New Zealand Numeracy Framework.

A pilot study created 39 items relative to the New Zealand Numeracy Framework and it was found that many items did not challenge highly skilled students to actually use a mental strategy (van Garderen, 2001). Further, it was recommended that procedures be developed to determine whether using a more basic strategy provides robust evidence that the student could not use more advanced strategies. A new set of items was developed with one of the Numeracy Project developers and trialled with students whose numeracy strategy ability had been determined by the recommended oral interview process by a highly regarded classroom teacher. That second study showed that students consistently demonstrated higher
or more complex strategy use in paper-and-pencil format than in the NumPA oral interview, suggesting that such items could be incorporated into asTTle should they be required (Hughes, Hattie, Brown, & Leeson, 2004). Consequently, both assessment methods were retained with the Ministry providing substantial resources and requirement for schools to use NumPA in the first 4-6 years of schooling. Nonetheless, this debate reveals a criterial vs. experiential tension played out not just in pedagogy and curriculum but in the very confines of the commissioning Ministry itself—perhaps a house divided is a good basis for development of resources teachers and policy makers actually want.

Technology and Curriculum

The asTTle system was deployed under the auspices of The Ministry of Education Curriculum division, since this group had responsibility for classroom and school assessment. The international tests (e.g., PISA, PIRLS, TIMSS) which New Zealand participated in were administered through the Research Division and New Zealand’s school qualifications system was administered through the quasi-independent body the New Zealand Qualifications Authority. In this case, it should be apparent that the asTTle development team with their software solutions developer were clearly better positioned to address criterial expectations about how assessment and education technology should be developed and deployed, while the Ministry curriculum officials would have relied much more on experiential bases for understanding the development and deployment of resources to help schools.

At first, this was a happy arrangement until it became apparent that asTTle would require considerable financial obligations beyond its development cost. For asTTle to operate successfully in schools was quite a different financial experience than releasing a set of text books or course materials; the software had to be maintained in the face of multiple operating systems and continually changing software environments. Indeed, not only did the Ministry have to supply assessment development services (AtoL) around the country, but they also had to provide a free-phone IT Help Desk to support teachers and schools struggling with technical issues (e.g., navigating anti-virus systems, communicating with school management systems, etc.).

A few examples illustrate unexpected challenges for a group of Curriculum officers concerned for the quality of teaching, learning, and resources for schools created by the digital nature of asTTle.

- When initially developed, asTTle was required to operate on Windows 95 and XT and Mac OS9 and 10 operating systems. Not long after the completion of asTTle v2 in December 2002 and v4 in December 2004, the MAC operating system changed how third-party applications (e.g., Adobe Reader) were launched meaning that the asTTle system was perceived in January of 2003 and 2005 respectively as ‘broken’. This meant the Ministry had to purchase a fix for the asTTle system since the change took place after formal acceptance of the software. This problem, while predictable to anyone in the IT industry, was outside the experience basis of ministry officials used to print-based or fixed-form technologies (e.g., CD).
- Very quickly after the release of V2, schools wanted local area network functionality so that tests and data could be shared by school leaders, syndicate or department managers, and classroom teachers. This was funded by the Ministry to support certain Linux, Mac, and Windows server environments. However, after release of V4, a small group of high schools, operating Novell networks, were not able to install asTTle on their systems. This escalated into a request to the Minister who authorised purchase of an extension of the asTTle system to cope with this additional environment. Here, asTTle was delivered according to criterial procedures to a Ministry that actually lacked full experiential understanding of the school IT
environments. This may have arisen partly as a function of the decentralisation of school administration in the 1989 Tomorrow’s Schools policy (Department of Education, 1988; Fiske & Ladd, 2000; Lange, 1988; Wylie, 1997) such that ministry officials could no longer claim definitive knowledge about school infrastructure.

- A persistent complaint of schools was that using asTTle tests was extremely expensive for classroom teachers who had to print and copy write-on tests for each student. In accordance with recommendations about test design, the asTTle team had purposely prioritised much white-space around test stimulus materials and ensured that marking guides were based on this expansive presentation. However, in order to economise on budgets, teachers were wont to reduce pages. Being aware of this, the asTTle user manual published with each version of the software warned teachers to adjust scoring guides to comply with any photo-reduction that they carried out. Indeed, I answered a complaint from a school’s deputy principal that the answer guide was wrong for a set of measurement questions, which turned out to be a consequence of having reduced the size of the questions from the original without adjusting the answer guide to fit the new size. This illustrates tensions between criteria derived from the Ministry’s product specifications and the experience-based approach of schools; perhaps it also illustrates the tension teachers have of adapting to standardised tests that require professional judgement.

It is not surprising, then that officials, charged with responsibility for maintaining and extending asTTle, appear to have gradually distanced themselves from asTTle. It may have been that the realisation that asTTle would require continuous financial support simply to maintain it as technologies constantly changed. It may also have been that, as the Ministry developed a much more explicit policy that assessment was ‘in the moment’ and ‘on the fly’ (Ministry of Education, 2007), the role of a standardised test was more easily seen as summative and evaluative, rather than diagnostic and formative. As indicated earlier, asTTle received strong support from the Minister of Education which may have elicited some resentment over a politician treading upon decisions that more properly belong to the Ministry officials. I understand that during one annual budget run the Minister of Education rejected the Ministry plan because it did not include a budget line for extending asTTle to include the development of a fully electronically administered version ‘e-asTTle’. This came then as an urgent request to the University of Auckland asTTle team to prepare a budgeted proposal for e-asTTle that had to be ready within just two-three weeks.

It should also be noted that the development of asTTle v1-4 took place without involvement from or accountability to the Ministry’s Information Technology Governance Board. It is normal within government circles to insist on full specification of deliverables prior to the authorisation of a contract for development of information technology systems. Up to the conclusion of V4, the asTTle team worked with a software development team and the Ministry of Education using a ‘time-box’ methodology. Time-boxes (approximately 3 months each) allow developers and clients to agree on a set of priorities (ranked Must, Should, Could, or Would) to appear in a product by the end of the time-box. The software developers guarantee that ALL Musts will be delivered within the set time period and the agreed budget, noting that some Should priorities will be delivered depending on success with the Musts. This approach allows flexibility so that every 3 months priorities can be adjusted according to experience with the product and its eventual users. However, the deliverables are somewhat uncertain since only the Musts are guaranteed, along with the timeline and budget. This approach worked very successfully over the four-years that our software team worked on asTTle v1-v4. However, when approaching a new contract for e-
asTTle, we had to be very persuasive that our experiential criteria derived from a methodology that successfully delivered four versions of the software on time and on budget, were not subverted by the criterial approach prioritised by standard and conventional approaches to software development. It also helped that asTTle was rare—almost no government software development projects are ever delivered on time and on budget and meeting expectations.

However, these technical issues and the struggle to rein in the burgeoning costs of delivering all that schools wanted has meant that the Ministry of Education has sought to commercialise asTTle. The Ministry of Education in 2012 publicly sought tenders from private providers to commercialise the asTTle service both domestically and internationally. It remains highly doubtful that a buyer can be found since schools are unlikely to be willing to pay for a tool that has up-to-now been provided free.

**Psychometric debates: Alternative criterial bases**

Criticism from various parties in the education sector emerged. For example, the decision to merge six attitude items into a single scale was criticised for inappropriately merging two separate constructs (i.e., liking and self-efficacy) and using an inappropriate statistical method to report student attitude in mathematics (Irwin & Irwin, 2005). This criticism arose because within the psychometric community there is an apparent disagreement around appropriate methods and models for item analysis. The traditional view (classical test theory), held by most teachers and test developers up to the 1980s, approaches items by determining each item’s correlation with the total score ($r_{pb}$) and the tendency of items to inter-correlate with each other (scale reliability often indicated by Cronbach’s alpha). The dominant view of contemporary test developers is that item quality is determined by a probabilistic approach (item response theory) that uses logistic analysis to identify the difficulty of an item, its discrimination, and degree of chance for selected response items (Hambleton, Swaminathan, & Rogers, 1991). Within the IRT community there is a view (e.g., Bond & Fox, 2007) that the one-parameter model (Rasch) is superior to multi-parameter models since only the Rasch model provides true measurement. In contrast, the position of the asTTle team was that the Rasch model was sufficient for reporting performance to students, only after items with poor or negative discrimination and high guessing levels were removed; hence, a multi-parameter IRT methodology had been implemented in preparing items, but a one-parameter model was used to score and report student performance.

The criticism was based on a one-parameter logistic (Rasch) analysis of the six attitude items administered to about 300 students, in contrast to the asTTle team which used maximum likelihood estimated factor analysis of over 20,000 student responses to create two inter-correlated factors that were reported as the average of the six responses. Further, the asTTle attitude scale was criticised for having a correlation below .30 with performance on the asTTle test items. The asTTle team responded in an Advisory Report to the Ministry pointing out technical issues in the criticism and identifying similarities in the results (Hattie, Brown, Irving, & MacKay, 2005). Specifically, the asTTle team argued that the six items formed two identifiable constructs which were robustly correlated, had reasonable internal consistency for the number of items, and claimed that the low correlation with achievement did not discredit the items or scale but rather suggested that educators and researchers need to investigate the low association of attitude with achievement. Not surprisingly, this issue appears to have been ignored by the teacher community, perhaps because the subtleties of how attitude items are analysed does not matter in the experiential world of the classroom.

**Levels and Norms: Criterial and experiential approaches to defining progression**
Claims circulated that the asTTle system was out of sync with previously existing standardised achievement tests (i.e., Progressive Achievement Tests) which reported performance using norm-referenced stanine and percentile scores relative to year and age groups. Based on a normative interpretation of performance by year relative to expected curriculum levels, it was claimed that top students in Year 8 ought to be either in the top of Level 4 or the bottom of Level 5; however, asTTle test scores would often indicate that students, on the average, were only at the top of Level 3. This clearly suggested that the standards built into asTTle were too harsh. As a further illustration of this, in a conversation with a primary school teacher working with Year 6 students in a high socio-economic school, I discovered that she considered that her students could only be in Level 3 because of their age despite the high quality of work she could demonstrate that her students could do. Similarly, another teacher asked why her students were awarded Level 2 by asTTle when they clearly could read the Level 3 passages used in the asTTle reading tests; for her, it was expected that if you could read fluently the text, then you were performing at the level of the text. Further, since very high proportions of students in Year 11 (approx. 70%) got Achieved or better on the National Certificate of Educational Achievement Close Reading standard at Level 1 (which was supposedly set at Level 6 of the curriculum), there was a second experiential reason to consider that the cut score for curriculum levels had been set too high.

These confusions over levels and norms were addressed, in part, by two technical reports (Hattie et al., 2003; Hattie & Brown, 2003) as well as interaction with the various professional development teams employed by the ministry. In setting the cut-score for each curriculum level, a number of panels of experienced teachers were asked to judge using one of three different procedures how well students would have to do to be considered working within a level. After reviewing the teacher judgements, the team, in accordance with international conventions for standard setting (Cizek, 2001), adjusted downward the teacher recommendation so that the cut-score for being in a level required getting 65% accuracy on items within that level. In other words, to be Level 3 Basic, required getting more than 65% on Level 2 Advanced, and less than 65% accuracy on Level 3 Proficient. Furthermore, data from students with asTTle scores was investigated to discern why students might be getting lower levels than their rank-score on other measures might suggest. This analysis demonstrated that students could have Level 4 in the surface component of the asTTle reading test but be in Level 3 on the deep cognitive processing requirements; and thus awarded an overall score in Level 3. At the same time such a student was actually scoring in a very similar rank-order range as he or she was performing on general ability standardised tests. Further, the emphasis teachers placed on mapping levels to school years meant that they were dependent on student performance of the outcome objectives specified by the curriculum documents for each strand and level as was essential for the development of a test system aligned to the curriculum statements.

Thus, it would appear that the Ministry’s curriculum appeared to teachers as revolving around years and ages, while for assessment developers the curriculum was dependent on outcome statements concerning what students needed to know or do so as to make progress through the levels. Fundamentally, this meant teachers and educators had not moved from the traditional convention that high-rank equated with high levels of performance, despite this being the intention of the National Curriculum Framework introduced about ten years earlier (Ministry of Education, 1993).

Conclusion
While it is possible to argue that standardised tests linked to national curriculum objectives and levels are an externally-imposed, criterial system, this short description of asTTle’s place in the New Zealand curriculum system has shown that it is possible to develop
a test system that is experiential, local, and contextualized. The second conclusion is that this process exposed multiple, competing criterial bases for evaluating technology, test scores, and curriculum—debates were not simply criterial vs. experiential; rather they were your criterial standards vs. my criterial standards. This clearly points to problems in evaluating schools—everyone has a standard and criteria—they just may not be the same as anyone else’s.

Multiple factors conspired to permit this development, chief among which must be the professionalism of New Zealand teachers, the policy which made teachers the active managers of their own schools, the political skills of the project director (Prof John Hattie), the support of a strong Minister of Education, and the commitment of all education parties to prioritise educational improvement as the purpose of testing rather than school evaluation. Nonetheless, the process of making explicit what the curriculum framework meant in the asTTle test system raised issues about assessment processes, technology, finance, curriculum, and progression. It is only as the attempt to integrate criterial approaches into the pre-existing experiential frameworks was attempted that tensions and conflicts became evident. These contrasting bases of evaluation are not easily resolved, but there does seem to be a way through.

References


