8

Effective Reporting for Formative Assessment

The asTTle Case Example

Gavin T. L. Brown, Timothy M. O'Leary, and John A. C. Hattie

Assessment should have a purpose. As Zumbo (2009) stated, in the context of discussing validity, 'it is rare that that anyone measures for the sheer delight' (p. 66) going on to concede that measurement is 'something you do so that you can use the outcomes' (p. 66). Within educational contexts, there are many ways testing might be expected to used and improve schooling (Haertel, 2013), as well as many ways users might anticipate using test results (Hopster-den Otter, Wools, Eggen, & Veldkamp, 2016). One key use, perhaps the primary use, of educational assessment is the support of student learning (Popham, 2000). Given such improvement purposes for tests, validity requires that reports on student performance be well aligned to the test (and the test well aligned to the intended curricular goals) and well designed to ensure understanding (Tannenbaum, this volume).

In any system that expects teachers to monitor and respond to student learning, teachers are important users of test information. In such systems, the teacher's role is primarily to mediate test score information into appropriate instructional decisions (e.g., pace of progress, student grouping, task and activity design, selection of curricular resources, etc.). The focus of this chapter is on the communication of test results to teachers in ways that foster interpretations and actions that align with those intended. Shepard (2001, 2006) makes it clear that most educational assessment is carried out in classrooms by teachers and that significant improvements are needed in how testing might continue to play a part in that process. Teachers are expected to make a series of qualitative interpretations about observed student performances, as well as interpretations of test scores (Kane, 2006). These interpretations occur as teachers interact with students in the classroom and are not simply recorded for later interpretation. While modern directions in assessment design focus on ensuring that a robust theory of learning or cognition is present (Pellegrino, Chudowsky, Glaser, & National Research Council, 2001), it seems more appropriate in evaluating test reports for teachers to focus on theories of effective communication and instructional action.

Within educational settings, the first goal of a diagnostic test score report should be to ensure that the test reports inform teachers' decision-making about '*who needs to be taught what next*' (Brown & Hattie, 2012). Extensive research on feedback (Hattie & Timperley, 2007) shows that

DOI: 10.4324/9781351136501-11

in order to close the gap between where students are and intended curriculum goals and standards, tests have to describe diagnostically the current status (strengths and weaknesses) of a student and point to action that the teacher and/or the student can take to improve learning so as to maximise the probability of attaining the success criteria of the lessons. It is to reduce this gap between where they are and where we want them to be that leads to the importance of assessment. This means that effective educational tests have to provide more than total score or rank order information. In order to make instructional decisions about curriculum, reports need to specify, among other things, how scores can be used (AERA, APA, & NCME, 2014), though relatively little is contained in the *Standards* about ensuring that report readers make appropriate interpretations. Test developers seldom provide validation evidence as to what report readers see in the reports and what they do with the information (Hambleton & Zenisky, 2013; Hattie, 2014; Hattie & Brown, 2010). Yet, it is these two issues which will determine if test reports contribute to improved outcomes.

As a consequence, the second goal of such a test report is, or should be, to improve the quality of teacher instruction and student learning (Popham, 2000). This agenda has been made increasingly explicit with greater policy and research emphasis on a variety of approaches to assessment including: formative evaluation (Bloom, Hastings, & Madaus, 1971), school-based assessment (Torrance, 1986), classroom assessment (Crooks, 1988), performance assessment (Darling-Hammond, 1994), alternative assessment (Birenbaum, 1996), assessment for learning (Black & Wiliam, 1998), and assessment for teaching (Griffin, 2014). What these approaches have in common is that they situate the design, administration, scoring and interpretation of evaluative processes in the midst of the instructional environment, rather than external to it. This improvement-oriented process has to take place early enough so as to make a difference to outcomes (Scriven, 1991) and is methodologically catholic in that it does not privilege or denigrate tests versus other methods (e.g., performances, portfolios, peer or self-assessment, etc.). These approaches all focus on generating data and decision-making about learning outcomes, much in the manner of total quality management (Deming, 1986), by the people closest to and directly responsible for educational practices and processes (i.e., teachers and school leaders).

Parallel to this is the need for reports on test data to reach the teacher soon after the test has been administered so that the information is relevant to where the learning was when it was tested. There can be no doubt that a report that arrives from a central test agency some three months or so after the test date is unlikely to be valid or effective. As we have argued before:

> the potential for that information to actually shape meaningful learning activities is practically nil—the students have changed class or grade, the teachers have moved on to new material, the class may have been successfully taught that content, and so on.
>
> (Hattie & Brown, 2008, p. 195)

Prompt feedback to the teacher as to which children have which needs or strengths is a *sine qua non* in ensuring that standardised tests serve educational rather than administrative or policy goals. Indeed, another feature of rapid reporting to teachers is the assurance it gives that they are the first to read the reports; delayed reporting may have been monitored and inspected by superiors before it arrives, more so in jurisdictions that prioritise testing for school accountability. Rapid reporting allows teachers early access to both pleasing and disturbing data and the chance to respond to it before external stakeholders inspect the results (Brown & Hattie, 2012; Hattie & Brown, 2008). Hence, rapid reports to teachers connects the test information to their current teaching context and raises the probability that teachers will actively respond to the data.

An important policy consideration that will support accurate teacher interpretations and decisions from test reports has to do with consequences or stakes associated with the test.

Good tests can lead to educators discovering some very discomforting news (e.g., the class or school is well below expectations and averages). In an environment where there are negative consequences (e.g., league tables), there can be strong incentives to game or cheat the test to avoid 'unfair' consequences. Hence, a low-stakes environment, creating a sense of psychological safety, is often needed to ensure 'bad' news in a test report is read and acted upon (Hattie & Brown, 2008). Helping teachers embrace the 'bad news' of poor scores so that correct diagnosis of need and prescription of appropriate instruction are maximised is the legitimate goal of test reports. Hence, effective test reporting depends, in part, on the existence of a non-punitive professional environment anchored on educators using data to improve curriculum, instruction, and learning (Lai & Schildkamp, 2016).

Defining Score Reports

Hambleton and Zenisky (2013), the foremost of contemporary score report theorists, described score reports as the vehicle 'to convey how scores can be understood appropriately in the context of the assessment and what are the supported actions that can be taken using the results' (p. 482). Rankin (2016) defined a score report as communicating data, through tables, graphs and words in order to achieve a purpose, typically helping to turn data into actionable information, for an intended audience. Thus, score reports are the tangible communication used to disseminate scores, which are the summarised results or output of some observable phenomena (test performance), to an intended audience. A score report may be a stand-alone single report, a series of reports, it may be bespoke or automatically generated, it may be a static online reporting environment or even a dynamic online reporting system. A score report may be any combination of the above or much more. More than simply the manner in which the outcomes of testing is reported, score reports are the thin lens through which the outputs from the complex process of assessment are communicated to its audience. Indeed, score reports are, arguably, far more than simply the output of assessments; they are part of the assessment they are reporting (O'Leary, Hattie, & Griffin, 2017b).

Score reports are then of fundamental importance to the intended outcomes of testing. More than simply the afterthought to the test development process, score reports are the integral link or interface in the communication between test developers and test score users. Effectively, score reports are decision support tools (Dhaliwal & Dicerbo, 2015) and shoulder the responsibility for supporting accurate user interpretation and use of test scores. As such, their design should be focussed upon optimising user interpretation and use (Zapata-Rivera & Katz, 2014). How well a score report does, or does not, communicate its message and subsequently influence the decision and actions of their intended audience is then critical, and, arguably, as important to the notion of validity as the other psychometric properties traditionally considered when undertaking validation (Hattie & Brown, 2010). Indeed, score reports are, arguably, far more than simply the output of assessments; they are part of the assessment itself.

Accepting that score reports are the mechanism through which performance is conveyed to an intended audience, it is evident that score reports are a form of feedback to those receiving the reports. Within educational contexts, diagnostic or interim assessments are the vehicle through which teachers receive feedback about the students in their class to assist in answering the question '*who needs to be taught what next*' (Brown & Hattie, 2012). In order to close the gap between where students are and the intended curriculum goals and standards, tests have to describe diagnostically the strengths and weaknesses of a student and point to action that the teacher and/or the student can take to improve learning (Hattie & Timperley, 2007).

The Challenges of Score Reporting

Interpretation and use of scores are of critical importance to validation efforts and any subsequent claims about validity (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). However, interpretation and use of scores does not transpire purely because testing occurs. Interpretation and use are the conclusion of the complex process of testing and occurs solely because of audience engagement and interpretation of the output of test score reports. In fact, how well a score report does, or does not, communicate its message and subsequently influence the decision and actions of its intended audience is critical and as important to the notion of validity as the other psychometric properties traditionally considered when undertaking validation (Hattie, 2010).

Unfortunately, however, validity theory and validation practice rarely incorporate explicit references or guidance about how to deal with the actual (as opposed to the intended) interpretations made by report users, nor the consequential actions of score users' engagement with score reports. The literature on score report design date back almost three decades. That literature persistently identifies that test users have difficulty in understanding test scores as intended, across a range of report formats (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998; Van der Kleij & Eggen, 2013). The last 25 years has seen significant contributions to the design of test reports from the information display literature (Bertlin, 1983; Cleveland, 1994; Few, 2012; Kosslyn, 2006; Tufte, 1990, 2001; Wainer, 1997). For example, Tufte (2001) identified seven principles of graph design which are pertinent to any effort to represent test scores graphically:

1. Show the data
2. Direct the reader to think about data being presented rather than some other aspect of graph
3. Avoid distorting the data
4. Present data using the minimum of ink
5. Make large data sets coherent
6. Encourage the reader to compare different pieces of data
7. Reveal the underlying message of the data.

As a consequence of significant work (Aschbacher & Herman, 1991; Hambleton & Slater, 1997; Hambleton & Zenisky, 2013; Hattie, 2010; Impara, Divine, Bruce, Liverman, & Gay, 1991; Jaeger, 1998; Linn & Dunbar, 1992; Rankin, 2016; Zapata-Rivera & Van Winkle, 2010; Zenisky & Hambleton, 2012, 2015), there has been an evolution of guidelines relating to score reporting. These guidelines have been integrated with explicit notions of user validity (MacIver, Anderson, Costa, & Evers, 2014) and of score report interpretability as an aspect of validity (Van der Kleij, Eggen, & Engelen, 2014). The ongoing advancement of score reporting guidelines has seen a progression from recommendations about what and how to produce score reports through to iterative design methodology (Hambleton & Zenisky, 2013; Zapata-Rivera & Van Winkle, 2010).

Hattie (2010) enunciated 15 principles for the design of test reports which align in part with Tufte and also extend to address issues arising when test reports are embedded within software systems. For example, he recommends in accordance with Tufte that reports (Principle 6) minimise the amount of 'numbers' and maximise the amount of interpretations, (Principle 8) have a major theme, (Principle 10) minimise scrolling, be uncluttered, and maximise the 'seen' over the 'read'. In terms of deploying test reports within a software system, he recommends (Principle 3)

that readers of reports need a guarantee of safe passage from where they are in the system to where they want to go and (Principle 4) report readers need a guarantee of destination recovery; that is, the system must intuitively allow them to navigate among the various reports and tools within the human-computer interface. He also recommends (Principle 7) that reports be restricted in the amount of information displayed (i.e., the answer is never more than 7 plus or minus 2).

Current best practice is captured in the Hambleton and Zenisky model (2013) and comprehensively described by Zenisky and Hambleton (2015) in the *Handbook of Test Development* (Lane, Raymond, & Haladyna, 2015). This model is an iterative process of score report development and refinement. The process is conceptualised as a four step or phase model. The first phase is about laying an appropriate ground work. The second phase is about report development. The third phase is about field test and redesign. Finally, the fourth is about evaluation and maintenance. One of the key aspects that makes this model best practice is that it is focused on an ongoing process of improvement and refinement and not simply static guidelines. Consistent with the Hambleton and Zenisky model, Hattie (2010) recommended that (Principle 1) the validity of test reports be determined by the reader's correct and appropriate inferences and/or actions in response to the report, (Principle 2) evidence be obtained to demonstrate how readers interpret reports, (Principle 5) the focus be on maximising interpretations not displaying numbers, (Principle 11) reports be designed to address specific questions, (Principle 12) provide justifications that the test is fit for the specific applied purpose, and (Principle 15) reports be thought of as actions to take, not just screens to print or store.

With an eye towards the Hattie (2010) principles, O'Leary (2017) has proposed amendments to the Hambleton and Zenisky (2013) model, aimed at providing more explicit articulation to the evaluation phase of their model. The goal of those recommendations is to direct the collection of evidence concerning user comprehension of score reports. Two overarching design principles of evaluation (i.e., utility and clarity) with seven sub-domains have been promulgated (Table 8.1; O'Leary, Hattie, & Griffin, 2016b). Utility requires that score reports are designed with a clear purpose, actions, and outcomes in mind, while clarity expects score reports to be designed so that they are easily comprehensible to the target audience. These align with the proposed forms of validity evidence put forth by O'Leary, Hattie, & Griffin (2016a, 2017a). The purpose of these principles is to provide an outcomes focused lens through which score reports are considered. A rubric for evaluating the alignment of score report construction against these criteria has been developed (O'Leary, Hattie, & Griffin, 2016b) and subsequent empirical work has demonstrated

Table 8.1 Empirically derived design principles for outcomes focused evaluation of score reporting.

Utility	
Purpose	The purpose of a score report must be explicit.
Interpretation	The intended interpretations of scores must be explicit.
Actions	The intended consequences or actions of interpretation must be explicit.
Clarity	
Design Features	The design of score reports must be based upon current best practices inclusive of contemporary examples of best practices and guidelines and recommendations from within the literature.
Interpretive Guidance	Score reports must be designed to be stand-alone aiming to minimise additional work or tasks that are required to fully interpret the reported information.
Displays	Score reports must integrate multiple forms of data representation.
Language	The language used in a score report must be easily understood by the intended audience.

that the rubric is a reliable tool for obtaining evidence that 'better' designed score reports more effectively communicate their intended message (O'Leary, Hattie, & Griffin, 2017b).

These standards can be used to evaluate any test-based reports as attempts to communicate expert information to a lay end-user audience. Unsurprising then, the role of timing is not explicit. A separate argument about the importance of rapid or delayed reporting needs to be made but which could be subsumed under the notion of validity. As we have already indicated, test reports which are not available to teachers soon after test administration cannot guide instruction. Further, in light of quality management principles, providing additional insights about directions for improvement needs to be timely; and for classroom teachers timely is next Monday or tomorrow, not three months from now. Hence, well-designed test reports that arrive too late to make a difference are of little value to formative practice. The advantage of pure 'in-the-head' and 'in-the-moment' formative interactions between teachers and students (Swaffield, 2011) is that it happens immediately, while the need or opportunity is evident. Such interactions may be more error prone than tests, but they are immediate. Matching this timely facet matters to teachers and teaching; delayed reports are fundamentally purposeless.

The Audacious Example of asTTle

To illustrate the principles enunciated in this chapter, it is constructive to examine the development of New Zealand's Assessment Tools for Teaching and Learning (asTTle) test system (Hattie & Brown, 2008; Hattie, Brown, & Keegan, 2003; Hattie, Brown, Keegan, et al., 2004). The asTTle test system is an online standardised test system for reading comprehension, mathematics and writing (and the Māori language equivalents) used in New Zealand primary/elementary and secondary/high schools. The test materials have been calibrated to Levels 2 to 6 of the New Zealand national curriculum (Ministry of Education, 2007) and norms are available for students in grades 4 through to 12 (nominally ages 8 to 17). The asTTle system consists of:

1. an item bank of over 20,000 curriculum-objective and level calibrated and difficulty-calibrated multiple-choice and open-response tasks,
2. a teacher-controlled test design engine,
3. an automated test scoring engine that converted 1PL Rasch item scores to performance on achievement objectives and Curriculum Levels,
4. a reporting engine that permitted selection from a range of test reports concerning group and/or individual performance, and
5. an online catalogue of teaching resources indexed to the test reporting system.

This system was created in a policy environment that prioritised diagnostic testing for the explicit purpose of informing improved instruction and student learning outcomes (Ministry of Education, 1994). Indeed, the official policy and the rhetoric used around the research and development phases of asTTle made explicit that using the test system was a low-stakes activity; use was not required nor was reporting to government and there was no centrally determined test administration (Hattie & Brown, 2008). Furthermore, as was made clear by the Ministry of Education (2010), the system was designed to inform and support teachers by giving them access to externally-referenced norms and diagnostic curriculum-aligned reports, rather than a mechanism to be used by the Education Review Office or the media to judge or evaluate teachers and/or schools. This ensures that generation of data about student learning was done in a non-punitive manner; the goal was to inform improvement, while generating data that allowed teachers to understand how their own students compared to similar students drawn from a robust national norming (Brown & Hattie, 2012).

New Zealand primary school teachers tend to make extensive use of standardised diagnostic testing, especially at the beginning of the school year to inform within-class grouping (Crooks, 2010). It is important to note that none of the standardised tests available through the New Zealand assessment 'tool box' were compulsory or nationally administered as a national test (Brown, Irving, & Keegan, 2014); the use of all tests was completely voluntary with data retained at the school level. However, the standardised tests available before asTTle were general ability tests and reported only total score and rank order performance information. These limitations were overcome in asTTle because the system (a) allowed testing at any time, (b) allowed teachers to customise tests to classroom teaching, (c) calibration allowed different tests to be compared over time and over classes, and (d) reported performance on curriculum achievement objectives and levels, as well as normative performance.

Hence, the overall goal in designing the asTTle test report system was to give teachers a sufficiently accurate portrayal of student strengths and weaknesses so that teachers could make appropriate decisions about *who needs to be taught what's next*. This meant that the level of accuracy required in reporting a score was determined by whether the teacher would make a defensible decision about curriculum materials, pedagogical activities, or student grouping. In a sense, a principle similar to Goldilocks was used in that there were really only three options; curriculum content and material too easy, just right, or too hard. Since teachers already have a reasonably accurate sense of rank order within a class of students and have already made judgements about the curriculum level which they are teaching, a good test report system would have to go beyond this extant information. A good system would have to tell teachers something that they did not already know; teachers should be surprised rather than comforted.

Alpha Testing

To achieve this, a series of teacher interviews and focus groups were conducted early in the system development process to determine the administrative and educational goals that teachers and school leaders had for an assessment event (Meagher-Lundberg, 2000) (Note, all asTTle Technical Reports are signified * in the reference list). That research identified that information comparing their own students to national norms was desired in order to report to a range of stakeholders (e.g., parents, trustees, staff), to inform school and staff self-appraisal and professional development, to plan teaching, and to target resource commitments. Additionally, teachers wanted descriptive information relative to curriculum levels and achievement objectives. With this information, the design of test reports was initiated. This involved collaborating with a graphic artist who created simulated screen shots for a set of report templates that might achieve the various goals identified by the teachers and which were deemed to be feasible through an objectively-scored test. These report templates were iteratively presented to teacher focus groups (Meagher-Lundberg, 2001a, 2001b) to ascertain that teachers could make the intended interpretations. Initial reaction to the designs indicated a strong need for clarity as to how navigation between reports would be conducted. Teachers indicated initial designs lacked clarity as to the meaning of various communicative devices such as coloured fields depicting normative information, arrows, dials, numeric scales, labels, and the position and salience of explanatory terms (Meagher-Lundberg, 2001a).

In light of this feedback, further revisions were created taking advantage of graphical communication insights obtained from the research literature (Brown, 2001). The navigation problem was successfully addressed subsequently by placing the report images in a browser window (Meagher-Lundberg, 2001b); taking advantage of existing end-user preferences and knowledge about how software operated (Spolsky, 2001). Changes to the devices used to communicate information were generally successful according to the second focus group. This led

to the design of a report engine that included a menu system to navigate to one of the following report templates: (a) a group or cohort achievement comparison console; (b) individual and group 'kid maps'; and (c) a curriculum level achievement 'skyline' showing proportions of group performing at each level. Additional features for reporting cognitive processing against the SOLO taxonomy (Biggs & Collis, 1982) and attitudes towards tested subjects were identified for integration into either individual or group reports. Note that the sample reports shown below (Figures 8.1 to 8.4) are from the current e-asTTle version (e-asTTle Project Team, 2009). As can be seen from the following example reports, each report provided interpretive guidance on-screen, addressed a single, clear educational purpose, with the goal of supporting actual teacher decision-making.

Note that the achievement cohort comparison console (Figure 8.1) drew heavily in its design on the previously deployed CRESST Quality School Portfolio report (Baker, 1999). That report used a series of gauges and dials to capture various quality aspects of schools (e.g., safety, technology, attendance, standardised test performance, etc.) and made use of traffic light colours to indicate level of concern (i.e., red= below average; yellow=average; green=above average). This report was intended more for the cohort, subject, or school leader who needed an overview of performance relative to national normative performance. In e-asTTle, a key is used to remind the reader that the normative performance of the related comparison group is the edge between the blue field and the white space and the performance of the tested group is shown as red pointers or box plots. Any aspects of the curriculum not covered by the test are greyed out to focus interpretation on the aspects for which there was sufficient information on which to base decisions and actions. Because it may be unfair to compare 'my school' with the whole nation, especially if my school's population is drawn from either the tails or tops of the socio-economic distribution, users are able to specify the type of comparative norm by selecting either student (e.g., sex, ethnicity) or school information (i.e., school cluster). The point of this selection is not only to drill down into performance of students meriting specific attention but also to remove the obstructive claim that my students cannot achieve because they are disadvantaged; if the average for similar students or school types is higher than one's own, then it does not hold that such factors in and of themselves prevent improvement.

Likewise, the 'kidmap' reports (Figure 8.2) draw on the work of Wright and Stone (1979) in which performance is classified into one of four spatial fields or categories; that is, (a) correct and easy, (b) correct but hard, (c) incorrect but easy, and (d) incorrect and hard. This is achieved through comparison of student accuracy on the item (i.e., correct vs. incorrect) according to the difficulty of the item relative to the student's overall performance. Rather than listing items in each space, the asTTle system reports achievement objectives in each field, supplemented by item numbers in order to maximise attention on the teaching of learning outcomes rather than test items. Clearly, this report was designed for the classroom teacher or counsellor who needed to discuss with a parent or guardian the specifics of an individual child. The report uses the same conventions as the Console report to indicate overall performance relative to the same grade level norm both in terms of subject performance and motivation or attitudes. This permits partnership discussions between teacher and parent with the student to identify priorities for both work at home, as well as work in class.

To cater for the reality that teacher planning has to address groups of students (e.g., classes, grade cohorts, or special categories), the individual kidmap learning pathways report was transposed using the same colour coding to point teachers to the proportion of students having strengths, mastery, weaknesses, or gaps according to curriculum objectives (Figure 8.3). To enable priority-making decisions, teachers had only to look for objectives for which the blue space (i.e., to be achieved) were large and those in which the green space (i.e., achieved) were large. The

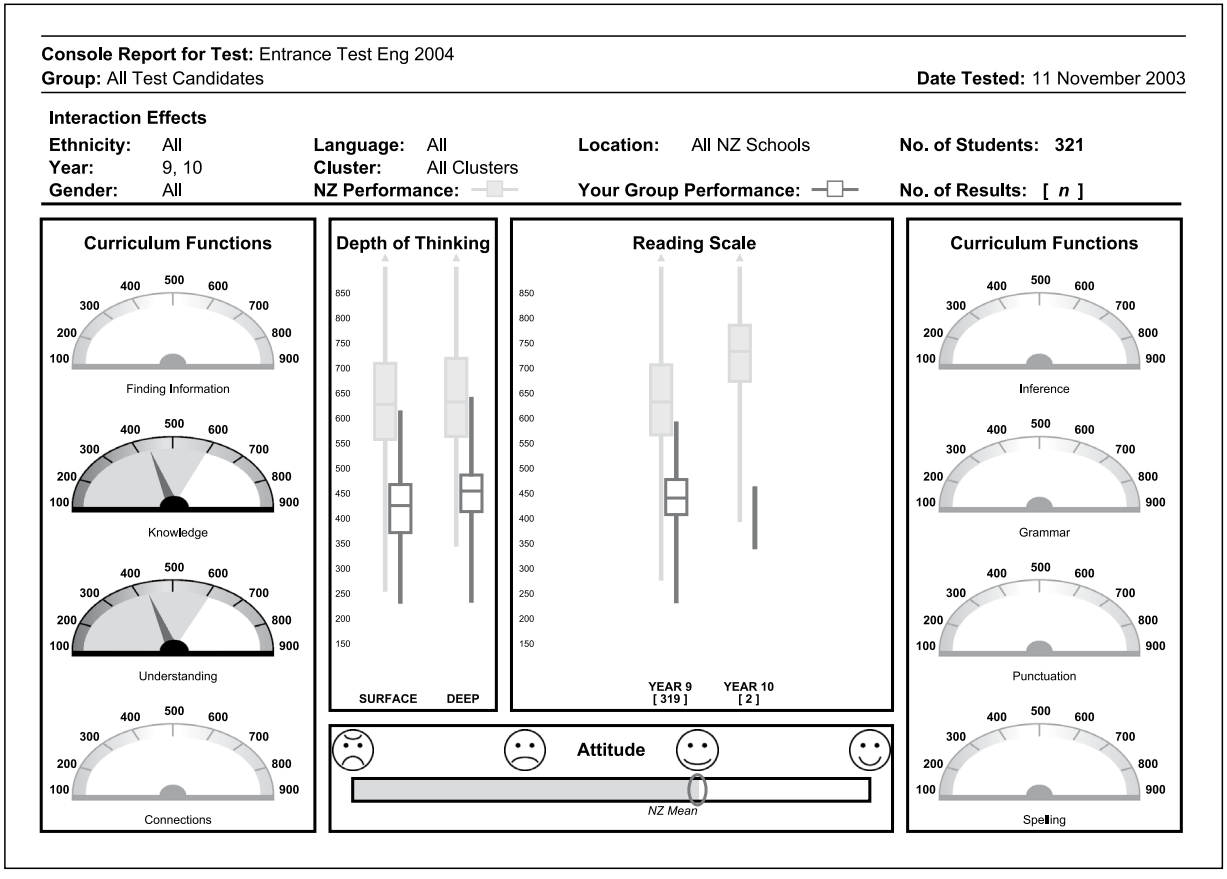

Figure 8.1 e-asTTle console report.

(e-asTTle Project Team, 2009, p. 78)

Note: Light gray is displayed as blue on-screen, the pointers and dark gray boxes are red.

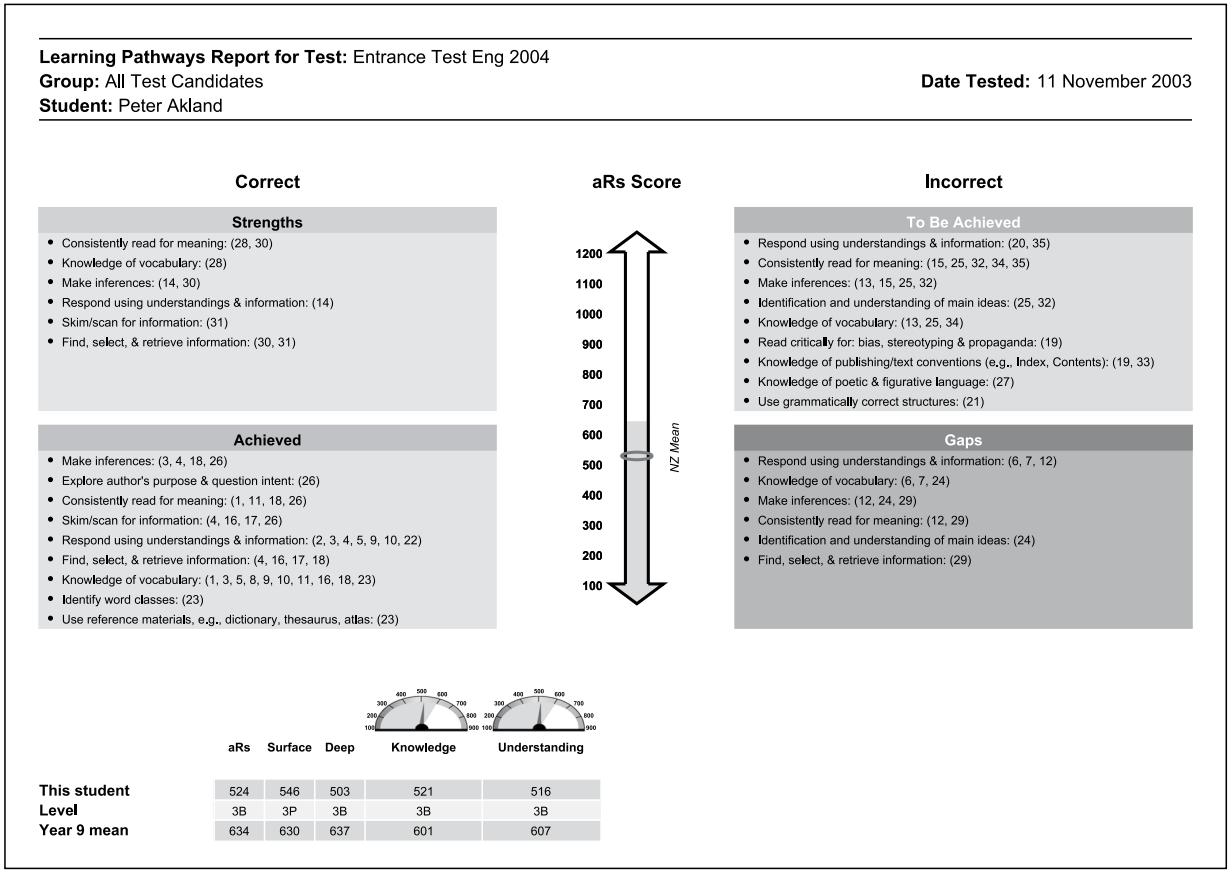

Figure 8.2 e-asTTle learning pathways report for an individual student.

(e-asTTle Project Team, 2009, p. 84)

Note: Achieved is displayed as green on-screen; Strengths is yellow; Gaps are red; and To Be Achieved is blue. The ellipse and pointers are red on a blue background.

Group Learning Pathways Report for Test: Entrance Test Eng 2004
Group: All Test Candidates
Group Size: 321
Date Tested: 11 November 2003

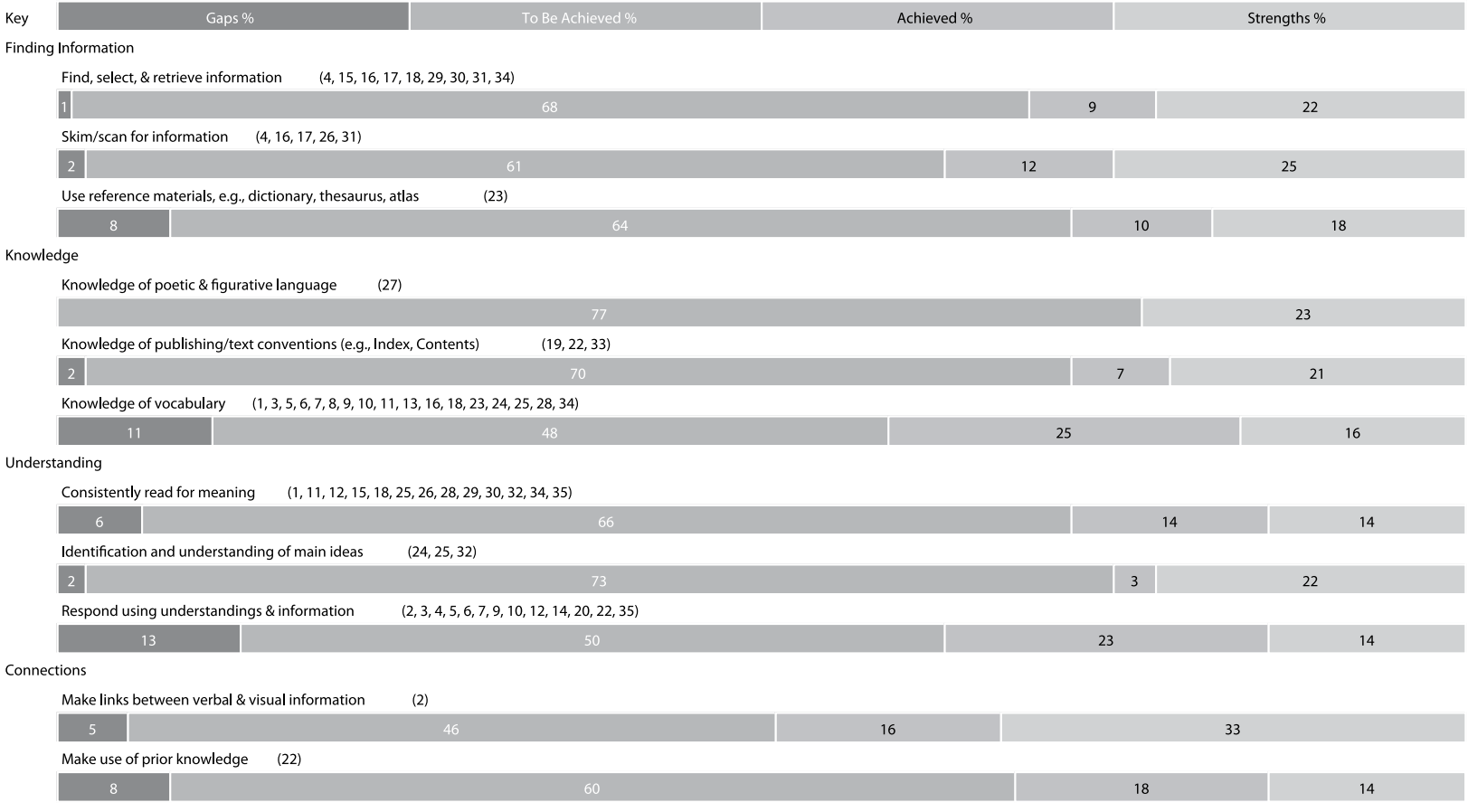

Figure 8.3 e-asTTle group learning pathways report.
(e-asTTle Project Team, 2009, p. 88)
Note: Achieved is displayed as green on-screen; Strengths is yellow; Gaps are red; and To Be Achieved is blue.

former indicated content that a high proportion of students needed to be taught, while the latter indicated material on which a high proportion needed no further instruction or practice.

Unsurprisingly, teaching to the mean will disguise the distribution of performance. Hence, the system provides a distribution of performance report (i.e., Curriculum Levels Report; Figure 8.4), which reveals both central tendency and distribution. Because New Zealand primary school teachers practice considerable within-class ability grouping, each 'skyline', when selected, displays the names of students in each performance group. This allows teachers to move children into different grouping combinations according to identified needs, rather than create persistent groups across all learning areas. In fact, this ability to differentiate for grouping was noted by early adopting teachers and their students as a positive facet of the system (Archer & Brown, 2013).

We suggest that the suite of reports and the ability to customise those for the multiple purposes of classroom teachers and school leaders meant that the system complies with the expectations of good reporting outlined in Table 8.1 and earlier.

Beta Testing

Having established through 'alpha' testing reasonably robust communicative test reports, these designs were further refined through 'beta' feedback from (a) Ministry of Education officials who were the funders and sponsors of the asTTle system, (b) the software engineering team who advised on feasibility and cost of various design options, (c) pilot testing by teachers who were exposed to a mock-up of the system, and (d) acceptance testing of asTTle version 1, containing materials for reading and writing only, which was deployed to 110 primary schools. As each stage of beta testing was conducted, formative changes were made to the asTTle system to achieve the curricular goal of helping teachers know what to teach to which students.

The evaluation of the pilot implementation of asTTle (v1) into 110 New Zealand schools (Ward, Hattie, & Brown, 2003) used a survey to ascertain, among others, the ability of teachers to accurately interpret asTTle reports. The survey included a set of report reading comprehension items, partially inspired by Hambleton and Slater (1997) and Linn and Dunbar (1992). Results indicated that in general, the Console Reports and the What Next reports had reasonably high levels of correct interpretation, whereas the means were much lower for Individual Learning Pathways and Curriculum Levels reports (Hattie, Brown, Ward, Irving, & Keegan, 2006). These results were incorporated into a structural equation model as dependent variables. The model proposed that attitudes towards computers, ICT, assessment and professional development would predict the level of involvement teachers had with the asTTle test system, which, in turn, would predict the teacher's evaluation of asTTle and their ability to answer the report reading comprehension questions. Indicating a belief that assessment is powerful for improving teaching, rather than for evaluating schools, and seeing the asTTle software as positive were clear predictors of accuracy in report interpretation (Hattie et al., 2006). The major messages were that professional development needed to be oriented most towards encouraging a positive attitude towards using ICT-based assessment as part of teaching and learning. 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.

Based on these studies, the Ministry of Education funded for several years multiple mechanisms to support teacher learning in making use of the asTTle system. A free-phone technology-oriented help desk was deployed so that callers using asTTle and e-asTTle on their local work-stations, school-based servers, and eventually the internet could have prompt help. When

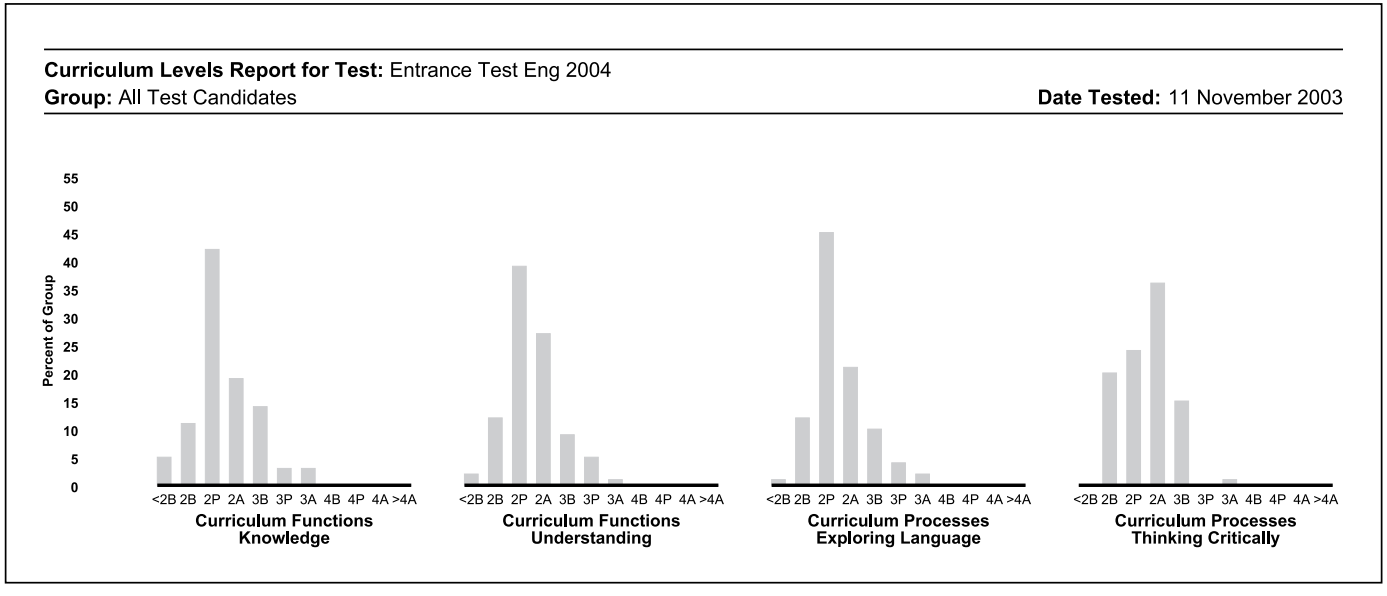

Figure 8.4 e-asTTle curriculum levels report.
(Hattie, Brown, Keegan, et al., 2004, pp. 3–18)

installed, the asTTle system provided user manuals and technical reports to support understanding and use of the system. These documents were also available online from the Ministry repository of reports and documents (http://e-asttle.tki.org.nz/). Throughout the nation, assessment-focused teacher professional development teams (Assess to Learn; AtoL) were commissioned and funded to provide within school services focused on the logic of using the asTTle reports to improve and guide instruction and reporting. The asTTle Project development team provided initial briefing to AtoL teams but was explicitly excluded from the delivery of school-based training. Nonetheless, it was apparent that the effectiveness of AtoL teams depended, in part, on the existing conceptions teachers had of the purpose of assessment—the more they considered assessment was for accountability, the less use they made of asTTle for improvement (Brown & Harris, 2009).

Extension to Secondary/High School

The original requirements brief for the asTTle system was focused solely on primary/elementary schooling; that is, Curriculum Levels 2 to 4, with norms for students in Years 5–8 only. However, given the success of asTTle v2 in primary schools, the Ministry of Education received vigorous requests from secondary/high school teachers and their union, the Post-Primary Teachers Association for extension of the system to include their students (Brown, 2013). The logic was reasonably simple: although the curriculum framework expects that Level 4 will be completed by the end of primary schooling (Year 8), empirical realities are such that many students arrive at high school still functioning at Levels 2 to 4 (Satherley, 2006). An environmental constraint in secondary schooling, not present in primary schooling, is the important role secondary schools play in preparing students for and administering formal qualifications assessments (i.e., the National Certificate of Educational Achievement-NCEA) (Crooks, 2010).

The NCEA begins with Level 1 in Year 11, culminating in Year 13 with Level 3. Nominally, NCEA Level 1 is equated to Curriculum Level 6, though some achievement objectives for Level 6 are taught in Year 12 rather than Year 11. The NCEA system evaluates student learning using a criterion-referenced, standards-based grade system (i.e., Not Achieved, Achieved, Merit, Excellence), somewhat akin to more conventional letter grade systems (i.e., D/F, C, B, A). NCEA also structures the curriculum objectives around units of work known as standards; this means that alignment of test items to NCEA standards might be of value to secondary teachers. This high-stakes evaluation system predominates educational assessment in New Zealand secondary school systems and so the possibility that the asTTle reports could be modified to accommodate this alternative system to the curriculum levels framework was explored. Additionally, within the framework of beta testing asTTle v3 in 55 secondary schools, accuracy and sufficiency of the test reports was conducted through a mixture of surveys, telephone interviews, and focus groups (Hattie, Brown, Irving, et al., 2004).

As reported in Hattie et al. (2006), secondary school teachers were positive about the asTTle reports, expressing satisfaction with the amount of detail on reports and the relevance of the reports to their needs. Teachers reported significant help from the formative and diagnostic reporting functions at both aggregated and disaggregated levels of reporting, especially in the Group Learning Pathways Report and the Individual Learning Pathways Report. In addition, they found benefit from the aggregated data in the Tabular, Curriculum Levels, and Console reports. Several enhancements based on feedback on asTTle v2 were evaluated positively. For example, instead of just reporting group means on the console report with an ellipse (i.e., *M*+/-*se*), a box-and-whisker plot showed the distribution of scores for the group being reported. The display of the national norm score as a coloured field within the dials instead of as a number below was also seen as an enhancement.

Secondary teachers indicated value in two new types of report. Focus group participants indicated value in longitudinal reports that showed how individuals or cohorts had been progressing over time. While the asTTle system had already included the ability to compare scores to similar students (i.e., schools like mine; Hattie, 2002), the ability to compare performance to different rather than similar categories (e.g., higher performing clusters or ethnicities) was seen as valuable.

Nonetheless, secondary teachers indicated significant concern about the correct or accurate interpretation of the asTTle reports. These concerns were obtained both from the Ministry Telephone Helpdesk as well as directly from the evaluation study. Confidence that reports were being understood and acted upon appropriately mattered to the teachers and needed to be addressed through modifications to the Ministry's professional development support services and asTTle documentation. Although most of the information sought by asTTle V3 users about report interpretation was available through the PDF manuals included with the asTTle V3 software, it was decided to develop an online tutorial system on understanding asTTle reports that could be used by individuals or schools as a supplement or alternative to professional development (Hattie, Brown, Irving, MacKay, & Campbell, 2005). Unlike later online tutorials that used video (Zapata-Rivera, Zwick, & Vezzu, 2016), these tutorials were slide presentations with voice over scripted dialogue that could be controlled by the user.

Lack of alignment to the NCEA system beginning in Year 11 meant that most secondary teachers had implemented asTTle V3 with students only in Years 9 and 10. However, when shown the possible reports and tests that asTTle might be able to generate for them as indicators of NCEA performance, teachers were quite enthusiastic. Teachers indicated that it was important or very important to know how the curriculum-level indexed items in an asTTle test related to the NCEA system. Of special interest to the participants would be the ability to create a test aligned to the various standards of the NCEA system, rather than to the achievement objectives of the curriculum. Despite the strong endorsement of the sampled teachers for adjustments of the asTTle reports to align with the official qualifications framework, the Ministry of Education sponsors declined to fund such research or developments. Perhaps, because the NCEA system is administered by a separate quasi-autonomous body (i.e., New Zealand Qualifications Authority; NZQA), such a development funded by the Ministry may have been seen as a breach of NZQA's autonomy and responsibility.

Aside from any systemic 'turf' issues, this last point raises some interesting challenges around alignment of formative and summative purposes. It may be that refusal to adapt asTTle to align with high-stakes qualifications system was that this may constitute a threat to the intention that asTTle serve goals related to diagnostic formative improvement of teaching and learning (Brown, 2004). When teachers perceive that assessments are for accountability purposes, our own studies have found that it is a rare teacher who can balance the tension between improvement of my teaching and evaluation of school quality (Brown & Harris, 2009). Indeed, this tension between the purposes or goals of assessment for improvement and assessment for accountability seems to remain more or less unresolved (Barnes, Fives, & Dacey, 2015; Bonner, 2016). As long as test systems are used by interested stakeholders to evaluate the work of teachers and school leaders, we can expect that there will be greater attention and effort paid to raising scores than improving instruction (Nichols & Harris, 2016). Hence, while the technical capacity exists to align formative and summative systems and information into a single test reporting framework, the real obstacles lie in political factors that may subvert well-meaning integration. Unless policy makers are willing to partner with the teachers and respect their legitimate concerns by attaching low-stakes to tests, it seems highly implausible that improvement and accountability can be effective bed-partners. Indeed, as long as high-stakes testing or examination or school accountability testing dominate the educational landscape, policies to support or require formative assessment are unlikely to be seen as 'the real thing' (Kennedy, Chan, & Fok, 2011).

Conclusion

This chapter has outlined the major challenges that face developing and validating test reports for teachers. The field has developed a reasonably robust understanding of why this has to be done and how it can be done. However, few test systems have conducted such time and resource-consuming programmes of formative evaluation and documented them as has the New Zealand asTTle system. This system is an exemplar of how accuracy in interpretation of reports and subsequent actions can be established. Clearly, the field needs more such studies that establish the validity of tests for use by communities of educators, each of which share different standards and approaches to assessment, ICT, and schooling in general.

References

*All asTTle reports are retrieved from https://e-asttle.tki.org.nz/Reports-and-research/asTTle-technical-reports

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Educational Research Association.

Archer, E., & Brown, G. T. L. (2013). Beyond rhetoric: Leveraging learning from New Zealand's assessment tools for teaching and learning for South Africa. *Education as Change*, *17*(1), 131–147. doi:10.1080/16823206.2013.773932

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Technical Report 326). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Baker, E. L. (1999). *Technology: Something's coming-something good* (CRESST Policy Brief 2). Los Angeles: UCLA Graduate School of Education & Information Studies, National Center for Research on Evaluation, Standards, and Student Testing.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. Gregoire Gill (Eds.), *International handbook of research on teacher beliefs* (pp. 284–300). New York, NY: Routledge.

Bertlin, J. (1983). *Semiology of graphics*. Madison, WI: The University of Wisconsin Press.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York, NY: Academic Press.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3–29). Dordrecht, The Netherlands: Springer.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw Hill.

Bonner, S. M. (2016). Teachers' perceptions about assessment: Competing narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 21–39). New York, NY: Routledge.

*Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of project asTTle outputs design* (asTTle Tech. Rep. #15). Auckland, NZ: University of Auckland, Project asTTle.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 301–318. doi:10.1080/0969594042000304609

Brown, G. T. L. (2013). asTTle—A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In M. Lai & S. Kushner (Eds.), *A national developmental and negotiated approach to school and curriculum evaluation* (pp. 39–56). London: Emerald Group Publishing. doi:10.1108/S1474-7863(2013)0000014003

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation*, *6*(12), 68–91.

Brown, G. T. L., & Hattie, J. A. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.), *Contemporary debates in childhood education and development* (pp. 287–292). London: Routledge.

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2014). *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (3rd ed.). Auckland, NZ: Dunmore Publishing.

Cleveland, W. S. (1994). *The elements of graphing data*. Monterey, CA: Wadsworth.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481.

Crooks, T. J. (2010). Classroom assessment in policy context (New Zealand). In B. McGraw, P. Peterson, & E. L. Baker (Eds.), *The international encyclopedia of education* (3rd ed., pp. 443–448). Oxford: Elsevier.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5–31. doi:10.17763/haer.64.1.j57n353226536276

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dhaliwal, T., & Dicerbo, K. E. (2015, April). *Presenting assessment data to inform instructional decisions*. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

e-asTTle Project Team. (2009). *Generation 2: e-asTTle year three educator manual*. Auckland, NZ: Visible Learning Lab, Auckland UniSevices, Ltd.

Few, S. (2012). *Show me the numbers: Designing tables and graphs to* en*lighten*. Oakland, MA: Analytic Press.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Griffin, P. (2014). *Assessment for teaching*. Port Melbourne: Cambridge University Press.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 1–18. doi:10.1080/15366367.2013.783752

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policymakers and educators?* (Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

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

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). Washington, DC: American Psychological Association.

Hattie, J. A. (2010). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1–15. Retrieved from www.oerj.org/View?action=viewPaper&paper=6.

Hattie, J. A., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, *36*(2), 189–201. doi:10.2190/ET.36.2.g

Hattie, J. A., & Brown, G. T. L. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: Concepts, research and challenges* (pp. 102–117). Abingdon: Routledge.

*Hattie, J. A. C., Brown, G. T. L., Irving, S. E., Keegan, P. J., Sussex, K., Cutforth, S., . . . MacKay, A. J. (2004, September). *Use of asTTle in secondary schools: Evaluation of the pilot release of asTTle V3* (asTTle Tech. Rep. #47), Auckland, NZ: University of Auckland/Ministry of Education.

*Hattie, J. A., Brown, G. T. L., Irving, S. E., MacKay, A. J., & Campbell, A. (2005). *Using asTTle: A teachers' guide* (asTTle Tutorial). Retrieved from www.breezeserver.co.nz/p85512844

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.

*Hattie, J. A., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., . . . Yu, J. (2004, December). *Assessment tools for teaching and learning (asTTle) manual* (Version 4, 2005). Wellington, NZ: University of Auckland/ Ministry of Education/ Learning Media.

Hattie, J. A., Brown, G. T. L., Ward, L., Irving, S. E., & Keegan, P. J. (2006). Formative evaluation of an educational assessment technology innovation: Developers' insights into assessment tools for teaching and learning (asTTle). *Journal of MultiDisciplinary Evaluation*, *5*(3), 1–54.

Hattie, J. A. C. (2014). The last of the 20th century test standards. *Educational Measurement: Issues and Practice*, *33*(4), 34–35.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Formative use of test results: A user perspective. *Studies in Educational Evaluation*, *52*, 12–23.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, *10*(4), 16–18.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kennedy, K. J., Chan, J. K. S., & Fok, P. K. (2011). Holding policy-makers to account: Exploring 'soft' and 'hard' policy and the implications for curriculum reform. *London Review of Education*, *9*(1), 41–54. doi:10.1080/14748460.2011.550433

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York, NY: Oxford University Press.

Lai, M. K., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 77–94). New York, NY: Routledge.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.) (2015). *Handbook of test development*. New York, NY: Routledge.

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.

MacIver, R., Anderson, N., Costa, A-C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*(2), 149–164. doi:10.1111/ijsa.12065

*Meagher-Lundberg, P. (2000). *Report on comparison groups/variable for use in analysing assessment results* (asTTle Tech. Rep. #1). Auckland, NZ: University of Auckland, Project asTTle.

*Meagher-Lundberg, P. (2001a). *Report on output reporting design: Focus group 1* (asTTle Tech. Rep. #9). Auckland, NZ: University of Auckland, Project asTTle.

*Meagher-Lundberg, P. (2001b). *Report output reporting design: Focus group 2* (asTTle Tech. Rep. #10). Auckland, NZ: University of Auckland, Project asTTle.

Ministry of Education. (1994). *Assessment: Policy to practice*. Wellington, NZ: Learning Media.

Ministry of Education. (2007). *The New Zealand curriculum for English-medium teaching and learning in years 1–13*. Wellington, NZ: Learning Media.

Ministry of Education. (2010). *OECD review on evaluation and assessment frameworks for improving school outcomes: New Zealand country background report 2010*. Wellington, NZ: Ministry of Education.

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 40–56). New York, NY: Routledge.

O'Leary, T. M. (2017). *Effective score reporting: Establishing evidence informed design principles for outcomes focused score reports*. Unpublished Ph.D. thesis, The University of Melbourne, Melbourne, Australia.

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017a). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, *36*(2), 16–23. doi:10.1111/emip.12141

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017b, April). *Evaluating the effectiveness of score reports: Do better designed score reports result in better interpretation?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

O'Leary, T. M., Hattie, J., & Griffin, P. (2016a, April). *Reconceptualising validity evidence including evidence of user interpretation*. Paper presented to Annual Conference of the NCME, Washington, DC.

O'Leary, T. M., Hattie, J., & Griffin, P. (2016b, July). *Design principles for action and outcome focused score report design*. Presentation at the biennial meeting of the International Test Commission, Vancouver, BC.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & National Research Council (U.S.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6th ed.). Boston, MA: Allyn & Bacon.

Rankin, J. G. (2016). *Standards for reporting data to educators: What educational leaders should know and demand*. New York, NY: Routledge.

Satherley, P. (2006). *Student outcome overview 2001–2005: Research findings on student achievement in reading, writing and mathematics in New Zealand schools*. Wellington, NZ: Ministry of Education, Research Division.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation & education: At quarter century* (Vol. 90, Part II, pp. 19–64). Chicago, IL: NSSE.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066–1101). Washington, DC: American Educational Research Association.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: Praeger.

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

Swaffield, S. (2011). Getting to the heart of authentic Assessment for Learning. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 433–449. doi:10.1080/0969594X.2011.582838

Torrance, H. (1986). Expanding school-based assessment: Issues, problems and future possibilities. *Research Papers in Education*, *1*(1), 48–59. doi:10.1080/0267152860010104

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretations of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation*, *39*, 144–152.

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation*, *43*, 24–39. doi:10.1016/j.stueduc.2014.04.004

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York, NY: Copernicus Books.

*Ward, L., Hattie, J. A., & Brown, G. T. (2003, June). *The evaluation of asTTle in schools: The power of professional development* (asTTle Tech. Rep. #35). Auckland, NZ: University of Auckland/Ministry of Education.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Zapata-Rivera, D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (Research Memorandum 10–01). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229. doi:10.1080/10627197.2016.1202110

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463. doi:10.1080/0969594X.2014.936357

Zenisky, A., & Hambleton, R. K. (2015). Good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 585–602). New York, NY: Routledge.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. Educational Measurement: Issues and Practice, *31*(2), 21–26.

Zumbo, B. D. (2009). Validity as a contextualised and pragmatic explanation, and its implications for validation practice. In R. W. Lizzitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: IAP-Information Age Publishing, Inc.