The psychology of evaluation: Addressing the pressures teachers face

Gavin T. L. Brown[1]

Abstract

Evaluation processes are used to formatively inform improved instruction and summatively to determine the worth or value of teaching, schools, and students. These uses create pressures for teachers because they have conflicting goals around improvement vs. accountability. This perspective paper overviews the author's understanding of evaluation, the challenges accountability presents to the conceptions teachers have about assessment and identifies a major obstacle to the impact of evaluation in the low information level of most so-called educational tests. A solution to the dilemma is recommended in the example of the Assessment Tools for Teaching and Learning software deployed in New Zealand schools.

Keywords: Teacher psychology; Educational testing; Design of test reports.

Resumo

A psicologia da avaliação: abordando as pressões que os professores enfrentam

Os processos de avaliação são usados para informar formalmente melhorias e, sumariamente, para determinar o valor do ensino, das escolas e dos alunos. Esses usos criam pressões sobre os professores porque eles têm objetivos conflitantes em relação à melhoria da aprendizagem dos estudantes e responsabilização dos docentes. Este artigo apresenta uma visão geral da compreensão do autor sobre a avaliação, os desafios que a responsabilização apresenta às concepções que os professores têm sobre a avaliação e identifica como seu maior obstáculo a reduzida informação sobre as avaliações educacionais. Uma solução para este dilema é apresentada ao descrever o software Assessment Tools for Teaching and Learning (em português, Ferramentas de Avaliação para Ensino e Aprendizagem), utilizado em escolas da Nova Zelândia.

Palavras-chave: Psicologia dos professores; testes educacionais; elaboração de relatórios educacionais.

Resumen

La psicología de la evaluación: abordando las presiones que enfrentan los docentes

Los procesos de evaluación se utilizan para informar formalmente las mejoras y para determinar el valor de la enseñanza, las escuelas y los estudiantes. Estos usos crean presiones sobre los maestros porque tienen objetivos contradictorios con respecto a mejorar el aprendizaje de los estudiantes y la responsabilidad de los maestros. Este artículo presenta una descripción del autor sobre la evaluación, los desafíos de la rendición de cuentas para las concepciones de evaluación de los docentes

[1] The University of Auckland

e identifica como obstaculo a las evaluaciones educativas su información reducida. Se presenta una solución a este dilema describiendo el software Assessment Tools for Teaching and Learning que se utiliza en las escuelas de Nueva Zelanda.
Palabras clave: Psicología de la educación; evaluaciones educativas; elaboración de informes educativos.

Most societies use student testing for selection of students for privileges such as graduation and scholarship. In the last half century, many societies use student test responses to evaluate school and teacher quality, often with severe negative consequences attached. At the same time, teachers are expected to monitor student learning and the efficacy of their teaching by assessing student progress and performance long before substantial consequences are attached to those results for either the student or the instructor. This tension in the purposes of assessment or evaluation create substantial challenges or pressures for teachers. In this perspective paper, I want to first describe what evaluation in education involves and how it relates to assessment and testing. I want to clarify the pressures evaluative assessment generates for educators. My analysis then leads me to focus on the inadequacies of tests used in education in terms of what they cannot tell instructors about what they need so as to do their job better. Consequently, I give an overview of an assessment system used in New Zealand that shows how curriculum-aligned, standardized, computer-assisted testing can help teachers fulfill the formative function of assessment while satisfying the accountability concerns of administrators. In the model I describe teachers are treated as professionals who need better tools.

Understanding evaluation

The term "assessment" has traditionally been situated in the Educational Research Information Center – ERIC (2001) thesaurus as "evaluation". It is this term that it is used predominantly around the world, outside the English-speaking world, to refer to educational assessment. Thus, the idea of assessment is inside evaluation. Hence, the term "evaluation" refers to all and any means of gathering information about performance or needs/strengths and the subsequent interpretations or uses to which that data are put (Brown, 2018).

Evaluative methods include school or classroom-based events and formal processes such as tests, exams, (e-)portfolios, observations, performances, peer or self-assessment, and so on. The term also covers formal public examinations and

qualifications assessments used to determine entry into competitive programs or levels of schooling and for the award of important societal awards such as graduation.

Evaluation both describes the qualities (i.e., strengths/weaknesses) of work products or processes but also makes judgments about the value, merit, or worth (e.g., does it meet expectations, should it be a pass or fail, or is it excellent vs satisfactory) of those same products or processes. Evaluations are needed to inform decisions about whether a product or process is good enough or not and if it needs changes, and what those needs are (Brown, 2018). This leads us to the general question "what is the purpose of assessment"?

A technological framing (Cheung, 2000) derived from how evaluation is used in engineering provides useful insights to what is needed in education. When devising a crewed flight to the moon, engineers need to know that their goal is to get humans to the moon and bring them back safely (i.e., the Apollo project of the 1960s). All the assessments had to be aligned across the various stages of the process to this goal for them to be valid. When the "test" is applied it needs to provide information as to whether work done to date is "on course" (i.e., Are we likely to reach the moon based on where we are now?). Whatever the answer, assessments need to also provide information as to what should be done to either get back on course or stay on course. As the capsule travels to the moon decisions are made based on readings leading to decisions as to whether a long or short fuel burn is necessary and in which direction that burn should be.

Furthermore, engineers had to know in advance what would happen with each action. The most likely result of any course of action based on technologically-designed assessments is known—it's not a "black" box of unknowns. Engineers can be confident that their actions will produce predictable results. Naturally, this is much easier with astrophysics than teaching; but teachers need assessment to fulfill similar qualities. They need to know their curricular goals; they need assessments that tell them where classes and individuals are; and they need robust knowledge about what they should do in response to assessment results and what is likely to happen if they take a certain course of action.

Formative and summative

As Scriven (1967) made clear there are two key timepoints at which evaluations can be implemented. Any time before the end evaluations that could inform

decisions about potential changes to the process to improve outcomes are called "formative" evaluation. The key idea here is to check if changes are needed before it is too late. That meant formative assessments collected data about student learning to inform teacher planning and instructional activities to ensure students mastered what they were expected to learn, understand, or do (Bloom, Hastings, & Madaus, 1971). To correctly inform instructors, administrators, and students about what they needed to change, the quality of formative assessments must be high, otherwise decisions that are meant to support improvement will be wrong (Hattie & Brown, 2010).

The second timepoint is at the end, called "summative", in that it sums up both a description and a quality appraisal of the value of whatever was being evaluated, with no further opportunity to change things. It is here that the word evaluation reveals the key notion of "value", implying worth or merit. It is most often that at this point overall quality or value appraisal takes place (i.e., *taking all things into account this student's work is satisfactory but not good so we award grade C*). Again, because the consequence of a summative evaluation is real, the quality of assessments used to make summative decisions must be high, otherwise evaluative decisions will be wrong.

Of course, formative assessments should be so well done that there will be few surprises when a summative decision is made. Furthermore, as Stake (Scriven, 1991) made clear, summative evaluations can become themselves formative if test-users choose to exploit that information. There are countries which have used their *Programme for International Student Assessment* (PISA) results, in conjunction with other factors, to make substantial and beneficial changes to policy and practice (Teltemann & Klieme, 2016). Hence, the fundamental difference between formative and summative evaluations, according to Scriven, reflects a difference in timing, rather than format or quality characteristics.

This stands in contrast to the view Sadler (1989) created, in which he claimed that there were different formats between formative and summative evaluations. In associating summative purposes with formal tests and examinations, Sadler correctly sought to remove high-stakes interpretations from the process of improving the quality of student learning or teacher instruction. But in doing so, he also removed formal testing from the toolbox of formative assessment. This trend continued into the assessment *for* learning (Black & Wiliam, 1998) movement in which formative assessment became a set of informal interactions between learner and instructor and simultaneously among learners (Leahy, Lyon, Thompson, & Wiliam, 2005). This

pedagogical view of assessment (Stobart, 2006) focused on what learners did within assessment, greatly reducing the role the teacher plays in evaluation. Much more recently, especially in higher education (Yan & Boud, 2022), where learners are expected as adults to be responsible for their own outcomes, theorists have positioned assessment *as* learning; a stance that seems inseparable from self-regulation of learning (Brown, 2022). Interested readers may wish to consult Brown (2021) for an integrated view of how these different approaches to assessment play out in teacher education.

Multiple purposes of assessment

A view of assessment or evaluation promulgated in New Zealand (Brown, 2004; 2006a; 2008) can be summed up in four purposes. Succinctly, assessment can be formative but needs to provide distinct kinds of information to the two parties of education (i.e., instructor and learner). Students need information about where they are, what they are doing well or poorly on, and where they need to focus their own study. Instructors, on the other hand, need to know, not just where individuals are but also where whole classes are, what the priority needs are, what students have mastered and no longer need to be taught, and who in a group has similar needs. Of course, this improvement-oriented use of educational assessment depends on the assessments being both valid and accurate indicators of student knowledge, skill, and development.

Parallel to this formative use, lie summative evaluations to be drawn from assessment that have different applications for students and instructors. Classifying where students are accurately (e.g., low, medium, high) allows validity in consequential award decisions, such as who graduates, who gets a scholarship, or who needs to repeat a year of schooling. Similarly, student performance on summative evaluations can be used to infer the quality of schools and teachers. Indeed, in many jurisdictions the quality of instruction is evaluated by testing the learners. Consider how governments and media treat results from the PISA or Trends in International Mathematics and Science Study (TIMSS) international large-scale assessment systems as a way of evaluating the quality of the country's entire educational system. Similarly, overall quality of curricular efforts can be evaluated in monitoring systems (e.g., the United States National Assessment of Educational Progress – NAEP or New Zealand's National Monitoring Study of Student Achievement – NMSSA) that use samples of students to infer what is working or not. Other jurisdictions implement systematic census testing of all students (e.g., UK's Key Stage testing) to determine if schools

are ensuring all students meet curricular expectations. Hence, through summative assessments, both students and schools/teachers can be held to account for the resources invested in them and for their use with those opportunities.

Largely in response to the summative uses of assessment, many educators have a perspective that I have called "irrelevance". They see assessments as being bad for learners (e.g., labelling them as failures), unfair to those from deprived backgrounds who find mastering the curriculum at the recommended pace difficult, and unnecessary. Such teachers perceive the formal acts and uses of evaluation as something they can ignore because they can tell already what students know, can do, and need. They might implement formal assessments because it is a requirement of their job, but they do not put stock in such acts, because they already know about and care for their learners.

It is worth noting that Remesal (2011) also captures this tension between a focus on teachers vs. students and between formative vs summative purposes. Consequently, evaluation policy and practice must resolve tensions between formative and summative uses, whether the focus is on judging learners or teachers, and whether the users of information are students or teachers. Regardless of these policy decisions, it is important to consider how evaluative decisions impact participants. It is not surprising if the key participant is the teacher who is responsible for leading and managing the classroom's activities and progress through the curriculum.

The challenge of accountability

Unfortunately, for those who dislike assigning merit, it is an inconvenient truth that as a society we need assessments (both descriptive and evaluative). Assessments need to be robust by measuring what we really value, instead of the easy to test material. For example, the quality of a composition is not really determined by the accuracy of spelling, punctuation, or grammar even though those are the easiest to identify. Robust assessments measure the skill, knowledge, or ability so well that we can have confidence that the consequential decisions about who should be rewarded can be defended. This is the core notion of validity depending on reliability (Cizek, 2020).

This means that robust assessments get it right more than they get it wrong. The cost of passing those who did not really master the expected material is to the

detriment of social good. That means we need robust assessments to assure society that those who did sufficiently well can be given privileges (e.g., driver's license, certification as an engineer, doctor, accountant, etc.) while those who do not are prevented from exercising the rights and privileges pertaining to success. As Lingard and Lewis (2016, p. 400) point out:

> Outright opposition to accountability is NOT a justifiable position. Rather, we need a progressive reconceptualization of accountability that acknowledges the broader societal purposes for schooling, and which holds systems, schools, and communities to account, in both bottom-up and top-down ways [capitalisation added for emphasis].

Applying this approach to educational assessment, Resnick and Resnick (1989) expounded the "What you test is what you get" (WYTIWYG) framework. In WYTIWYG, the assumption is made that if something is valuable, it should be assessed, otherwise it will be seen as optional and may end up not being taught or learned. If it is not in a formal assessment, society cannot be sure key knowledge or skills are being delivered. They further argued that without consequences to such tests, the importance of that knowledge can be discounted. The consequences usually relied upon included changing schools, teachers, or leaders (e.g., replace, make redundant, or close). While intuitively appealing, the implementation of this logic in state accountability testing in the United States has revealed undesirable outcomes. Nichols and Harris' (2016) review showed that testing students to judge teacher quality is a counter-productive idea. Generally, this approach to accountability leads to distortion of teaching, curriculum, and teacher professionalism, oppresses minority students, bores highly able students, and encourages cheating on those accountability tests.

This outcome is completely in line with Lerner and Tetlock's (1999) review of the known effects of accountability. The first effect is that subordinates ensure the outcomes specified by superiors are delivered. When superiors want higher test scores and better pass rates, teachers do their best to achieve that. Unfortunately, this may lead to "justified" unethical or unprofessional behavior by teachers and schools to manage how superiors perceive their work. For example, more than 30 years ago, Cannell (1989) documented above-average performance on standardised tests in each state of the United States for a number of reasons including: (1) teachers taught students directly what was going to be on the test, (2) schools used old tests that had invalid comparison norms, (3) teachers taught items in class based on an early copy of the official test, (4) schools encouraged low performing students to NOT come to

school on testing days, (5) teachers gave students hints during testing, and (6) schools corrected student test responses before sending the tests to a central agency. This latter process resulted in teachers and school leaders being jailed in Atlanta, USA for falsifying student performances on school accountability tests[2].

Strong negative consequences for low scores seem to be seen by many teachers as fundamentally unfair, in part because the school system is usually not fair. For example, some schools have high proportions of children whose parents are poor or uneducated. Other schools are poor, lacking essential resources or are in locations where it is difficult to attract highly skilled teachers. Psychologically, most teachers object to using standardized tests to label students as failures or to determine a child's worth. With these beliefs it seems that "cheating" to raise test scores may be a highly rational, albeit unprofessional, response. It seems highly likely under a regime of accountability, which ignores the many complicating factors that limit student test performance, resistance and subversion of the regime will occur and be potentially defensible.

This means that knowing how well a school system is doing by testing students to judge teachers and schools is a highly fraught exercise. As Bonner (2016) elegantly points out the desire teachers have to do a good job and see students learn is pressed by external mandates, policies, and resource levels. So, before we blame teachers for wanting to hide from the "bad" news that their teaching does not always work, we need to examine how teachers conceive of and use assessment. It may be that teachers are not to blame; rather it could be that the nature and quality of assessment resources provided to teachers do not support a diagnostic, formative goal (Brown & Hattie, 2012).

Teachers conceptions of assessment

Based on a four-concept analysis of assessment purposes, the Teachers Conceptions of Assessment inventory (Brown, 2006b) was developed in New Zealand with primary school teachers. Because New Zealand had a strongly child-centered primary school system without any formal tests or examinations, it was not surprising to note that teachers were strongly in favor of assessment for improvement,

[2] https://en.wikipedia.org/wiki/Atlanta_Public_Schools_cheating_scandal.

rejected the idea that assessment was irrelevant, cautiously agreed that assessment could evaluate students, and rejected the idea that assessment could identify school quality (Brown, 2004).

The conception that assessment is for improvement has been found to elicit the strongest endorsement in many contexts (i.e., jurisdiction and level of schooling). Primary and secondary teachers in New Zealand (Brown, 2011), Hong Kong (Brown, Kennedy, Fok, Chan, & Yu, 2009), Queensland (Brown, Lake, & Matters, 2011), Cyprus (Brown & Michaelides, 2011), Ecuador (Brown & Remesal, 2017), and Egypt (Gebril & Brown, 2014) all endorsed assessment for improvement the most. Across those same jurisdictions teachers rejected most strongly either assessment for school accountability (i.e., New Zealand, Egypt, and Queensland) or assessment is irrelevant (i.e., Hong Kong, Ecuador, and Cyprus). Teachers in India gave highest endorsement to improvement and diagnostic assessment (Brown, Chaudhry, & Dhamija, 2015). It is important to note that among teachers the priority of assessment for improvement is noticeable regardless of culture or consequence culture attached to assessment.

It also turned out that different conceptions of assessment influenced the type of assessment practices teachers used (Brown, 2009). The choice to use informal assessment formats, as opposed to formal testing controlled by the teacher was influenced almost equally by the inversely correlated conceptions that assessment was irrelevant and assessment was for improvement. This result suggested teachers treated formal assessments as irrelevant and considered that informal assessments gave dependable information about learning. At the same time, teacher conceptions that assessment was about making students accountable predicted use of formal assessments that tested surface level knowledge. In contrast, deep level knowledge was predicted by the conception that assessment showed the quality of schools. Together these indicate New Zealand teachers thought externally developed tests for students measured gaining knowledge, while high quality schools delivered deep, relational, critical thinking.

Overall, teachers believe assessment exists to support improved teaching and learning. They reject the idea that it is irrelevant. They are aware that it evaluates students and have mixed feelings about that. It seems that teacher beliefs are not the problem if students are not doing well on formal, standardized qualifications assessments and examinations. Recommendations to improve teachers' understanding

through professional development are easy to make, but perhaps the challenge is that the assessment and testing system stands in the way.

So, what's wrong in evaluation?

The most widespread source of information for evaluation consists of tests and examinations. These have the virtue of being the same across all test-takers and require simple assumptions. First, that test scores are "uncontaminated measures, assuming that a student's responses to test items reflect only his or her knowledge, skills, and abilities relevant to the target measurement domain" (Wise & Smith, 2016, p. 207). Second, as Thorndike (1924) put it "all our measurements assume that the individual in question tries as hard as he [sic] can to make as high a score as possible" (p. 228). Hence, any glitches or momentary lapses in attention, motivation, or self- thoughts, are just part of the random background of "error". From this, parents presume that test scores tell the truth of how what students know and how hard they have worked before and during a test (Harris & Brown, 2016). It should be obvious that these assumptions are simply wrong.

Tests and exams rarely go beyond giving a total score, most often as a percentage correct. Once we have a score, it is easy to create a rank-order score such as position in class (e.g., first or last). Standardized tests tend to report rank order position relative to a norming sample using methods such as percentile or stanine. Standardized tests often report performance with a transformed score (e.g., the IQ scale has Mean = 100 and Standard Deviation = 15). These scores do not give strong diagnostic or formative information, even if they are used early in the school year or during course of instruction.

One reason they lack this power comes from the problem of sum of items correct scoring methods. This classical test theory approach gives every item equal weight. In addition to the simplicity of adding up the number right and dividing by the total number of marks, there can be reasons for tests to produce inflated scores. Tests with many easy tasks or questions produce high average scores with many students getting above an important cut-scores (e.g., pass-fail). This can make the teacher look good and lead students to believing they are doing well on the expected work. Unless scores can be adjusted by item difficulty as they are in item response theory, teachers and students will be seduced into complacency if high scores are

obtained from tests full of easy questions. Only when tests have challenging tasks is it possible to know who has mastered only the easy stuff or who can do the hard stuff; consequently, giving scores weighted by item difficulty means it is possible to identify easy stuff that just needs practice versus hard stuff that needs instruction and where students are in relation to difficulty.

Rank order scores are easy to determine and understand in a classroom setting. Only three can stand on the podium as winners of the test race. However, order does not tell the teacher, student, or the parent what each learner needs to learn next. An unfortunate side-effect of rank order is that it leads to the assumption that those at the bottom cannot learn while that those at the top cannot be taught more. Both are false conclusions and anti-educational. Related to this, teachers tend to normalize on their own populations, assuming that the best and worst performer in their own group is equal to the truly best or worst in the nation. This is where standardized tests can help by comparing students to a nationally representative previously tested population. That way we can say things like "this boy is worst in the class, but he is actually performing equal to the top third of the nation. His problems are not severe".

Another reason total score is educationally ineffective is that diagnosis comes from knowing the profile of performance across the teachable sub-scores or components of the curriculum that teachers need to teach. No teacher teaches just a subject; they plan lessons for sequentially ordered material that builds on previously taught material. That means when assessing our instruction, we need to know which parts of the curriculum we can stop teaching because students have got it and which parts need new instruction or possibly return for revision. Unless test tasks and questions are mapped to the curriculum, its objectives, strands, or components that teachers must teach, they will only reinforce what teachers already know. Teachers tend to know very quickly who is best and worst in a class, but they do not know "who needs to be taught what". That is where well-designed tests mapped to important teaching objectives come to the rescue (Brown & Hattie, 2012).

A good educational system provides to its teachers tools they can use that provide educationally useful information. Diagnostic, standardized tests can tell teachers what they need to know, such as:

- which objectives students have mastered and do not need to be taught anymore;

- which objectives they did well on and for which they can be praised;
- which objectives they need to revise, review, and practice; and,
- which objectives students cannot yet do and for which they need instruction.

However, designing and analyzing tests in this way in a timely fashion requires skills and resources that most teachers do not have. Despite the many textbooks and courses that teach test development and analysis, experience shows us that most teachers lack the skills to write high quality tests. They also do not have the time or priority to crunch the numbers related to making sense of test scores. While it may be tempting to disregard formal testing as a tool that teachers can use, no other mechanism allows teachers to find out so efficiently who knows what and who needs to be taught what. Also, as research into teachers' beliefs has shown teachers are not really the problem. The real problem seems to lie in poorly designed educational tests.

The New Zealand response: asTTle

To address this challenge a technological evaluation system was developed 20 years ago in New Zealand and is still being operated by schools. The Assessment Tools for Teaching and Learning (e-asTTle[3]) system provides standardized testing for reading comprehension, mathematics, and writing in two languages (i.e., English and te reo Māori, the indigenous language of Aotearoa). With substantial funding from the government, we developed at the University of Auckland a computer-assisted, school-controlled testing platform that gradually evolved from stand-alone computers to local area networks to, now, a fully functional web application with computer adaptive testing functionality.

Curriculum aligned

What makes the system powerful is not its computing or statistical technologies (Hattie & Brown, 2008). Rather, the core principles of responsible educational testing were deployed in the design from the beginning. These include linking every task, question, and report to the curriculum's objectives, strands, and progression

[3] https://e-asttle.tki.org.nz/.

levels. The system gives teachers and schools choice over when they test, who they test, what content is in the test, and how they use the data. To achieve this, a wizard system allows teachers to specify what curriculum areas and difficulty levels should be contained in a test created from an item bank of pre-calibrated test questions. Teachers get to see what the test will look like before administering to ensure that it fits their intentions and recent classroom teaching. If the test creation engine does not produce what the teacher had in mind, the test can be altered by changing content and difficulty settings.

All the performance information is clearly communicated after extensive psychometric analyses using item response theory to give weight to success on more demanding tasks and questions. Norms derived from 30,000 students in reading, 25,000 in mathematics, and 20,000 in writing are provided for school years 4–12 and for curriculum levels 2–6. The difficulty of an item was derived from the norming information using item response theory (IRT) methods (Embretson & Reise, 2000). Items that few students got right have higher difficulty levels than those that most students got right. But instead of just giving a number, performance scores were mapped to curriculum levels using panels of experienced teachers. That way a teacher knows where in the curriculum progression a child or class is and whether they are ahead of or behind relevant year norms. This gives the ability to see who is where and where they are compared to others; an especially useful set of insights to share with parents and families. Users can have confidence in the alignment to both curriculum standards and achievement norms used to create reports for their own students.

To support the challenge of assessing the curriculum, the asTTle system does the donkeywork of test design, data entry, item scoring, and aggregation into meaningful curriculum insights, leaving the teacher free to think about how to design and deliver appropriate instruction instead of trying to figure out "who got what on the test". To assure schools that these assessments were truly for diagnostic, formative purposes, schools keep control over the system and their data. That meant no central authority collected the data, could examine it, or say who had to be tested at any time. This meant that schools and teachers could see if there were any "bad news" stories in their own data and respond to it well before any external evaluators arrived on the scene (Hattie & Brown, 2008).

Data visualization

The crowning glory of asTTle lies in our efforts to communicate more powerfully to teachers the information they need to make sound and valid educational decisions about their teaching. A review of the literature made clear that most data visualizations around assessment are poorly understood by their intended readers (Brown, 2001). Our search for good examples encountered the CRESST Quality School Portfolio (Baker, 1999) which used a dashboard reporting system exploiting the power of traffic signals (i.e., red = warning, below average; yellow = average; green = good, above average) to communicate information to help teachers understand the data that schools have. Insights were collected from teachers concerning their information needs and how testing could meet them (Meagher-Lundberg, 2000). With the assistance of a graphic artist a series of mock-ups were piloted with teachers for whom the system was intended (Meagher-Lundberg, 2001a; 2001b), resulting in designs that have been used since the beginning (Hattie, Brown, & Keegan, 2003).

The system generates a selectable menu of reports that allow teachers and school leaders to evaluate, respond to, and monitor the effects of their work. Each report is designed to meet the information needs of a specific role in schooling (i.e., classroom teacher, department manager, school leader) and thus the information is displayed in a way designed for that use of the assessment results. Specifics of the reports and their underlying logic is given in Brown, O'Leary, and Hattie (2018). The reports visualize the strengths and weaknesses of student performance at both individual and group levels, such that they can be used by teachers to plan instruction, report to and with parents, and give feedback to students (Archer & Brown, 2013).

The overall performance of a student was determined by the IRT weight of items answered correctly without any penalty for guessing or error. This meant that student performance could be classified by correctness and difficulty of items answered. Following the Wright Map display method (Wright & Stone, 1979), the Individual Learning Pathways (ILP) report positioned items and their achievement objective in one of four quadrants on a chart relative to the individual student's overall performance. Note item numbers are displayed in brackets beside the achievement objective, to prioritize a focus on curriculum goals rather than test questions. The four quadrants also lead to appropriate teaching and feedback messages, as given in the table below. This report has been successfully used in schools as the basis of parent-teacher discussions about how a child is going and what will happen next.

Table. Individual learning pathways interpretive guide.

Color	Content	Interpretation
Green	Correct but easy	This material is mastered, so the teacher should not teach it anymore.
Yellow	Correct but hard	Congratulations! This shows what you can really achieve. The teacher should ensure more work of this level and type is done.
Red	Incorrect but easy	The student should be encouraged to do some practice, the teacher should check that it has been taught, but it should not be drilled because it will be easy for the student to correct.
Blue	Incorrect but hard	The teacher should plan to teach this material because it was answered incorrectly, suggesting it requires knowledge or skill that have not yet been taught. It is unlikely to be learned by osmosis or discovery methods.

Unlike the medical profession which treats one patient at a time, the classroom teacher deals with groups of 20 to 40 students in New Zealand, while in other countries that group could be much larger. Additionally, in many school systems, middle managers of year groups or teaching subjects want to see what the pattern of strengths and weaknesses is across a whole cohort. Hence, we developed an aggregate report that identifies proportion of the group for each achievement objective using the same coded color scheme as that used in the ILP. For planning purposes, this report easily shows the size of the Blue zone across the group against the number of items for each objective. Interpretation is similar to the ILP; teachers need to plan to teach the Blue zone by halting teaching in the Green zone. Anecdotal reports indicate school departments that do this see robust growth in student achievement; in other words, students learn what we teach them, and assessment can show that.

In New Zealand at least, within class grouping is a common practice (Wilkinson & Townsend, 2000), as well as making extensive use of peer tutoring (Roscoe & Chi, 2007). To serve those goals, a distributional chart was developed in which the proportion of the group in each relevant curriculum level was displayed, giving the teacher a sense of how disparate the performance levels are. Upon clicking the chart, a list of names is generated in each curriculum sub-level, allowing teachers the ability to see who could be grouped together for differentiated instruction (Moon, 2016) or who could be partnered for peer support.

Instead of just giving comparison to national norms, school leaders can select comparison groups for schools with similar demographic characteristic (i.e., Schools Like Mine; Hattie, 2002). In this way, a more credible comparison is made; however, with the caveat that if other schools like mine are doing better on average than my school, I can no longer say it is because of my community. In this way, schools are challenged to stop blaming external factors and instead consider what other schools are doing that might help their own school improve.

Acceptability

To protect the formative potential of asTTle, it was made clear throughout its development, to both the Ministry of Education and to the school community, that the system contained both the questions and answers. That meant it would be possible to "cheat" to make a class or school look good. However, there would be no educational value in doing this because it is only when educators look at what students fail to know or be able to do, that teachers can begin to address those challenges.

Other mechanisms that supported acceptability include incremental development and transparency. The asTTle system developed through multiple technological variations in accordance with the school infrastructure system to now be delivered online as a web application (Brown, 2019). The asTTle system acknowledged the contribution of the 1,000s of teachers who were involved as panelists, item triallists, and reviewers; a step that helped gain acceptability across the nation. Furthermore, technical reports were posted on a website[4] giving unrestricted access to what the asTTle project had done, what we had learned, and why we had made certain design choices. This transparency helped teachers accept the value of the system as a formative tool.

Conclusion

In this approach to formative evaluation, tests are not the problem. Instead, poorly designed tests with limited reporting are the problem. The system in New Zealand makes it possible to create an educationally effective testing system that helps

[4] https://e-asttle.tki.org.nz/Reports-and-research/asTTle-technical-reports.

schools and teachers do their job, but also meet accountability expectations. This can be done because teachers and their attitudes are not at the core of the problem. Teachers want to teach better but are often left adrift with evaluation systems that are so limited they cannot help teachers identify who needs to be taught what next. Taking advantage of the positive improvement-oriented beliefs of teachers, rather than treating them as the problem is a major step forward. A possible slogan is "if you respect teachers and want to improve outcomes, give them a tool like asTTle". This is valid because, rather than being about technology or summative scoring, the system is designed as an educational technology that helps teachers do what teachers need to do.

Any society that wants to improve student learning needs to create an evaluation system that goes beyond the status quo. The system must focus on telling teachers what teachers need to improve their instruction and classroom activities. Assessments need to be tightly integrated with the curriculum so that they can serve teaching goals. The asTTle example in New Zealand is a successful example of using evaluative practices embedded in technology to improve teaching.

References

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. https://doi.org/10.1080/16823206.2013.773932

Baker, E. L. (1999, July). Technology: something's coming-something good. *CRESST Policy Brief*, 2. Retrieved on April 14, 2023 from https://cresst.org/publications/cresst-publication-3245/

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

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

Brown, G. T. (2019). Technologies and infrastructure: costs and obstacles in developing large-scale computer-based testing. *Education Inquiry*, *10*(1), 4-20. https://doi.org/10.1080/20004508.2018.1529528

Brown, G. T. L. (2001, Aug.). Reporting assessment information to teachers: report of project asTTle outputs design. *Technical Report 15: Reporting Assessment Information.* Retrieved on April 14, 2023 from http://tinyurl.com/q69ew9a

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

Brown, G. T. L. (2006a). Teachers' conceptions of assessment: validation of an abridged version. *Psychological Reports*, *99*(1), 166-170. https://doi.org/10.2466/pr0.99.1.166-170

Brown, G. T. L. (2006b). *Teachers' conceptions of assessment inventory-Abridged (TCoA-IIIA-version 3-abridged).* Auckland: University of Auckland. https://doi.org/10.17608/k6.auckland.3199543.v1

Brown, G. T. L. (2008). *Conceptions of assessment: understanding what assessment means to teachers and students.* New York: Nova Science.

Brown, G. T. L. (2009). Teachers' self-reported assessment practices and conceptions: using structural equation modelling to examine measurement and structural models. In T. Teo, & M. S. Khine (Eds.), *Structural equation modelling in educational research: concepts and applications* (pp. 243-266). Rotterdam: Sense.

Brown, G. T. L. (2011). Teachers' conceptions of assessment: comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45-70. https://doi.org/10.18296/am.0097

Brown, G. T. L. (2018). *Assessment of student achievement.* London: Routledge.

Brown, G. T. L. (2021). Evaluation in teacher training: an analysis of theoretical conceptions. In W. Santos, & R. Stieg (Eds.), *Evaluación educativa en la formación de profesores: Brasil, Colombia, Chile, España, Inglaterra, México, Nueva Zelanda y Uruguay* (pp. 53-67). Curitiba: Appris.

Brown, G. T. L. (2022). Assessments cause and contribute to learning: if only we let them. In Z. Yan, & L. Yan (Eds.), *Assessment as learning: maximising opportunities for student learning and achievement* (pp. 38-52). London: Routledge.

Brown, G. T. L., & Michaelides, M. P. (2011). Ecological rationality in teachers' conceptions of assessment across samples from Cyprus and New Zealand. *European Journal of Psychology of Education*, *26*(3), 319-337. https://doi.org/10.1007/s10212-010-0052-3

Brown, G. T. L., & Remesal, A. (2017). Teachers' conceptions of assessment: comparing two inventories with Ecuadorian teachers. *Studies in Educational Evaluation*, *55*, 68-74. https://doi.org/10.1016/j.stueduc.2017.07.003

Brown, G. T. L., Chaudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: a quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research*, *71*, 50-64. https://doi.org/10.1016/j.ijer.2015.03.001

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 347-363. https://doi.org/10.1080/09695940903319737

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: the impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, *27*(1), 210-220. https://doi.org/10.1016/j.tate.2010.08.003

Brown, G. T. L., O'Leary, T. M., & Hattie, J. A. C. (2018). Effective reporting for formative assessment: the asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting: research and applications* (pp. 107-125). London: Routledge.

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

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests.* Wheaton: Friends for Education.

Cheung, D. (2000). Measuring teachers' meta-orientations to curriculum: application of hierarchical confirmatory analysis. *Journal of Experimental Education*, *68*(2), 149-165. https://doi.org/10.1080/00220970009598500

Cizek, G. J. (2020). *Validity: an integrated approach to test score meaning and use.* London: Routledge.

Educational Research Information Center – ERIC. (2001). *Thesaurus of ERIC descriptors* (14th ed.). Phoenix: Oryx.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Gebril, A., & Brown, G. T. L. (2014). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice, 21*(1), 16-33. https://doi.org/10.1080/0969594X.2013.831030

Harris, L. R., & Brown, G. T. L. (2016). Assessment and parents. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1-6). Berlin: Springer.

Hattie, J. A. (2002). Schools like mine: cluster analysis of New Zealand schools. *Technical Report 14, Project asTTle*. Retrieved on April 14, 2023 from https://e-asttle.tki.org.nz/content/download/1470/5943/version/1/file/14.+Cluster+analysis+of+NZ+schools+2002.pdf

Hattie, J. A. C., 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. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189-201. https://doi.org/10.2190/ET.36.2.g

Hattie, J. A., & Brown, G. T. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: concepts, research and challenges* (pp. 116-131). London: Routledge.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255-275. https://doi.org/10.1037/0033-2909.125.2.255ra

Lingard, B., & Lewis, S. (2016). Globalization of the Anglo-American approach to top-down, test-based educational accountability. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 387-403). London: Routledge.

Meagher-Lundberg, P. (2000, Oct.). Comparison variables useful to teachers in analysing assessment results. *Technical Report 1: Project asTTle*. Retrieved on April 14, 2023 from https://e-asttle.tki.org.nz/content/download/1454/5895/version/1/file/1.+Comparison+Variables+2000.pdf

Meagher-Lundberg, P. (2001a). Output reporting design: focus group 1. *Technical Report 9: Project asTTle*. Retrieved on April 14, 2023 from https://e-asttle.tki.org.nz/content/download/1464/5925/version/1/file/9.+Output+reporting+design+Focus+Group+1+2001.pdf

Meagher-Lundberg, P. (2001b). Output reporting design: focus group 2. *Technical Report 10: Output Focus Group 2*. Retrieved on April 14, 2023 from https://e-asttle.tki.org.nz/content/download/1465/5928/version/1/file/10.+Output+reporting+design+Focus+Group+2+2001.pdf

Moon, T. R. (2016). Differentiated instruction and assessment: an approach to classroom assessment in conditions of student diversity. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 284-301). London: Routledge.

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). London: Routledge.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teaching and Teacher Education, 27*(2), 472-482. https://doi.org/10.1016/j.tate.2010.09.017

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534-574. https://doi.org/10.3102/0034654307309920

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. https://doi.org/10.1007/BF00117714

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Chicago: Rand McNally.

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: National Society for the Study of Education.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). Thousand Oaks: Sage.

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 369-386). London: Routledge.

Wilkinson, I. A. G., & Townsend, M. A. R. (2000). From Rata to Rimu: grouping for instruction in best practice New Zealand classrooms. *The Reading Teacher*, *53*(6), 460-471.

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204-220). London: Routledge.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. San Diego: Mesa.

Yan, Z., & Boud, D. (2022). Conceptualising assessment-as-learning. In Z. Yan, & L. Yang (Eds.), *Assessment as learning: maximising opportunities for student learning and achievement* (pp. 11–24). London: Routledge.

Submitted on: September 2022
Accepted on: October 2022

Sobre o autor

Gavin T L Brown
PhD, Director Quantitative Data Analysis and Research Unit, Faculty of Education & Social Work, The University of Auckland. Associerad Professor, Dept. of Applied Educational Sciences, *Umeå Universitet*, Sweden; Bualuang ASEAN Chair Professor, *Thammasat University*, Thailand; Honorary Professor, Dept. of Curriculum & Instruction, *Education University of Hong Kong Frontiers in Education;* Chief Section Editor: Assessment, Testing, and Applied Measurement.