frontiers in Education

PERSPECTIVE
published: 28 June 2019
doi: 10.3389/feduc.2019.00064

# Is Assessment for Learning Really Assessment?

*Gavin T. L. Brown* [1,2]*

[1] Umeå University, Umeå, Sweden, [2] The University of Auckland, Auckland, New Zealand

This opinion piece questions the legitimacy of treating assessment for learning (A*f*L) as assessment. The distinction between testing and assessment is first made, then the defining characteristics of contemporary A*f*L are identified. While A*f*L claims to be assessment, my analysis argues that A*f*L is a pedagogical curriculum approach that has some process aspects of assessment. However, because of the interactive and in-the-moment characteristics of A*f*L, it fails to meet requirements of an assessment. Specifically, because the in-the-moment and on-the-fly aspects of effective classroom discussions and providing feedback happen in ephemeral contexts it is not possible to scrutinize the interpretations teachers make of student products and processes. Furthermore, we cannot know if those interpretations were sufficiently accurate to guide classroom interactions. Without social or statistical moderation, stakeholders cannot be assured that valid conclusions are reached. Additionally, the scale of error in both teacher and student judgment means that A*f*L practices cannot be relied upon for decision making beyond curriculum-embedded actions within a pedagogical process. Because teaching requires robust evidence to support decisions made about students and teachers, the practices commonly associated with A*f*L cannot provide sufficient evidence on which to base anything more than teaching interactions.

**Keywords: assessment, assessment for learning (A*f*L), error, verifiability, evaluation**

OPEN ACCESS

***Edited by:***
*Sarah M. Bonner,*
*Hunter College (CUNY), United States*

***Reviewed by:***
*Rolf Vegar Olsen,*
*University of Oslo, Norway*
*Peter Ralph Grainger,*
*University of the Sunshine Coast, Australia*

***Correspondence:**
*Gavin T. L. Brown*
*gavin.brown@umu.se;*
*gt.brown@auckland.ac.nz*

***Specialty section:***
*This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education*

***Received:*** *17 May 2019*
***Accepted:*** *17 June 2019*
***Published:*** *28 June 2019*

***Citation:***
*Brown GTL (2019) Is Assessment for Learning Really Assessment? Front. Educ. 4:64. doi: 10.3389/feduc.2019.00064*

## DEFINING ASSESSMENT AND TESTING

Assessment and evaluation are terms that have been bundled for a long time. When I was doing my master's degree in the early 1990s, the ERIC thesaurus placed assessment under evaluation. Indeed, many of my Chinese colleagues as recently as 10 years ago asked why I used assessment instead of evaluation. While many people use the terms comfortably, either interchangeably or distinctively, it will help the reader to know my perspective on these and related terms.

Evaluation, the much older term, has embedded within it the word "value"; hence, the term indicates processes for determining the merit, value, or worth of some product, process, program, personnel, etc. It is easy to see why assessment would fit under evaluation, when the only kinds of assessments were tests and examinations which were used to evaluate the quality of student achievement, rank candidates, and make selection for rewards and further opportunities. Testing and examination is what was meant by the authors and editors of the first *Handbook of Educational Measurement* (Lindquist, 1951).

Markus and Borsboom define testing as "any technique that involves systematically observing and scoring elicited responses of a person or object under some level of standardization" (Markus and Borsboom, 2013, p. 2) and contrast this to assessment which involves a broader set of non-systematic or non-standardized protocols. Thus, testing involves (a) a data collection

mechanism that samples appropriately from a domain of interest, (b) is administered to fairly to appropriate test-takers, (c) is scored according to replicable rules and procedures, and (d) from which inferences can be legitimately drawn about the quality of performance or ability, including identification of weaknesses, needs, or gaps. This is a technological approach to testing in which, like engineering, potential flaws in the process (design to collection to interpretation) are identified (Crooks et al., 1996) and mechanisms put into place to ensure accuracy, consistency, and reliability.

Hence, from my perspective, testing involves description of characteristics of performance, product, or process, which can lead to either diagnostic prescription of subsequent actions or a statement of value, merit, or worth. To achieve this description as a robust basis for subsequent decisions or actions, testing must demonstrate characteristics associated with trustworthiness. There needs to be empirical and theoretical evidence supporting the interpretations and decisions being made from the test (Messick, 1989). That evidence must cover the processes used to ensure the validity of the test design, administration, and interpretation—that is:

- *Can you show that the test itself validly represents the domain?*
- *Was it administered fairly and properly?* and
- *Were appropriate procedures used to evaluate the data arising from the assessment*?

Secondly, the evidence must show that the scoring processes were reliable, accurate, and credible. In other words, questions such as these need to be addressed:

- *Were the right and wrong answers given the right score?*
- *Would another judge give, following the same scoring protocol, the same or nearly the same score?* or
- *Was the scoring free from biases?*

Without evidence that the design, implementation, and scoring were done in a robust way, the testing process fails to meet fundamental requirements, and should not be used as the basis of decision-making. These are the standards and expectations the psychometric industry places on tests (AERA, APA, and NCME, 2014). My view of assessment is, notwithstanding its non-standardized or non-systematic procedures, if it is to be the basis for decisions about students (see Newton, 2007 for 17 different purposes or functions to which assessments can be put), that it needs to be judged against the criteria by which standardized tests are evaluated. Otherwise, assessment practices do not merit the term assessment.

## DEFINING ASSESSMENT *FOR* LEARNING

Assessment *for* Learning (A*f*L), in contrast, seems to be focused on classroom strategies and techniques that are associated with classroom learning (Black and Wiliam, 1998a,b). Their influential work identified two sequenced actions that could make evaluation practices "formative"; that is, learners needed awareness of the gap in current capabilities relative to the desired goal and need to take action to close the gap. Thus, formative assessments techniques were (a) choice of tasks that aligned with goals and had potential to reveal gaps, (b) open-ended teacher-student conversations, (c) use of deep thinking questions, (d) judicious use of testing, (e) the quality of feedback, and (f) involving students in assessment through peer and self-assessment.

As I have seen A*f*L promulgated in teacher education communities and in curriculum policy making in the English speaking world, it would appear that this pedagogical approach to A*f*L has been widely accepted. For example, Shavelson (2008) in the US defined assessment *for* learning this way:

> Teachers … use their knowledge of "the gap" to provide timely feedback to students as to how they might close that gap. … By embedding assessments that elicit students' explanations—formally within a unit, as part of a lesson plan, or as "on-the-fly" teachable moments occur—teachers would take this opportunity to close the gap in student understanding. As a consequence students' learning would be expected to improve (Shavelson, 2008, p. 293).

Here he points to the within classroom instruction context as the place in which assessment for learning takes place. Similarly, the New Zealand Ministry of Education curriculum framework clearly points to the interaction between and among teachers and students, as the place in which A*f*L takes place. Note the explicit attention to the constraint upon analysis and interpretation of data—it takes place instantaneously and in the mind of the teacher:

> Assessment for the purpose of improving student learning is best understood as an ongoing process that arises out of the interaction between teaching and learning. It involves the focused and timely gathering, analysis, interpretation, and use of information that can provide evidence of student progress. Much of this evidence is "of the moment." Analysis and interpretation often take place in the mind of the teacher, who then uses the insights gained to shape their actions as they continue to work with their students (Ministry of Education, 2007, p. 39).

Swaffield (2011), in England, went further in rejecting the notion of agreed criteria or even the central role of the teacher when she argued for a "pure formative" view of A*f*L:

> Assessment for learning when the learner is center stage is as much about fuzzy outcomes, horizons of possibilities, problem-solving, and evaluative objectives, as it is about tightly specified curriculum objectives matched to prescribed standards. It is the (mis)interpretation of A*f*L as a teacher driven mechanism for advancing students up a prescribed ladder of subject attainment that is the problem, not A*f*L itself (Swaffield, 2011, p. 440).

If A*f*L is "integral to teaching and learning and indeed a powerful form of learning itself" (Swaffield, 2011, p. 436), without agreed targets or goals, it seems even more distinct from an evaluative approach to assessment. This may constitute an extreme position that will seem alien to many, but this call for open-ended approaches to assessment in which learning is freed from the ties of standards, outcomes, or teachers has had

considerable influence (her 2001 article has nearly 200 Google Scholar citations at the time of writing), perhaps most notably in teacher education circles.

Stobart (2006) has suggested that the child-centric pedagogical processes of AfL are teaching, especially as understood in Anglo-centric primary school contexts. Stepping back from a "purist" perspective, AfL at least seems to focus on teachers engaging with learners in co-constructing new knowledge (Brookhart, 2016) and in that moment-to-moment process teachers interactively adjust their teaching, prompts, activities, groupings, questions, and feedback in response to the ideas, skills, or knowledge exhibited by the students. This pedagogical application of assessment principles within AfL has led to a reasonably universal set of practices that capture the essence of AfL, summarized succinctly by Leahy et al. (2005). Specifically, AfL seeks to ensure that students learn by (1) involving them in the processes of defining goals, (2) participating in open-ended tasks, (3) evaluating their own and their peers' work, and (4) giving to peers and receiving from peers and teacher feedback intended to improve their learning. Nonetheless, in general AfL practice, the teacher, informed by curriculum, sets the criteria, engineers discussions and activities, and provides feedback. Thus, it can be seen reasonably easily that AfL is a set of pedagogical practices for teaching.

My argument is that these AfL classroom practices are potentially powerful and important teaching techniques; Hattie (2009) has identified that many of these are substantial contributors to increased outcomes. However, I want to draw our attention to research into characteristic implementation of AfL that casts serious doubts as to its ability to provide veridical understanding of student learning. This does not mean I require all "assessment" to be tests; I am very supportive of a wide range of data elicitation methods, such as portfolios, authentic assessments, peer assessment, rubrics for judgments, self-assessments (Brown and Ngan, 2010; Brown et al., 2014; Brown, 2018). The difference is that because these assessment methods are meant to lead to actions and decisions about learning, outside the intuition of the teacher in-the-moment at hand, they need to demonstrate qualities that tests have to meet, albeit the standards may not be as high.

This of course does not mean that AfL as described here is the only version of formative assessment available to education. There are many variations on the theme of using assessment to support improved learning outcomes rather than just as summative judgments about students, teachers, or schools. In New Zealand, for example, standardized, computer-administered tests are used for diagnostic and accountability purposes and do so effectively because they are school-controlled, rather than externally administered, and provide teachers information as to "who needs to be taught what next' (Brown and Hattie, 2012). Other approaches, seen more often in New Zealand secondary schooling (Crooks, 2010), include a broad range of data elicitation techniques (e.g., direct observation of performances, portfolios, long constructed response products) and systematic ways of ensuring validity and reliability of judgments, including use of multiple raters for student work, external validation of ratings, and use of scoring rubrics with specified marking criteria. These systems also involve giving students access to learning intentions and criteria, peer and self-marking against those standards, and teacher feedback all indexed to the assessment criteria; see description of a New Zealand high school English teacher's practices in Harris and Brown (2013). Elsewhere, regular and repeated testing has been used as a way of generating feedback to students about the progress they have made and the needs they have (Roediger et al., 2011). An important key for all formative assessments, including AfL, is that they must have low-stakes consequences (Hattie and Brown, 2008), otherwise all the negative aspects of accountability testing will come to the fore.

Nonetheless, I wish to take issue with AfL as described here because of its popularity in teacher education and its prevalence in classroom assessment, at least in contexts that I have encountered, including NZ, Australia, and Sweden. My sense of AfL, as described here, is that it looks like teaching, not assessment that can reliably be depended upon for decision making. In this perspective piece, I want to discuss further what it is about this version of AfL that concerns me.

## AfL VS. ASSESSMENT

Under Scriven's (1967) evaluation terminology, AfL could be considered synonymous with using assessment formatively. Formative evaluation processes, such as classroom assessment, take place early enough to lead to improved processes and products before it is too late (i.e., the summative evaluation). In comparing the traditional processes of assessment (AERA, APA, and NCME, 2014) with the practices advocated by AfL, it seems there is some overlap with the more formal evaluative process, especially around design, data collection, and consideration of next steps (**Table 1**). However, the gaps between the processes are telling. The range of assessment methods is quite different and deliberate attention to validation of interpretation, communication of results, and validation of consequences is considerably greater in the formal assessment processes. It is as if the quality of student and teacher involvement in AfL need only focus on a different purpose and style of assessment rather than concern itself with the validity of the judgments being made.

The missing aspects, in my view are around the validity and reliability of the interpretations (sometimes grades or scores) teachers make of the performances and products students create in the classroom. Furthermore, while teachers communicate their interpretations to students on-the-spot, there is no guarantee that such feedback is correct or that it is grasped correctly by the students themselves. Furthermore, these practices are silent about longer-term monitoring of the impact of the interpretations and actions teachers take on the assessments they make.

There is no argument that good teaching is responsive to the student and interacts with the student in-the-moment and on-the-fly to make adjustments and give feedback. Good teachers behave this way because they are aware of how their plans and goals interact with student learning and how and when they need to be changed in light of those insights. Good teaching aims for rich conversations with learners in a way that opens learning to

**TABLE 1 |** Comparison of Assessment Processes and AfL Practices.

| Assessment as a Psychometric Process (AERA, APA, and NCME, 2014) | AfL practices Leahy et al., 2005 |
|---|---|
| Design of tasks or activities to elicit evidence of the desired and intended learning outcomes specified by a curriculum or test specification | Clarifying and sharing learning intentions and criteria for success as per the curriculum |
| Collection of evidence through those tasks, through a wide variety of means including tests, portfolios, homework, in-class individual and group work, classroom discussion, and so on. | Engineering effective classroom discussions, questions, and learning tasks.<br>Activating students as the owners of their own learning. |
| Validated scoring of tasks, tests, performances, etc. | |
| Interpretation of performances leading to identification of strengths and weaknesses at individual and group levels | |
| Reporting or communication of the interpretation and action to relevant parties (i.e., students, colleagues, parents) | |
| Monitoring impact of interpretations and actions to ensure intended positive outcomes are achieved without the introduction of unintended negative consequences Messick, 1989. | |
| | Providing feedback that moves learners forward<br>Activating students as instructional resources for one another |

and beyond the explicit curricular goals (Barnes, 1976; Torrance and Pryor, 1998). In considering how AfL has been described, I am very comfortable with it as an excellent template for effective teaching (Black and Wiliam, 2006; Stobart, 2006; Brown, 2013). Students learn effectively when they are given feedback that leads them toward achievable, challenging goals relative to their current standing or capacity are (Hattie and Timperley, 2007). However, does this make AfL actually an assessment process, given that some of the phases of evaluation have been left out?

## A NAGGING CONCERN: VERIFIABILITY

My fundamental premise is that assessments require robust theoretical and empirical evidence to lead to appropriate and valid interpretations and actions (as Messick defined validity in 1989). Robustness requires that the evidence and the inferential processes leading to decisions and actions are open to scrutiny such that we can be satisfied that an appropriate analysis and response has taken place. This is what takes place with standardized tests and formal examinations. There are mechanisms to ensure consistency and accuracy in marking or scoring. Standard setting processes are used to ensure that valid and appropriate decisions are made about the merit and worth of various levels, types, or categories of performance. Monitoring is put in place to ensure that as the evaluative system is deployed the consequences of the system are what was intended. Additionally, efforts are being made within the assessment industry to ensure that results from formal assessments are effectively communicated to relevant stakeholders (Hattie, 2010; Zapata-Rivera, 2018).

For example, classroom assessment of student writing in the English language arts often requires marking of student essays by two or more colleagues working in the same department. Rubrics are developed and referred to, disagreements are discussed, and consensus is reached; an example of social moderation. This means that the individuals concerned (student and teacher) can be confident that the comments and/or grades/scores are a fair reflection of the characteristics and quality of the work produced by the student. If two competent teachers reach similar interpretations or decisions given the same information, then we can say the assessment process has been robust. Consequentially, both the teacher and the student can make plans about what to do next based on a verified assessment of their work.

Assessment processes that take place solely in the head of the learner or teacher are difficult to scrutinize and validate, especially in light of how quickly they must happen and how little material evidence there can be of what led the teacher or a peer to respond as they did. Such processes are inevitable, unavoidable, and desirable in classroom action. However, without an opportunity to subject the classroom processes to validation, it is difficult to treat them as a basis for decision-making beyond classroom action.

Verifiability is necessary whenever there are any consequences attached to classroom assessment processes. Whenever there are risks in a process, the greater surety of credibility in the judgment or scoring processes there needs to be. Bridges have to be able to cope with the potential of collapse, injury, death, and destruction of property. Hence, engineers build them to be more than strong enough for the obligations and responsibilities put on them. Classrooms ought to be places in which little risk to student or teacher welfare occurs. However, risks abound simply in the sense that assessments have psychological, social, as well as educational, consequences. Getting a score or comment other than expected may be hard to receive and giving it may be hard too. If teachers have much time and many opportunities to adjust their teaching (as might be the case in the first term of a primary school year), then perhaps there is no need to check that a teacher's AfL pedagogical interactions in the flow of classroom life are comparable to another teacher's.

However, classroom assessments are designed to have consequences for students (e.g., assign harder or easier curriculum materials, put students in different learning contexts, motivate students to greater effort, diagnose learning difficulties, etc.). Such decisions have to be made on a regular basis in all schooling. Thus, it seems necessary that the risks of getting the decision wrong are not null; the impact on an individual could be severe. Hence, a more formal approach to assessment that insists on checking the validity of the data collection, interpretation, and responses seems warranted. Simply leaving assessment in the head or hands of a teacher prevents scrutiny, debate, or discussion as to the basis and legitimacy for the interpretation and actions. But, this is quite a different matter to the kinds of decisions imagined by assessment *for* learning. Thus, if decisions

have to be made about the best way to teach or if reports about learning need to be communicated to stakeholders, then it seems to me we need to be sure that the interpretations and actions arising from classroom assessment can be validated.

Related to this, as Scriven (1991) has pointed out evaluations that take place early in a process have to be as equally robust and trustworthy as those which take place at the end. If assessments in the formative phase of instruction are not verifiably accurate, there is a good chance they will provide wrong signals to participants about where things are relative to the destination and what has to be done to reach that goal. Poor quality assessments (i.e., those with much error in them) cannot possibly lead to improvement, except by accident. So if AfL is really meant to guide instruction, it needs to move beyond (but include) the intuition of the teacher. Assessment, as opposed to AfL, has to be a competent opinion based on inspection of multiple sources of evidence (including those that can be verified by another competent judge or from trustworthy sources of data) leading to agreement as to what the right interpretation and action might be. If these characteristics are not included in AfL, then it is difficult to perceive that AfL is assessment.

## CLASSROOMS ARE FULL OF HUMANS AFTER ALL

An important constraint upon AfL is that it takes place in the human and social context of the classroom. The humanity of teachers and students is what makes schooling interesting and difficult to manage. A schooling future of children learning alone with their computer tutors is not one that I would wish for my grandchildren. Nonetheless, we need to be realistic about the strengths and weaknesses of the humans in whose hands AfL is placed.

### Teachers

AfL depends on teachers having the ability to understand curriculum and pedagogy (which ought to be the case if well-trained and working within their field of expertise). More importantly they need to have the wit to notice, interpret, and respond appropriately in-the-moment to the contributions of anywhere from 20 to 40 students simultaneously. Though, this parameter can be much larger in places like China or India or in higher education. AfL requires teachers to design appropriate tasks, elicit good information, and respond to it appropriately all within seconds. Teachers are asked to do a job that most professionals do not have to do. Lawyers, doctors, dentists, and so on deal with one client at a time; teachers deal with 10s of students simultaneously.

In classroom interaction, teachers can easily misunderstand a student contribution without any malevolent intention, respond based on that misunderstanding, and wreak minor to massive consequences for a student. Getting it wrong seems to be the default position for teachers simply because they do not have time and resources to continually respond appropriately and accurately to all the students under their care, through all the moments and varying activities of the day. It's a hard job to teach, let alone assess in this way.

AfL requires teachers to be sensitive to what students are doing and thinking, and capable of guiding and responding to that with minimal error. Indeed, the AfL model seems to suggest that there is no error in the teacher input side of the interaction; this seems to be an extremely romantic and naïve expectation, in my opinion. Even when we give teachers time to evaluate student work, it is difficult for them to reach consensus as to the merits or needs of student work. Extensive work on teacher rating suggests that getting to agreement is very hard (Brown, 2009) and even when standardized test scores are available to inform, teacher judgments can be distorted by student characteristics (Meissel et al., 2017).

Getting it wrong is a characteristic of all assessment practices. Standardized test developers calculate and communicate the degree of error in their assessment processes. They may underestimate that error but they identify that there is error. However, in AfL it seems teachers are presumed to be error free in the questions they pose, how they understand student contributions, or even in the feedback they give to students. Their intentions are good, but they can be wrong. If we cannot admit the possibility of teacher error throughout AfL, then it simply cannot meet the expectation that assessments are trustworthy.

Thus, it is very hard to have confidence in a teacher's intuition, without either or both social and statistical moderation. Having time to consider and compare one's intuitive judgments with a colleague is a luxury afforded medical practitioners and professors, but rarely given to teachers. This could be grounds for smaller classes, continual professional development, and low-stakes for error in judgments. While most people can remember a good teacher, they always remember a teacher who got it wrong. This is why creating communities of evaluative judgement are essential to the ability of teachers to interact with students more effectively.

### Students

AfL is predicated on the active role students play in understanding criteria and targets, giving each other feedback, and making progress toward greater learning. Because recognizing the qualities of work is difficult, learners need insights from others; individuals are often too close to their own work to be able to properly consider its strengths or weaknesses. Furthermore, since there is usually only one teacher per class and many students, it makes sense that students would make use of each other as learning resources when it comes to evaluation and diagnosis and even prescription of next steps around their work. Indeed, insights from others can help correct both inappropriate overly optimistic or pessimistic considerations of work. Gaining the ability to realistically and veridically (to use Butler's, 2011 word) judge work characteristics is an important life and work skill. Getting students to actively engage in the process of recognizing features of work and how they relate to criteria and standards is an important learning objective in itself (Tai et al., 2018). Being involved in educational activities that mimic assessment, without the consequences normally associated with evaluation, is a valuable curricular activity

(Brown and Harris, 2014). Thus, the very activities that most look like student-involved assessments are actually valuable curricular activities rather than good assessments.

The active involvement of students in AfL requires them to be robustly honest about their own weaknesses and strongly supportive of others. This means they need to be incredibly mature and psychologically robust to handle peer feedback or tell the truth about themselves. The problem with this construction is simple; students are humans with psychological and social concerns as complex as adults. Lois Harris and I have written extensively about how students dissemble to each other and teachers in classroom assessment (Harris et al., 2009, 2014, 2015, 2018; Harris and Brown, 2013). Children in school are free-will beings who do not necessarily like or trust their classmates or teacher. They can feel threatened by the attitudes and behavior of those around them in the classroom environment. Students can and will lie about the qualities of their own work, their friend's work, or in response to injunctions from a teacher. None of this should surprise us since adults also look to protect themselves from harm and to maximize their self-worth. Even when asked to evaluate their own work, students will deceive themselves about how good their work is and lie to others about it (Brown and Harris, 2013). This can happen because, as novices (Kruger and Dunning, 1999), students are not as able to see quality as teachers, for example, and part of it comes from lack of safety and trust in the social environment.

Creating an environment in which students are encouraged and supported in telling and receiving the truth about their work without fear of recrimination or alienation from others is necessary and hard. AfL practices can and do contribute to the acquisition of those skills. However, it does so by being a curricular and pedagogical practice, not an assessment process.

## CONCLUDING THOUGHT

To conclude, assessment is a separate entity (i.e., a verifiable decision making process) from AfL which is an interactive, intuitive, expert based process embedded within curriculum-informed teaching and learning. I certainly want AfL to co-exist with assessment; but I consider AfL to be an insightful pedagogical practice that ought to lead to better learning outcomes and much more capable learners. However, given the threats to the validity of interpretations and judgments arising from these AfL practices, the approach to formative assessment embedded in AfL does not provide us with the verifiability and legitimacy that assessment requires.

In AfL, much of the inferential process about what to pay attention to, how to interpret it, and what response to make is located in the mind of the teacher; it simply isn't available to others. In contrast, assessments are expected to provide evidence of validity and reliability; this needs to be carried out in an open-space in which multiple eyes can examine the evidence and query the inferential processes behind the decisions. There simply is insufficient time in an AfL pedagogy for inspection of inferences, so AfL does not meet the standards implied by validity expectations of systematic evidence gathering about learning. AfL is an excellent but difficult teaching framework, but it is not assessment which depends upon verifiability for its legitimacy as a tool for decision-making.

## DATA AVAILABILITY

All datasets analyzed for this study are cited in the manuscript and the supplementary files.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## FUNDING

Support for the publication of this paper was received from the Publishing and Scholarly Services of the Umeå University Library.

## REFERENCES

American Educational Research Association [AERA], American Psychological Association [APA], and National Council for Measurement in Education [NCME]. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Barnes, D. (1976). *From Communication to Curriculum*. London: Penguin Press.

Black, P., and Wiliam, D. (1998a). Assessment and classroom learning. *Assess. Educ.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* 80, 139–144, 146–148.

Black, P., and Wiliam, D. (2006). "Developing a theory of formative assessment," in *Assessment and Learning,* ed J. Gardner (London: Sage), 81–100.

Brookhart, S. M. (2016). "Section discussion: building assessments that work in classrooms," in *Handbook of Human and Social Conditions in Assessment,* eds G. T. L. Brown and L. R. Harris (New York, NY: Routledge), 351–365.

Brown, G. T. L. (2009). "The reliability of essay scores: the necessity of rubrics and moderation," in *Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice and Research,* eds L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston, and M. Rees (Wellington, NZ: Ako Aotearoa), 40–48.

Brown, G. T. L. (2013). "Assessing assessment for learning: reconsidering the policy and practice," in *Making a Difference in Education and Social Policy,* eds M. East and S. May (Auckland, NZ: Pearson), 121–137.

Brown, G. T. L. (2018). *Assessment of Student Achievement*. New York, NY: Routledge. doi: 10.4324/9781315162058

Brown, G. T. L., and Harris, L. R. (2013). "Student self-assessmentm," in *The SAGE Handbook of Research on Classroom Assessment,* ed J. H. McMillan (Thousand Oaks, CA: Sage), 367–393. doi: 10.4135/9781452218649.n21

Brown, G. T. L., and Harris, L. R. (2014). The future of self-assessment in classroom practice: reframing self-assessment as a core competency. *Front. Learn. Res.* 3, 22–30. doi: 10.14786/flr.v2i1.24

Brown, G. T. L., and Hattie, J. A. (2012). "The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning," in *Contemporary Debates in Childhood Education and Development,* eds S. Suggate and E. Reese (London: Routledge), 287–292.

Brown, G. T. L., Irving, S. E., and Keegan, P. J. (2014). *An Introduction to Educational Assessment, Measurement and Evaluation: Improving the Quality of Teacher-Based Assessment, 3rd Edn*. Auckland, NZ: Dunmore Publishing.

Brown, G. T. L., and Ngan, M. Y. (2010). *Contemporary Educational Assessment: Practices, Principles, and Policies*. Singapore: Pearson Education South Asia.

Butler, R. (2011). Are positive illusions about academic competence always adaptive, under all circumstances: new results and future directions. *Int. J. Educ. Res.* 50, 251–256. doi: 10.1016/j.ijer.2011.08.006

Crooks, T. J. (2010). "Classroom assessment in policy context (New Zealand)," in *The International Encyclopedia of Education, 3rd Edn,* eds B. McGraw, P. Peterson, and E. L. Baker (Oxford: Elsevier), 443–448. doi: 10.1016/B978-0-08-044894-7.00343-2

Crooks, T. J., Kane, M. T., and Cohen, A. S. (1996). Threats to the valid use of assessments. *Assess. Educ.* 3, 265–285. doi: 10.1080/0969594960030302

Harris, L. R., and Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: case studies into teachers' implementation. *Teach. Teach. Educ.* 36, 101–111. doi: 10.1016/j.tate.2013.07.008

Harris, L. R., Brown, G. T. L., and Dargusch, J. (2018). Not playing the game: student assessment resistance as a form of agency. *Austr. Educ. Res.* 45, 125–140. doi: 10.1007/s13384-018-0264-0

Harris, L. R., Brown, G. T. L., and Harnett, J. A. (2014). Understanding classroom feedback practices: a study of New Zealand student experiences, perceptions, and emotional responses. *Educ. Assess. Eval. Account.* 26, 107–133. doi: 10.1007/s11092-013-9187-5

Harris, L. R., Brown, G. T. L., and Harnett, J. A. (2015). Analysis of New Zealand primary and secondary student peer- and self-assessment comments: applying Hattie and Timperley's feedback model. *Assess. Educ.* 22, 265–281. doi: 10.1080/0969594X.2014.976541

Harris, L. R., Harnett, J. A., and Brown, G. T. L. (2009). "'Drawing' out student conceptions: using pupils' pictures to examine their conceptions of assessment," in *Student Perspectives on Assessment: What Students Can Tell Us About Assessment for Learning,* eds D. M. McInerney, G. T. L. Brown, and G. A. D. Liem (Charlotte, NC: Information Age Publishing), 321–330.

Hattie, J. (2009). *Visible Learning: A Synthesis of Meta-Analyses in Education*. London: Routledge. doi: 10.4324/9780203887332

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hattie, J. A., and Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *J. Educ. Technol. Syst.* 36, 189–201. doi: 10.2190/ET.36.2.g

Hattie, J. A. C. (2010). Visibly learning from reports: the validity of score reports. *Online Educ. Res. J.* 1–15. Available online at: http://www.oerj.org/View?action=viewPaper&paper=6

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Leahy, S., Lyon, C., Thompson, M., and Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educ. Leadersh.* 63, 19–24. Available online at: http://www.ascd.org/publications/educational-leadership/nov05/vol63/num03/Classroom-Assessment@-Minute-by-Minute,-Day-by-Day.aspx

Lindquist, E. F. (ed.). (1951). *Educational Measurement*. Washington, DC: American Council on Education.

Markus, K. A., and Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York, NY: Routledge.

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* 65, 48–60. doi: 10.1016/j.tate.2017.02.021

Messick, S. (1989). "Validity," in *Educational Measurement, 3rd Edn*, ed R. L. Linn (Old Tappan, NJ: MacMillan), 13–103.

Ministry of Education (2007). *The New Zealand Curriculum for English-Medium Teaching and Learning in Years 1-13*. Wellington, NZ: Learning Media.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assess. Educ.* 14, 149–170. doi: 10.1080/09695940701478321

Roediger, III, H. L., Agarwal, P. K., McDaniel, M. A., and McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *J. Exp. Psychol.* 17, 382–395. doi: 10.1037/a0026252

Scriven, M. (1967). "The methodology of evaluation," in *Perspectives of Curriculum Evaluation, Vol. 1*, eds R. W. Tyler, R. M. Gagne, and M. Scriven (Chicago, IL: Rand McNally), 39–83.

Scriven, M. (1991). "Beyond formative and summative evaluation," in *Evaluation and Education: At Quarter Century, Vol. 90,* eds M. W. McLaughlin and D. C. Phillips (Chicago, IL: NSSE), 19–64.

Shavelson, R. J. (2008). Guest editor's introduction. *Appl. Measu. Educ.* 21, 293–294. doi: 10.1080/08957340802347613

Stobart, G. (2006). "The validity of formative assessment," in *Assessment and Learning,* ed J. Gardner (London: Sage), 133–146.

Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assess. Educ.* 18, 433–449. doi: 10.1080/0969594X.2011.582838

Tai, J., Ajjawi, R., Boud, D., Dawson, P., and Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *High. Educ.* 76, 467–481. doi: 10.1007/s10734-017-0220-3

Torrance, H., and Pryor, J. (1998). *Investigating Formative Assessment: Teaching, Learning and Assessment in the Classroom*. Buckingham, UK: Open University Press.

Zapata-Rivera, D. (ed.). (2018). *Score Reporting: Research and Applications*. New York, NY: Routledge.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.