DOI: https://doi.org/10.26686/nzaroe.v26.6854

Responding to Assessment *for* Learning: A pedagogical method, not assessment

Gavin T. L. Brown

University of Auckland

Assessment for learning (AfL) is a major approach to educational assessment that relies heavily on pedagogical practices, such as involving students in assessment, making transparent objectives and criteria, and asking open-ended questions that provoke higher order thinking. In this perspective piece, I argue that without the possibility of opening classroom activities to systematic and rigorous inspection and evaluation, AfL fails to be assessment. AfL activities happen ephemerally in classrooms, leading to in-the-moment and on-the-fly interpretations and decisions about student learning. In these contexts, determination of the degree of error in those judgements does not happen. Because human performance is so variable and because the samples teachers use to make judgements are not robustly representative, there is considerable error in their judgements about student learning. Nonetheless, despite the difficulties seen in putting AfL into practice, they appear to be good classroom teaching practices. In contrast, assessment proper requires careful inspection of data so that alternative explanations can be evaluated, leading to a preference for the most valid and reliable interpretation of performance evidence. Psychometric methods not only quantify amounts or qualities of performance, but also evaluate the degree to which judges agree with each other, leading to confidence in the validity and reliability of insights. Consequently, because AfL activities lack the essential characteristics of paying attention to error and methods of minimising its impact on interpretations, I recommend we stop thinking of AfL as assessment, and instead position it as good teaching.

Keywords: assessment for learning, psychometrics, pedagogical practices, error, assessment

Introduction

In this brief perspective, I want to present my thoughts on the Assessment for Learning (AfL) agenda in education, especially as it is implemented here in New Zealand. I want to compare and contrast what AfL appears to be with what I understand assessment to be. My perspective might lead some of you to share my concerns for how we are using AfL. This brief essay builds on previous papers in which I have shared my doubts about the validity of AfL as assessment (Brown, 2013, 2019). This issue seems important because much of what we know as AfL introduces many construct-irrelevant factors that invalidate decisions about student learning. Perhaps in an instructional context, this may not matter, but for assessment, I think it is extremely important.

My take on assessment is grounded in the notion that includes any act of interpreting and acting on information about student performance (Messick, 1989), collected through any of a multitude of means or practices (Gipps, Brown, McCallum, & McAlister, 1995). Given this breadth of purpose and form, it is possible to think almost any data collection about student learning qualifies as assessment. For me, that would be a mistake. As a basis for decision making, whether during instruction or in summing performance, the information collected has to be a valid sample and accurate description of what the learner knows or can do. At its simplest, this means that the quality of the data collection and interpretive processes has to be demonstrated. The assessment interpretations themselves and the people who make them need to be viewed by stakeholders (e.g., students, parents, administrators, etc.) as defensible judgements made by judges who are accepted as competent to make such judgements (Cresswell, 1996). Fundamentally, that means that unless processes are open to inspection and corroboration, they cannot qualify as assessment.

I attribute my approach to Popper's (1945) perspective on science, in which he argues that science does not look for facts which confirm pre-existing theories. It seems to me that teachers have theories about what learners (as groups or individuals) can do and know. One of the roles of assessment, then, is to test those theories in order to find flaws in our understanding of learner capability and performance. For an assessment to have the potential to falsify a teacher's theory of student learning, assessments have to have the potential to surprise or expose false perspectives on the learner. Further, the assessments have be robust in quality to serve such an important goal. Unless assessment interpretations and decisions are falsifiable, through testing against alternative perspectives, they do not qualify as assessments. Instead, they may act as legitimate signals for teaching, but unless we can test the interpretations of teachers in situ against other data sources or perspectives, the information cannot, for me, be assessment.

As background, allow me to give you some personal biography. I was a New Zealand high school teacher of English, English as an additional language, remedial reading, and social studies for a decade in west and south Auckland. Most of my professional career, however, has been spent becoming an expert in educational assessment and applied psychometric theory. I was an assessment research officer at the New Zealand Council for Educational Research helping to develop the *Assessment Resource Bank* and the *Essential Skills Assessments for Information Literacy*. Then, I spent five years running the *Assessment Tools for Teaching and Learning* (asTTle) project, which has morphed now into the e-asTTle system. That applied research focused on a simple idea: "*how to make a test worth taking that would tell a teacher something they needed to know about who needs to be taught what next.*" Since then, I've had an academic career working cross-culturally on assessments and the psychology of assessment. I've conducted research on how teachers and students experience and understand assessment and seen how that influences outcomes and behaviours. I've written about marking essays (Brown, 2009, 2010), student self-assessment (Brown & Harris, 2013, 2014; Harris & Brown, 2018), classroom assessment (Brown, 2018; Brown, Irving, & Keegan, 2014) and psychometric analysis of tests (Brown & Abdulnabi, 2017).

Assessment for Learning

Most basically, all assessment processes that are designed to contribute to greater, deeper, or more sophisticated student learning involve collecting evidence about the current state of learning early enough so that teacher action (i.e., teaching) and student activity (i.e., learning) change in a way that causes progress towards intended goals and targets. Progress involves being able to do things more quickly and more efficiently, it involves being able to do or know more, and it involves creating better quality performances and products. With this as the goal, AfL must happen formatively (Scriven, 1967). That means it must take place before the end so that it can contribute to improvement. It must lead to changes in both teaching and learning practices and result in progress. If these things are happening, then there should be evidence in student work, attitudes, and self-perceptions that improvements have occurred. Further, it should be possible to see that previous data collection has informed teachers' changed practices in terms of materials they use, sequences of activities deployed, grouping of students, and activities that students actually engage in. That means AfL must produce information that teachers can use to appropriately guide changes in their own practices.

This requires that we accept that there is an intense and intertwined relationship between assessment, learning, and teaching. Allow me to outline in bullet points four major ways of relating assessment with curriculum, teaching, and learning:

- Traditionally and in New Zealand during the days of School Certificate and University Entrance examinations, this relationship was purely linear: The curriculum told teachers what they should teach, teachers taught it, and an external qualifications board created formal examinations that were used to judge student consequences. Normally, teachers would use one year's examination results to adjust their teaching for the next cohort of students and so only distally was there a formative effect. It is also evident that students learn in this summative assessment of learning process (Brown, 2022).

- An intermediate step, seen in the National Certificate in Educational Achievement (NCEA) in New Zealand secondary schools, is for teachers to become the examiners for high-stakes qualifications. Teachers use a variety of data collection tools (e.g., direct observation of performances, portfolios, long constructed response products; Crooks, 2010) and systematic ways of ensuring validity and reliability of judgements (e.g., multiple raters, moderation among raters, and scoring rubrics with specified marking criteria). These systems allow teachers to make judgements that are used to award qualifications and to formatively provide feedback to their own students about quality and improvement. Future students may benefit from information teachers gain about what worked well or not. This system is formal, guided by externally devised and defined curricular goals with the invidious potential for qualifications and standards to become the curriculum. This approach requires teachers to shift identity and role from being purely instructors to also being judges of quality, a challenge for many teachers (Xu & Brown, 2016).

- More commonly in the grades prior to the NCEA qualifications (i.e., junior high school, intermediate, primary, and early childhood settings), teachers exercise considerable freedom around broadly defined curricular objectives to interact with learners as they develop. These systems depend heavily on AfL practices (Black & Wiliam, 1998; Leahy, Lyon, Thompson, & Wiliam, 2005) such as, giving students access to learning intentions and criteria, peer and self-assessment against those standards, and teacher feedback and questions indexed to the assessment criteria. This approach has very much taken hold in the child-centric pedagogical processes of Anglo-centric primary school contexts (Stobart, 2006). Unlike other jurisdictions, in New Zealand diagnostic tests are reasonably prevalent (Crooks, 2010) but are school rather than externally controlled and provide teachers information as to "who needs to be taught what next" (Brown & Hattie, 2012). In this approach the curriculum frames the learning and the teaching, and subordinates assessment to a supportive or quasi-invisible role.

- The strongest AfL position is one which is free of intended learning intentions, pre-specified learning goals, or teacher-centric educational practices (Swaffield, 2011). This approach, perhaps more pronounced in higher education (Boud & Associates, 2010; Bourke, 2017), puts the learner at the centre alongside personally defined learning outcomes and divergent conversations. The emphasis here is interaction among the teacher and students and in which assessment is an in-the-moment, subjective and intuitive approach (Ministry of Education, 2007) to adjusting classroom activities to the needs and strengths of the learner (Hill, 2000). The ultimate goal of this 'learner-centred' approach to assessment seems to be the satisfaction of the learner's needs rather than those of the mandated curriculum, the priorities of the instructor, or the authority of assessment practices.

The last pair of definitions make it clear that AfL seems to reject the idea that assessment only occurs at the end of instruction and raises doubts about the validity of tests and exams per se. A positive outcome of AfL is the use of greater diversity of assessment methods and an explicit attention to using assessment processes to deliberately improve the quality of student learning. However, there is a challenge when AfL relies on data collection methods that are conducted on-the-fly and in-the-moment (Ministry of Education, 2007). The challenge is simple: *how do we know if the interpretations made in this way are correct and do they lead to valid decisions and actions*? Of course, there are those who would doubt the need to be concerned about these questions. But allow me to make a case for a more traditional view informed by psychometric science.

A psychometric view

Based on classical test theory (Novick, 1966; Traub, 1997), any assessed score is deemed to consist of two components: the *true* score of a person's ability and the *error* involved in conducting the assessment. Good tests estimate the amount of unexplained variance there is in a test. For example, the diagnostic tests available for use in New Zealand classrooms (e.g., PAT and e-asTTle) have estimated the degree of error in their reported scores. The PAT margin of error is +/- 3 points on scale scores to indicate the degree of accuracy in the scale scores for each subject tested. For e-asTTle the standard error is 22 points; meaning that score differences less than 22 points are statistically not significant at a 68% confidence interval, and at a 95% confidence interval, differences less than 44 points are not statistically significant. That 44-point range is as big as two years' normative gains in the middle primary school years (Ministry of Education, 2010).

Student performance is difficult to measure accurately, in part because humans are so variable. What learners know now may not be accessible tomorrow because of random factors such as the effect of sleep, diet, noisy environments, and so on. In general, this random variability in human performance has been shown to be considerably less influential on score variation than systematic errors introduced by human scoring of performance (Brennan, 1996). So, while learners may perform differently on tests without any substantive change in their capabilities, there is robust evidence that systematic errors exist in all educational assessment processes. In assessments scored, rated, or judged by humans, we see considerable error introduced to estimates of student performance through the act of raters and judges (Black & Wiliam, 2012). Human judges are distractable, emotional, inconsistent, and prone to misjudging phenomena (Brown, 2009, 2017).

The challenge for AfL comes in estimating the error in human judgement processes. Published data on the degree to which New Zealand teachers agree or disagree with each other in making curriculum level, overall teacher judgements (OTJs), or grades for NCEA assessments is generally lacking. Hence, it is difficult to know if in operation teachers have high or low agreement when assigning marks or grades. It may be that there is high agreement (e.g., absolute consensus >70%, inter-marker correlation >.70; Stemler, 2004), but evidence of this is absent. Thus, it could be that standardised tests have more error than teacher judgements, but this is unlikely. And more importantly, if teacher judgements are more accurate than tests, evidence of this is not easily obtained.

To illustrate the problem, consider Meissel et al.'s (2017) detailed analysis of New Zealand teacher OTJs relative to standardised test scores. At the time of the study, teachers were required to report how students were performing in reading, writing, and mathematics relative to year level expectations. A student was deemed to be 'well-below' if they were working two or more years lower than their year level, 'below' if they were one year behind, 'at' if they were within the normative range for their year, and 'above' if they were one year ahead of their year group. Student performance on standardised tests with population norms could also be classified using the same quality categories by adjusting test performance by the available year norms. Meissel et al. (2017) reported that teacher OTJs were consistently lower than the test performance of male, non-white, and special education students. If the judges had been free from bias, the degree of mismatch between judgement and score should have been small *and* constant across all groups. A number of explanations for this discrepancy exist, including variable teacher expectations by student demographic factors, students' variable familiarity and practice with standardised tests, mismatch of teaching to test content, lack of OTJ moderation among teachers, or even unconscious bias (Shah et al., 2020). 'Otunuku and Brown (2007) even suggested that teachers may be delivering a curriculum that disguises from Māori and Pasifika students that their performance is well below expectations.

To overcome this inconsistency, "careful selection and training of markers, detailed protocols for markers to follow, and monitoring performance of markers through scrutiny by a more experienced marker of a sample of their marking" (Black & Wiliam, 2012, p. 246) are all needed. Thus, in every human act of using student performances to form an opinion about educational needs and strengths, we humans need to admit the possibility that we got it wrong as to the students' learning needs and abilities. Ignoring the possibility of human error may make for efficient teacher activity, but it doesn't bode well for trustworthy, dependable decision-making. From this perspective, until AfL attempts to estimate and control the error in human judgement, it will not meet a fundamental requirement of assessment.

AfL as pedagogy

Not taking the possibility of error into account makes assessment in AfL something that is not assessment. An allowance for error not only requires humility, but also calls out for mechanisms by which human judgements can be tested for validity. In other words, for assessments to be assessments they have to be verifiable or testable. This, of course, is a fundamental principle of the scientific method; results are evaluated by peer review and disclosed for replication (Stanovich, 2009). Thus, if stakeholders (especially parents and students) want to trust teacher interpretations based on ephemeral events in a classroom, they need some mechanism assuring that the data are captured, and the judgement processes evaluated for validity and accuracy. This is unlikely without distorting the quality of classroom experiences.

As suggested by Black and Wiliam (2006), AfL is properly a set of pedagogical practices intended to improve the quality of teaching and the depth and speed of student learning. It is certainly desirable for students to have clarity about learning goals or intentions, to have opportunities to reflect on their own and their peers' progress, to be asked questions that challenge their thinking, and to receive feedback that helps them develop around task, processes, and self-regulation (Hattie & Timperley, 2007). These factors are part of what Hattie (2009) describes as activator strategies; good teaching does these things.

My own research has shown that problems inherent in various AfL practices mean that they can't be treated as verifiable assessments. Consider:

- student self-serving inaccuracy within self-assessment (Brown & Harris, 2014)
- the inter-personal relationships that distort peer assessments and feedback (Harris & Brown, 2013)
- the difficulties in getting students to do the formative work we set (Harris, Brown, & Dargusch, 2018)
- the reluctance of students to use the feedback we provide (Harper & Brown, 2017)
- the reluctance of students to classify AfL practices as assessments that link to their performance (Brown, Irving, et al., 2009; Brown, Peterson, & Irving, 2009), and
- the tendency of students to pay attention only to assessment practices that have consequences for their own future (Brown, 2021)

Thus, I consider these AfL teaching practices are not assessments, partly because they happen too quickly to be properly documented for public inspection and subsequent verification of the validity of the interpretations being made by instructors. Students are sensitive to teacher assessment practices (Peterson & Irving, 2008; Remesal, 2009), but independent of systematic student evaluations of teaching, their perspectives are not normally available to correct inappropriate practices. Any error in the conclusions that teachers reach in classrooms may be overcome by the time teachers have with students. Theoretically, teachers who have more time with students have more opportunity to correct any inaccurate evaluations of student capabilities. However, even when primary school teachers have 20-25 hours per week with a group of students, each student might receive just a few minutes per week of individual interaction with the teacher. Thus, over the long term, as Meissel et al. (2017) suggest, errors may accumulate into large discrepancies that generate substantial disservice to students and families. Brown and Harris (2016) suggested that while educators and society might be concerned about the negative impact of high-stakes testing (e.g., promotion, graduation, selection for awards, etc.), "the cumulative effect of 'garbage in, garbage out' factors on the many low-stakes decisions implemented in classrooms (e.g., placement in an inappropriate within-class reading group, assignment of class work which is too easy/difficult, etc.) may have serious impacts on learners' motivation, effort, progress, and desire to learn" (p. 508).

Given that classroom assessments are prone to many construct-irrelevant sources that distort the accuracy and validity of decisions, it may be best to treat such activities as teaching rather than assessment. That is certainly how Lois Harris and I have recommended teachers think of self-assessment, because of the construct-irrelevant sources of error (Brown & Harris, 2014; Harris & Brown, 2018). We concluded (Brown &

Harris, 2016) that "it may not be too much of a stretch to argue that all classroom interactions, despite their similarity to formal assessment practices, should be treated rather as teaching-learning or pedagogical interactions" (p. 514). Consequently, it seems much safer to call AfL good teaching and avoid thinking of it as assessment at all. With that in mind, we need to focus on the real goal of educational assessment; identifying who needs to be taught what next (Brown & Hattie, 2012). AfL has potential to generate insights that inform teaching and learning, but for the most part such insights seem to be accepted as correct, despite their credibility being far from certain. Nevertheless, the goal of AfL is changing both student learning and teacher instruction; as such those are good ambitions. I still maintain that AfL is not assessment, as I have understood it.

References

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

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81-100). London: Sage.

Black, P., & Wiliam, D. (2012). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 243-263). London: Sage.

Boud, D., & Associates. (2010). *Assessment 2020: Seven propositions for assessment reform in higher education*. https://www.uts.edu.au/sites/default/files/Assessment-2020_propositions_final.pdf

Bourke, R. (2017). Self-assessment to incite learning in higher education: Developing ontological awareness. *Assessment & Evaluation in Higher Education*, 1-13. https://doi.org/10.1080/02602938.2017.1411881

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment (NCES 96-802)* (pp. 19-58). Washington, DC: National Center for Education Statistics.

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston, & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Wellington: Ako Aotearoa.

Brown, G. T. L. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly, 64*(3), 276–291. https://doi.org/10.1111/j.1468-2273.2010.00460.x

Brown, G. T. L. (2013). Assessing Assessment for Learning: Reconsidering the policy and practice. In M. East & S. May (Eds.), *Making a difference in education and social policy* (121-137). Auckland: Pearson.

Brown, G. T. L. (2017). The future of assessment as a human and social endeavor: Addressing the inconvenient truth of error. *Frontiers in Education, 2*(3). https://doi.org/10.3389/feduc.2017.00003

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

Brown, G. T. L. (2019). Is assessment for learning really assessment? *Frontiers in Education, 4*(64). https://doi.org/10.3389/feduc.2019.00064

Brown, G. T. L. (2021). Student conceptions of assessment: Understandable responses to our practices*. ECNU Review of Education*. https://doi.org/10.1177/20965311211007869

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*. London: Routledge.

Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education, 2*(24). https://doi.org/10.3389/feduc.2017.00024

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). Thousand Oaks: SAGE. https://doi.org/10.4135/9781452218649.n21

Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 3*, 22-30. https://doi.org/10.14786/flr.v2i1.24

Brown, G. T. L., & Harris, L. R. (2016). Volume conclusion: The future of assessment as a human and social endeavour. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 506-524). New York: Routledge.

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

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

Brown, G. T. L., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. F. (2009). Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97-111. https://doi.org/10.1016/j.learninstruc.2008.02.003

Brown, G. T. L., Peterson, E. R., & Irving, S. E. (2009). Beliefs that make a difference: Adaptive and maladaptive self-regulation in students' conceptions of assessment. In D.

M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning.* (pp. 159-186). Charlotte, NC: Information Age Publishing.

Cresswell, M. J. (1996, September). *What are examination standards? The role of values in large scale assessment.* Paper presented at 22nd Annual IAEA Conference, Beijing, China.

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

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition or evidence? Teachers and national assessment of seven-year-olds*. Buckingham: Open University Press.

Harper, A., & Brown, G. T. L. (2017). Students' use of online feedback in a first year tertiary biology course. *Assessment Matters, 11*, 99-121. https://doi.org/10.18296/am.0026

Harris, L. R., & 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. *Teaching and Teacher Education, 36*, 101-111. https://doi.org/10.1016/j.tate.2013.07.008

Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning*. New York: Routledge.

Harris, L. R., Brown, G. T. L., & Dargusch, J. (2018). Not playing the game: Student assessment resistance as a form of agency. *The Australian Educational Researcher*. https://doi.org/10.1007/s13384-018-0264-0

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses in education*. London: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Hill, M. (2000). Dot, slash, cross: How assessment can drive teachers to ticking instead of teaching. *set: research information for teachers* (1), 21-25. https://doi.org/10.18296/set.0779

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

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48-60. https://doi.org/10.1016/j.tate.2017.02.021

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: MacMillan.

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

Ministry of Education. (2010). *e-asTTle norms and curriculum expectations by quarter: Reading and mathematics*. Retrieved 2 March 2021 from Wellington: https://e-asttle.tki.org.nz/content/download/1229/4816/version/4/file/e-asTTle+norm+tables+Sept+2010.doc

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

'Otunuku, M., & Brown, G. T. L. (2007). Tongan students' attitudes towards their subjects in New Zealand relative to their academic achievement. *Asia Pacific Education Review, 8*(1), 117-128. https://doi.org/10.1007/BF03025838

Peterson, E. R., & Irving, S. E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction, 18*(3), 238-250. https://doi.org/10.1016/j.learninstruc.2007.05.001

Popper, K. (1945). *The open society and its enemies: The high tide of prophecy: Hegel, Marx, and the aftermath* (Vol. II). London: George Routledge & Sons.

Remesal, A. (2009). Accessing primary pupils' conceptions of daily classroom assessment practices. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 25-51). Charlotte: Information Age Publishing.

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, IL: Rand McNally.

Shah, R., Brown, G. T. L., Keegan, P. J., Burakevych, N., Harding, J. E., & McKinlay, C. (2020). Teacher rating versus measured academic achievement: Implications for paediatric research. *Journal of Pediatrics and Child Health, 56*(7), 1090-1096. https://doi.org/10.1111/jpc.14824

Stanovich, K. E. (2009). *How to think straight about psychology* (9th ed.). New York: Pearson Education.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). https://doi.org/10.7275/96jp-xz07

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

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

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Gavin Brown is a Professor in the Faculty of Education and Social Work at the University of Auckland. He is also Associerad Professor in Educational Sciences at Umeå University, Sweden and an Honorary Professor in Curriculum and Instruction at the Education University of Hong Kong. He is a world leader in educational assessment, testing, and applied measurement. He has over 200 research publications, examining teacher and student psychology of assessment in cross-cultural contexts. His most recent books include *Using self-assessment to improve student learning* and *Assessment of student achievement,* both Routledge (2018).

Email: gt.brown@auckland.ac.nz

ORCiD: https://orcid.org/0000-0002-8352-2351