The future of assessment as a human and social endeavour: Addressing the inconvenient truth of error

Gavin T L Brown

The University of Auckland

Assessment faces continuing challenges. These challenges arise predominantly due to the inherent errors we make when designing, administering, analysing, and interpreting assessments. A widely-held assumption is that our psychometric methods lead to reliable and valid scores; however, this premise depends on students exercising 100% effort throughout a test event, with no cheating, and having had sufficient personal environmental support to produce best possible results (Dorans, 2012).

Inconveniently, research makes clear that cheating and lack of effort contaminate scores (Murdock, Stephens, & Groteweil, 2016; Wise & Smith, 2016). This is especially the case in low-stakes testing situations, such as institutional evaluations (Wise & Cotten, 2009), leading to inappropriate conclusions about the state of an organisation or jurisdiction. Hence, while it is convenient to presume that statistical advances will account for such systematic sources of error, the reality is that much assessment takes place both 'in vivo' and 'in situ' during classroom activities (Zumbo, 2015). Thus, while psychometric methods work reasonably well in high-stakes examination or standardised testing contexts (i.e., 'in vitro'), there is little guarantee that these assumptions hold true for what happens in classroom contexts. Thus, the psychometric and testing industry has much to do to develop methods of describing and accounting for the myriad complexities of classroom or school based dynamics.

This matters because a widespread policy framework of using assessment to guide or inform improvement (i.e., ‘assessment for learning’ or ‘formative assessment’) requires teachers to assess students so as to identify the quality of student learning and appropriate

changes to classroom practices. UK experts tend to argue that this can only be done through teacher-student interaction in the classroom or by involving students in the process of considering the merits of their own or peers' work (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Harlen, 2007; Swaffield, 2011). Others consider that tests can contribute information about changes to teaching that lead to better learning outcomes, provided the tests go beyond rank order or total score reporting (Brown & Hattie, 2012) or if teachers spend time analysing strengths and weaknesses (Carless, 2011).

Regardless of the type of assessment method, it is very difficult for pre-service teachers to learn how to assess formatively (Hill & Eyers, 2016). Indeed, even practicing teachers need expertise in curriculum and pedagogy to exercise command of multiple methods of assessment in such a way that all learners are helped to overcome the, sometimes idiosyncratic, challenges they face (Cowie & Harrison, 2016; Moon, 2016). Teachers in New Zealand and Netherlands have learned to use achievement data to guide school-wide improvements, provided experts give them help (Lai & Schildkamp, 2016). However, such efforts often take 2-3 years before changes can be seen in student performance. Thus, despite multiple studies which show that teachers believe in using assessment formatively (Barnes, Fives, & Dacey, 2015; Bonner, 2016), putting in place policy and resources to support formative assessment is difficult, meaning formative assessment is not a quick fix for improving outcomes for all learners.

The formative assessment policy agenda challenges the dominance of formal testing and teacher-centric methods of assessment, with expectations that effective learning takes place as students engage with learning targets, outcomes, or objectives, take ownership of their work, cooperate with peers, understand more deeply what quality is, and receive and generate appropriate feedback (Leahy, Lyon, Thompson, & Wiliam, 2005). Inconveniently, involving students in assessment presents considerable challenge due to psychological and

social factors that interfere with the student's ability to accurately self-evaluate (Andrade & Brown, 2016) or to constructively peer evaluate and collaborate (Panadero, 2016; Strijbos, 2016). Indeed, evidence that student involvement in assessment develops self-regulatory abilities is weak (Dinsmore & Wilson, 2016). Feedback processes are complex, belying the simple notion that student 'horses' will automatically learn once they are led to the 'water' of feedback (Lipnevich, Berg, & Smith, 2016). While novelty in assessment methods is being developed, especially through introduction of ICT (Katz & Gorin, 2016), it is true that students are not necessarily fans of new ways of being assessed for fear their performance will be impacted (Struyven & Devesa, 2016).

A second wide-spread policy initiative is to use assessments, especially standardised tests, to evaluate teachers, schools, and systems (Lingard & Lewis, 2016; Teltemann & Klieme, 2016). It is clear that such policies tend to have largely negative impact on the quality of teaching (Hamilton, 2003; Nichols & Harris, 2016), and perhaps more so among minority and lower socio-economic communities. Nonetheless, public acceptance of the legitimacy of using assessment scores to ascertain quality in schooling is reasonably high (Buckendahl, 2016). Using tests to evaluate schools and teaching is a relatively quick and low-cost political process (Linn, 2000). However, summative accountability use of assessments creates tensions for teachers (Bonner, 2016), with many teachers in high-stakes accountability environments having very negative views of such uses (Deneen & Brown, 2016). Using assessments formatively requires discovery of what students have ‘failed’ to be good at, so as to inform further instruction (Hattie & Brown, 2008). This implies that a formative assessment ought to reveal lack of success, a problematic event if external accountability consequences are attached to the same result. Thus, if consequences for low scores are seen as unfair, then it is not surprising if teachers use multiple methods to ensure

that scores increase. If accountability assessment scores are inflated through construct-irrelevant processes, then the meaning of an accountability assessment is problematic.

The choice of policy priorities within different jurisdictions strongly shapes the nature and power of assessment practices. For example, both Arabic and Chinese language societies strongly prioritize memorization of content as the dominant model of schooling and attach substantial social and economic benefits for successful performance on formal examinations (Gebril, 2016; Hargreaves, 1997; Kennedy, 2016; OECD, 2011). Anglo-Commonwealth countries strongly prioritize a child-centered, student-involved approach (Stobart, 2006), in which interactive teacher assessment practices have been prioritised as means of improving learning outcomes (Black & Wiliam, 1998). The United States has strong legal protection for special needs students (IDEA, 1997) who are entitled to differentiated assessment and evaluation practices (Tomlinson, 1999). These differences in social uses and styles of assessment complicate the meaning of a grade or score, and create challenges for psychometric models that attempt to create universal explanations of performance.

Within societies that are highly homogenous in terms of ethnic and linguistic make-up (e.g., Finland, Japan, China), it may be reasonable to expect that common psychological and social factors influence assessment. This simplifies predicting and modeling those factors. However, when comparisons are made among culturally-distinct groups in multicultural societies, which is more the case in economically-developed societies and nations (van de Vijver, 2016), the psychological factors influencing student response, teacher judgments, or test performance can vary significantly. For example, tendencies to self-effacement or self-enhancement are not equal across cultural groups (Suzuki, Davis, & Greenfield, 2008), so the meaning of self-assessment has to be carefully evaluated (i.e., among collectivist groups modest self-reporting enhances group belongingness). In multicultural contexts, assessments that depend on classroom interactions between and among students and teachers is likely to

be impacted by these different cultural standards as to the best way to communicate an evaluation of work. The capacity of teachers to appropriately collect, analyse, and plan in response to both formal and informal assessment data is generally weak (Xu & Brown, 2016). Quite prolonged and intensive professional development is needed to generate 'assessment capable' teachers (Smith, Hill, Cowie, & Gilmore, 2014). Thus, assessors and assessments are challenged by the varying and subtle differences created by cultural difference.

Even the introduction of technological solutions that increase the authenticity, diversity, and efficiency of formal testing (Csapó, Ainley, Bennett, Latour, & Law, 2012; Katz & Gorin, 2016) does not necessarily improve student performance or solve problems in scoring. Students' enthusiasm for a computerized activity does not automatically lead to valid conclusions about their proficiency. Students are often concerned that novel assessment practices (including peer assessment, self-assessment, portfolio, performance, or computer-based assessments) will have negative impacts on their performance simply because they are unsure as to how well they will do on a new method of evaluation (Struyven & Devesa, 2016). Consequently, students tend to retreat into a strong preference for conventional assessment practices (e.g., essays or multiple-choice questions). Furthermore, technology now permits data sharing and long-term tracking of student performance, which ought to improve our understanding of how students are improving in which areas. However, the existence of these electronic data raises concerns about privacy and protection; imagine possible negative implications if early poor performance is kept on record and used in evaluative decision-making, despite substantial subsequent progress (Tierney & Koch, 2016).

Thus, inconveniently, the field of testing, applied psychometrics or measurement, and assessment is faced with complex problems, which are not restricted to any one form of assessment or any one society in which assessment is deployed. The inconveniences outlined here are especially the case if we accept that the goal of assessment is to inform improvement

and make valid decisions about learners and teachers. The need for accurate diagnostic prescriptions that teachers, students, and/or parents could use to inform improvement is paramount. These prescriptions need to occur close to and responsive to the real-time processes of classroom learning and teaching, which is a substantial problem. The great contribution of psychometrics to the field of education has been an explicit attention to the problem of error in all testing, measurement, and assessment processes. However, few tools are currently available to robustly estimate and account for the kinds of error that occur in real-time classroom observations, interactions, and interpretations. The inconvenient challenge for educators who would minimise the role assessment plays in curriculum is that high-quality tests and measurements are necessary for justice, fairness, and the well-being of individuals and society. The inconvenient challenge for policy-makers is that many assessment processes are not reliable or dependable (e.g., essay examinations; Brown, 2010), nor do they account well for the many factors outlined here. Thus, many policy decisions based on inadequate tools or processes are invalid.

The future of assessment requires that we no longer ignore these inconvenient problems facing assessment, testing, and applied measurement. Rather, assessment has to turn constructively to deeply insightful investigations into these perennial problems. Teachers and students need to know where learning is and what is next. Policy makers and parents have a right to know what is working, who is learning, who needs help, what needs to change, and so on. Assessment and testing is how we as humans discover the answers to these questions. Hence, good schooling and good education need good testing or assessment, both in the sense of high-quality and rightly done.

Leaning heavily on validity theory (Kane, 2006; Messick, 1989), good assessment leads to defensible interpretations and actions. These uses depend on robust arguments based on relevant theories of curriculum, teaching, learning, and measurement and on trustworthy

empirical evidence that has been subjected to scrutiny (i.e., statistical and/or social moderation). The need to bring greater skill and insight into assessments that inform classroom practice is essential. The success of the whole superstructure of schooling relies on the quality of judgements and evaluations carried out in the millions of classrooms of the world on an everyday basis. If this work is not done well, and if we do not know that it is not done well, we fail.

Hence, engaging in the difficult challenges of how assessment can help education, while also making a credible case for the scores or judgements generated by assessments, needs to be reported. Leaving this only to educational statisticians would be a mistake. Testing and measurement needs to integrate with classroom teaching, learning, and curriculum if it is to support schooling and prevent politicians from making simplistic but wrong interpretations and uses of assessment. This is the Grand Challenge for this Section of the journal *Frontiers in Education*. How can assessment be made flexible enough to support real learning in vivo, while fulfilling all the diverse expectations society has for it? As Section Editor, I look forward to your contributions.

Acknowledgement

This paper draws heavily on Brown, G. T. L. & Harris, L. R. (2016). The future of assessment research as a human and social endeavour. In G. T. L. Brown & L. R. Harris (Eds.). *Handbook of Human and Social Factors in Assessment* (pp. 506-523). Routledge: New York. An earlier version of this paper, presented as an inaugural professorial lecture at the Faculty of Education & Social Work, The University of Auckland, can be seen at doi: 10.17608/k6.auckland.4238792.v1.

References

Andrade, H. L., & Brown, G. T. L. (2016). Student self-assessment in the classroom. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 319-334). New York: Routledge.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. Gregoire Gill (Eds.), *International Handbook of Research on Teacher Beliefs* (pp. 284-300). New York: Routledge.

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

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, UK: Open University Press.

Bonner, S. M. (2016). Teachers' Perceptions about Assessment: Competing Narratives. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 21-39). New York: Routledge.

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

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.

Buckendahl, C. W. (2016). Public perceptions about assessment in education. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 454-471). New York: Routledge.

Carless, D. (2011). *From testing to productive student learning: Implementing formative assessment in Confucian-Heritage settings*. London: Routledge.

Cowie, B., & Harrison, C. (2016). Classroom processes that support effective assessment. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 335-350). New York: Routledge.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 143-230). Dordrecht, NL: Springer.

Deneen, C. C., & Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education, 3*, 1225380. doi:10.1080/2331186X.2016.1225380

Dinsmore, D. L., & Wilson, H. E. (2016). Student participation in assessment: Does it influence self-regulation? In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Factors in Assessment* (pp. 145-168). New York: Routledge.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20-37. doi:10.1111/j.1745-3992.2012.00250.x

Gebril, A. (2016). Educational assessment in Muslim countries: Values, polices, and practices. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 420-435). New York: Routledge.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education, 27*(1), 25-68.

Hargreaves, E. (1997). The diploma disease in Egypt: Learning, teaching and the monster of the secondary leaving certificate. *Assessment in Education: Principles, Policy & Practice, 4*(1), 161-176. doi:10.1080/0969594970040111

Harlen, W. (2007). *Assessment of learning*. Los Angeles: Sage.

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

Hill, M. F., & Eyers, G. (2016). Moving from student to teacher: Changing perspectives about assessment through teacher education. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 57-76). New York: Routledge.

Individuals with Disabilities Act, Pub.L. 101-476 C.F.R. § §1400 et seq. (1997).

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

Katz, I. R., & Gorin, J. S. (2016). Computerising assessment: Impacts on education stakeholders. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 472-489). New York: Routledge.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian-Heritage Cultures. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 404-419). New York: Routledge.

Lai, M. K., & Schildkamp, K. (2016). In-service Teacher Professional Learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 77-94). New York: Routledge.

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

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). New York: Routledge.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lipnevich, A. A., Berg, D. A. G., & Smith, J. K. (2016). Toward a Model of Student Response to Feedback. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 169-185). New York: Routledge.

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

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). New York: Routledge.

Murdock, T. B., Stephens, J. M., & Groteweil, M. M. (2016). Student dishonesty in the face of assessment: Who, why, and what we can do about it. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 186-203). New York: 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). New York: Routledge.

OECD. (2011). *Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States.* Paris, FR: OECD Publishing.

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 247-266). New York: Routledge.

Smith, L. F., Hill, M. F., Cowie, B., & Gilmore, A. (2014). Preparing Teachers to Use the Enabling Power of Assessment. In C. M. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing Assessment for Quality Learning* (pp. 303-323). Dordrecht, NL: Springer.

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

Strijbos, J. W. (2016). Assessment of collaborative learning. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 302-318). New York: Routledge.

Struyven, K., & Devesa, J. (2016). Students' perceptions of novel forms of assessment In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 129-144). New York: Routledge.

Suzuki, L. K., Davis, H. M., & Greenfield, P. M. (2008). Self-Enhancement and Self-Effacement in Reaction to Praise and Criticism: The Case of Multiethnic Youth. *Ethos, 36*(1), 78-97. 10.1111/j.1548-1352.2008.00005.x

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

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). New York: Routledge.

Tierney, R. D., & Koch, M. J. (2016). Privacy in classroom assessment. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 267-283). New York: Routledge.

Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners.* Alexandria, VA: ASCD.

Van De Vijver, F. (2016). Assessment in education in multicultural populations. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 436-453). New York: Routledge.

Wise, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity: The influence of student 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. 187-205). Charlotte, NC: Information Age Publishing.

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). New York: Routledge.

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

Zumbo, B. D. (2015). *Consequences, side effects and the ecology of testing: Keys to considering assessment in Vivo*. Plenary address to the 2015 annual conference of the Association for Educational Assessment—Europe (AEA-E), Glasgow, Scotland.