An Integrated Approach to Assessment as a Co-ordinating Framework

Gavin T. L. Brown

The University of Auckland

Abstract

All assessments are events within a process that has the goal of making decisions about instruction, learning, curriculum, students, institutions, and consequences. Three underlying disciplines (i.e., psychometrics, psychology, and sociology) inform the evaluation of assessments. Error is ubiquitous in the selection of tasks that constitute an assessment, the administration, marking, reporting, and decisions contain a non-ignorable component of error that has to be mitigated. Psychological ego factors in the marker and the assessee can be maladaptive in generating responses to assessment demands; awareness of this validity threat is needed to support participants into adaptive effort. Environments, from cultural norms to government policies, create contexts that can contribute greater error and psychological ego into the system, especially when high-stakes accountability pressures, as opposed to low-consequence formative support, are implemented. I conclude with suggestions of how integration of these underlying disciplines can improve the credibility and quality of higher education assessment.

Keywords. Psychometrics, psychology, technology, reliability, validity, process

Introduction

All assessments consist of events that take place within a process. The goal of the 'design-administer-score-interpret' process is to make decisions about instruction, learning, curriculum, classroom organisation, and consequences. An integrated approach to assessment needs to keep an eye on the quality characteristics of events and threats to the legitimacy of processes. Assessment as an applied engineering response to problems does not have a robust

theoretical basis. However, it can be understood effectively by consideration of three theoretical domains. The first is psychometric and statistical research into estimating the accuracy of scores that presumes there is error in all measurements and estimations of quality or quantity. That awareness has been extended by psychological research into the human factors that impinge upon both the validity and reliability of assessments. It is clear the human with ego is a major threat to assessment processes. The third field, informed on social psychological or sociological theories, identifies the impact of environmental factors of policy, culture, and society upon the validity of assessments in education. In this chapter, I will attempt create a coherent framework that focuses on error, ego, and environment that impact on our understanding and use of assessments and assessment processes.

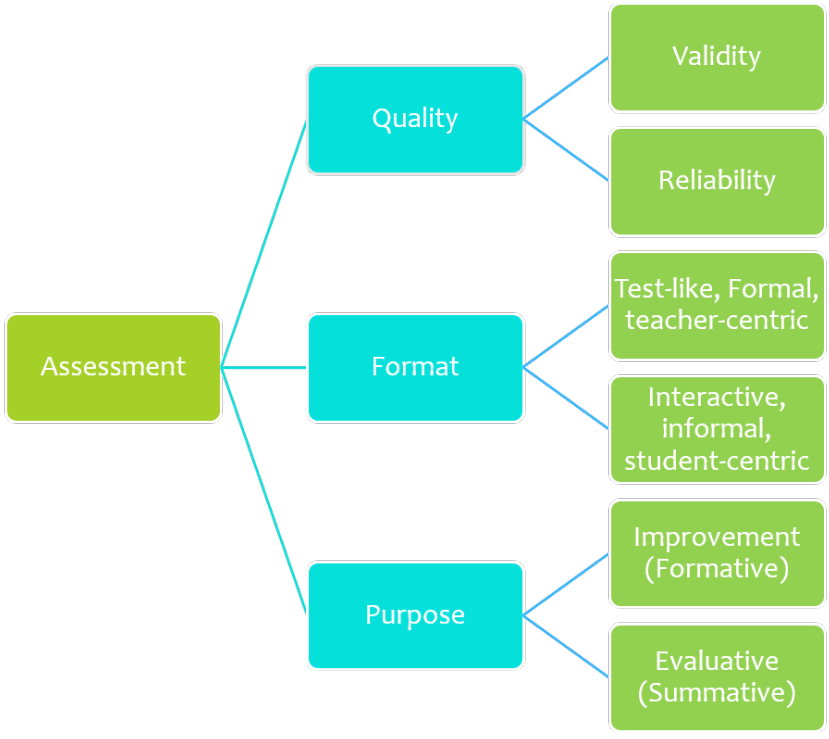
Integration

To establish an integrated framework, I consider it necessary to examine the features of assessment events within the overall process of conducting an assessment. Each separate assessment event is evaluated by three major concerns, displayed in Figure 1.

1. The quality of an assessment depends on its validity (Messick, 1989) which is the theoretical and empirical evidence that supports decisions about a learner based on their performance. Fundamentally, a valid assessment is one for which there is a persuasive argument (Kane, 2006) that how the assessment processes were implemented were sufficiently robust to give confidence to making decisions with the assessment information (Crooks, Kane, & Cohen, 1996). Understandably, the quality of scores arising from an assessment have to be reliable or consistent (Haertel, 2006); if another marker would not give very similar scores or grades, the decisions that follow are on shaky grounds. Hence, arguments and evidence for the validity and reliability of scoring for an assessment event have to be present.

2. Assessments come in a variety of formats ranging from formal examinations under invigilation to interactive student-involved practices such as peer and self-assessment. Each of these methods has strengths and weaknesses (Brown, 2018). In order to overcome weaknesses, multiple opportunities and samples of student performance need to be taken to arrive at defensible decision making (Brennan, 1996).
3. Colouring all of these decisions is the underlying purpose which assessment is meant to fulfil. Using Scriven's (1967) distinction, assessments can be formative (i.e., while instruction is being implemented with the intention of helping learners improve) or summative (i.e., terminal at the end of instruction with the intention of determining merit, worth, or value of performance).

Figure 1.

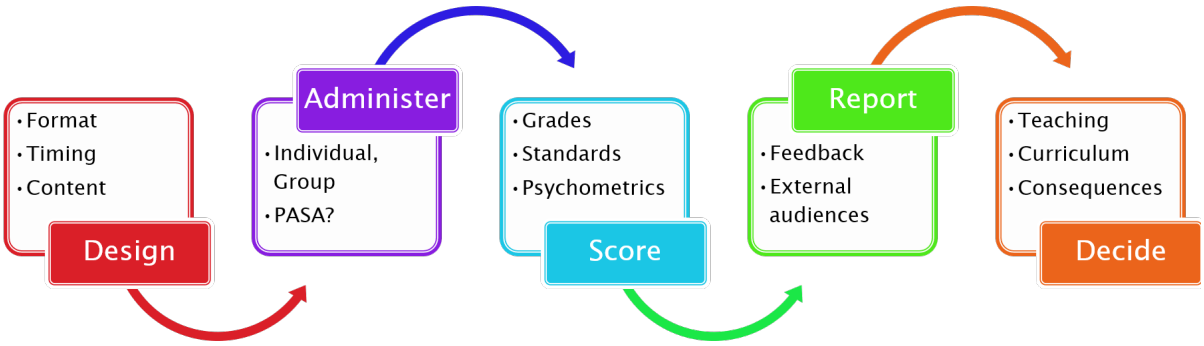


Key Characteristics of Assessment

Having outlined the characteristics of assessment, it is important to keep in mind that assessment is a process (Figure 2). Assessment begins with design decisions as to what an assessment will cover, how long it will be, and what format it will take. Then assessments

have to be administered with decisions made about whether it will be individual or group, invigilated or not, and whether it will be an interactive student-involved activity. Then, assessments have to be scored, marked, or at least read by either a machine or a human so as to generate an estimate of performance quality or quantity. Those results, whether numeric or commentaries, have to be reported to learners and administration in some form or another. Finally, decisions have to be made with that information, ranging from store it as a record, through using it to inform changes to the teaching/learning processes, to using it to evaluate the learner, the course, the instructor, or even the institution.

Figure 2.



The Assessment Process

Error

Error exists in all measurements, not just those of psychological, cognitive, or affective domains. The metre is meant to be 1/10,000,000th of the distance from the equator to the north pole; modern satellite-based measurements indicate that the accepted measure is actually 0.2mm short (Alder, 2002). Hence, all observations of student learning or performance are by definition imprecise.

The psychometric tradition assumes that all assessments are estimates of a person's true ability influenced by some amount of, hopefully random, error. Expressed formulaically, the classical test theory (Novick, 1966) states, that the observed score (O) consists of the true score (T) and a random error (e) which could inflate or decrease the observed score:

$$O = T +/- e \quad (1)$$

Validity errors occur when designs fail to generate robust evidence for the competence intended by the assessment (e.g., evaluating hair cutting by asking for an essay on types of hair; Brown, 2018). It is absolutely critical that the learning objectives, goals, or intentions shaping the design of any assessment be focused on the valuable aspects (Popham, 2000); there is no real justification for using instructional time by assessing trivial things. An important feature of value is ensuring that assessments require higher-order thinking skills alongside the important skills of remember and recall; a useful approach could be the adoption of the Structure of Observed Learning Outcome taxonomy (Brown, 2010a). Most importantly, the validity of any assessment design is restricted in that it is a sample of all tasks or questions that could possibly be posed about a domain. An assessment is rarely ever a census of the domain; in a constrained period of time it is not possible to get any assessee to do every aspect of a domain. While authenticity in task design, especially exploiting digital technologies used in the workplace (Jisc, 2020), works toward greater verisimilitude of assessments to work and life beyond education (Ashford-Rowe, Herrington, & Brown, 2013) any such assessments are, not only samples, but also conducted within the context of an educational institution with its formative orientation (Frey et al., 2012)—assessment in educational settings simply is not the real world.

The validity of assessments is clearly reduced if administration circumstances (e.g., disruptive noise) interfere with assessees' ability to demonstrate their best effort. Administration needs to follow recommended procedures for ensuring that performances are genuinely those of the person who is meant to be assessed; detecting cheating and ensuring security of digital assessments is essential (Dawson, 2021) if results are to be seen as credible.

The scoring or marking stage of any assessment introduces probably the largest source of error (Brennan, 1996). Markers are notoriously inconsistent, even with marking rubrics, within themselves (intra-marker) and between each other (inter-marker), especially with essays (Brown, 2009, 2010b). Between marking times or persons, results are rarely sufficiently robust enough to support decisions that have important consequences; recommended thresholds for decision making include: 70% identical scores, 90% equivalent scores (i.e., +/- 1), or correlations between score sets of $r \geq .70$ (Stemler, 2004).

The ability to concentrate on the important features of performance over multiple performances and over a long period of time is difficult to maintain. Marker consistency and validity is reduced by uncomfortable chairs, lack of food, consumption of alcohol, time of night, and mood. The discrepancy between how teachers judge student performance and how the same students perform on standardised tests makes clear that the accuracy of overall judgments is poor (Meissel, et al., 2017; Shah et al., 2020). Hence, markers need to (a) understand that error is inevitable as a human and (b) implement strategies and processes to minimise introduction of error and maximise consistency in the signal that their grades or marks give to learners and systems.

A number of strategies enhance consistency within and between markers. These include: 1) pre-specified marking rubrics, 2) annotated exemplars, 3) moderation between multiple markers, 4) marking one task or question across all candidates before proceeding to the next task/question, 5) self-testing one's own marking by re-scoring previously marked performances, and 6) training (Brown, 2018; Brown, Glasswell, & Harland, 2004). While humans struggle to achieve accurate scoring of complex performances, automated marking of essays has shown itself to be more consistent than human marking (Page, 1968; Shermis, 2014). With the advent of natural language processing (e.g., spoken commands to Google, Siri, or Alexa), the possibility of automated marking of essays is becoming real (Bulut, in

press). Nonetheless, all machine marking depends on learning from humans in order to mimic them; such mimicry may perpetuate human error in marking.

Having created marks, grades, or text commentary, that information has to be communicated to relevant end-users. Of course, the learner's own experience within the assessment event constitutes itself information that can tell the learner something about the quality of their work (Brown, 2022). The processes of selecting what to say, how to frame it, which modality to use, when to give it, and so on are not robustly agreed upon in the literature (Lipnevich et al., 2016), except for the requirement that feedback, in whatever form, be dependably accurate (Smith & Lipnevich, 2018). That implies that error in feedback, from its content to its delivery, is always possible and likely (e.g., too little/too much information; too harsh/lenient; too supportive/insufficient support, etc.). Indeed, the degree of inaccuracy in primary school teacher reports to parents is substantial (Hattie & Peddie, 2003; Robinson et al., 2003), meaning that parents as decision-makers were unable to do anything that might have helped their children. Given the enormity of factors impacting on feedback processes, it is not surprising that it is easy to get it wrong when reporting results or advising recipients.

The ultimate decision-making process itself, assuming that all the prior processes are defensible, is itself susceptible to error. While differential item functioning (Zumbo, 1999) and invariance (Brown et al., 2017) testing methods can detect bias in test scores, the validity of the consequences imposed on assessee populations requires a wider approach. If the proportion of people from different demographic groups awarded benefits or assigned to reduced opportunities has a +/-20% difference in membership, then there is evidence that the consequences are disproportionate (Biddle, 2005). To illustrate this approach, consider the National Research Council (2002) report that, as of 1998, Black Americans were more much more likely (OR = 2.24) to be assigned to special education compared to white children and were less likely to be assigned to gifted or talented programs (OR = 0.41). These

discrepancies show a real, not perceived, outcome difference that exceed the 20% threshold. This means that there can be error in consequential decisions, and unfortunately these tend to be visible especially for minority or indigenous children who are more likely to grow up in relative poverty.

Hence, throughout the assessment process, errors arise from multiple sources and these are generally not ignorable (Brown, 2017b). Work in the psychometric tradition to estimate the degree of error originated in evaluations of standardised educational and psychological tests (Lindquist, 1951). A wide panoply of statistical methods have been developed to account for and mitigate rater effects, establish validity of diagnostic sub-scores, and establish equivalence between groups and tests, as well as over time (Brennan, 2006). New technologies now permit assessments in new formats (e.g., game-based, simulations, and video), automated item development, more accurate ability estimation with adaptive testing, and delivery on mobile devices; while raising challenges for data and test security (Drasgow, 2016).

While reliability estimates of $r > .80$ are seen in published standardised tests or those used to evaluate entry into university or postgraduate professional programs, the quality of instructor-made tests, which are commonplace in higher education, is not so well attested. Quite often high-stakes examinations are not even evaluated for item or overall test quality; but, when done so even examinations that have been through thorough pre-administration judgment processes can be extremely poor quality (Brown & Abdulnabi, 2017). Nonetheless, efforts around estimating test quality ought to give confidence in the information generated by those tests and examinations.

However, efforts to estimate error in other forms of assessment are extremely difficult given that non-formal, student-involved, classroom interactions are so ephemeral and therefore not available to detailed inspection. Evaluations have examined student-involved

assessment practices such as self-assessment and peer assessment and calibrated them against test scores, tutor, or instructor marks. Self-assessments are generally only weakly consistent with other measures (Boud & Falchikov, 1989; Brown & Harris, 2013). Peer assessments in higher education have shown relatively consistently that student awarded scores tend to be somewhat higher than those awarded by tutors or instructors (Ashenafi, 2015). Despite the promise and allure of e-portfolio technologies as a basis for assessment, little evidence exists as yet that they are an effective strategy for eliciting high quality insights into student learning; their relative novelty in part contributes to this dearth of research (Struyven & Devesa, 2016). Nonetheless, students who hold positive views of e-portfolio as a way to improve their learning seem to have greater intention to use the technology that way, which contributes to greater self-reported academic performance (Deneen et al., 2018).

An interesting conundrum of course in judging these student-involved judgments relates to if there is a gold-standard by which self- or peer reports should be judged. If we accept that the 'self' knows the 'self' best, then on what grounds might we consider the self-assessment to be in error? Given the status of learners as relative novices in their subject domains, it might not be surprising at all that their judgments are strongly impacted by other psychological and social factors which I will address in the next section.

Ego

The psychology of teachers and students under, within, and in response to assessment should be an obvious feature of any integrated framework. Assessments have consequences for both teachers and students. For the latter assessments, are used summatively to accredit, fail, award scholarships, and so on (Newton, 2007); among the former, the results can be used, often under freedom of information principles, to create league tables according to success rates. Consequently, the public (Buckendahl, 2016) and officials use that data to evaluate the quality of teachers and schools. Even when the purpose of official response is to provide

support to schools, this can be seen as a negative outcome by school leaders (Ngan, Lee, & Brown, 2010). It is worth noting that international large-scale assessment programs (e.g., PISA, TIMSS) are used by the media, the public, and government officials to evaluate and change system processes (e.g., teacher qualifications, curriculum, etc.) (Teltemann & Klieme, 2016). This is done despite extensive evidence that the scores of such tests lack psychometric equivalence across multiple jurisdictions (Asil & Brown, 2016).

With educational assessment results being used to judge or evaluate teachers and schools (Nichols & Harris, 2016), it is not surprising that, where consequences are imposed by authorities, teachers experience substantial tension around using assessments for administrative reasons instead of for improvement-oriented goals (Bonner, 2016). The obvious reason for this lies in the power of accountability to inspire compliance with the expectations of superiors (Lerner & Tetlock, 1999). By maximising student test scores against which they will be judged, teachers are protecting themselves from what could arguably be seen as inappropriate measures of their competence and efficacy (Cannell, 1989; Ravitch, 2013). Participant positive perceptions of organisational justice plays a significant part in subsequent work performance (Konradt et al., 2017) and if subordinates perceive unethical conduct from their superiors or perceive that the institution is unfair, then less ethical behaviour arises (Cropanzano & Stein, 2009) . Hence, an understandable response to unfair consequences arising from student testing is to exercise ego-protection by falsifying results; *the consequences are not fair, so inflating scores to ensure we are acceptable is valid*.

Teachers are not alone in responding to assessment with ego-protective attitudes and behaviours. The dual cycle model of self-regulation of learning (Boekaerts & Corno, 2005; Boekaerts & Niemivirta, 2000) provides a framework around which student responses to assessment can be understood. One cycle is growth-oriented in which students embrace

information, even when it is critical or negative, and implement appropriate strategies and approaches that improve their learning outcomes. In contrast, the ego-protective cycle stimulates maladaptive motivations, attributions, and behaviours, which shut off the 'self' from criticism, but which perpetuate poor performance. Empirical evidence exists for deleterious effects of ego-protective psychological characteristics; for example, external as opposed to internal attribution (Weiner, 1985), performance as opposed to mastery motivation (Elliot & Dweck, 1988), less as opposed to greater confidence in one's ability to perform a task (Bandura, 1982), just enough versus doing my best effort (Meyer et al., 2009), fixed versus malleable beliefs about intelligence (Mangels et al., 2006), and surface versus deep or transformational beliefs and approaches to learning (Entwistle, 1991).

The human tendency to protecting the 'self' from perceived harm arising from negative results is complicated by the ignorance most learners have about the quality of their own work, how to improve, and how to exercise their cognitive and metacognitive strategies to achieve maximal outcomes. Realism in recognising the quality of one's own work or that of a peer is a complex task which requires expertise and courage to admit one's own faults. Unfortunately, humans are notoriously bad at knowing the truth of how good or poor their work is (Dunning et al., 2004), even though without that knowledge appropriate improvement cannot be identified, let alone implemented. This intra-personal tendency to have an unrealistic evaluation of one's own work has led us to conclude that self-assessment can be considered a valuable curricular goal (i.e., evaluative judgment; Tai et al., 2018) but should not be treated as useful assessment information (Brown & Harris, 2014). Inter-personal factors impinge on our ability to tell to or receive the truth from a peer; there has to be trust, humility, and shared values between and among peers involved in assessment around a growth or learning goal (Panadero, 2016). Otherwise, students have shown that they will

dissemble and outright lie in order to protect their relationships and their 'self' (Harris et al., 2018).

While this section speaks to the challenge assessed performance or learning have for the ego of both teachers and students, it should be borne in mind that much of this 'ego' effect is done invisibly. In part the lack of visibility comes from lack of awareness of or sensitivity to the many adaptive things that capable learners implement intuitively. Highly capable learners know how to cope with the challenges of assessment (Brown & Eklöf, 2018) and feedback (Brown et al., 2016). Nevertheless, sensitivity to the social and psychological factors in how students learn is necessary. This extends to how teachers are expected to cope with the accountability demands put on them. What may seem like maladaptive responses, beliefs, attitudes, or behaviours may simply be appropriate and rational responses to the pressures of the environment in which assessment takes place; the matter which I turn to next.

Environment

Humans exist in societies that differ in how schooling is understood (Brown & Harris, 2009). Those cultures become contexts that mould educational stakeholders (i.e., families, children, and teachers) through socialisation into the normative beliefs, attitudes, and practices of a context (Rogoff, 1991). At the same time, forms of government determine policies and norms that set the educational sector conditions which strongly frame traditions and conventions in how teaching, learning, curriculum, and assessment take place (Bourdieu & Passeron, 1976).

An insightful frame for cultural impact upon education comes from LeVine and White's (1986) life chances analysis around two key principles embedded into societal choices; that is, options and ligatures. They argue that economically developed societies prioritise giving children options to pursue study and careers that require advanced levels of literacy and numeracy as preparation for jobs that may not yet exist; this means in general that with educational success, young people do not follow the same careers of their parents,

nor do they always end up living nearby. In contrast, feudal, agrarian, and highly traditional societies privilege education that emphasises belonging to and identity within the society; in this case, children only need a few years of schooling to master basic literacy and numeracy and then are inducted in apprenticeship fashion into careers with their relatives.

In light of this, indigenous and minority children often do not experience assessments as a positive experience (Bishop & Berryman, 2006) and may resist loss of their own identity by resistance against assessment of achievement (Fordham & Ogbu, 1986). Other evidence of the impact of cultural references on achievement can be seen in the stereotype threat in which members of a usually less academically successful group are reminded of that status before an assessment and consequently perform worse than when not prompted with the stereotype (Steele & Aronson, 1995). Jones (1991) demonstrated that there were very different teaching, learning, and assessment preparation activities between white girls from privileged homes and Pacific Island girls; the former learned by debate and discussion with the teacher, while the latter learned by memorisation and regurgitation. These brief examples show that children are socialised into an understanding of what assessment is, what it is for, and how one prepares and responds to it.

The greater number and significance of further life opportunities (e.g., enrolment in senior secondary school, access to scholarships, enrolment in university) controlled by assessments, the greater power assessment has to shape stakeholders. The experiences of progress through assessment systems creates what Ecclestone and Pryor (2003) call an ‘assessment career’ in contrast to a ‘learning career’. Societies that privilege high-stakes examinations as a selection tool (e.g., East Asia; Islamic nations; Latin America, South Asia; Brown, 2017c) tend to generate environments in which assessment is solely summative, teaching is highly didactic, and performance depends on repeated practice, persistence, and recall.

Likewise, teachers can be socialised into a summative examination mindset, especially where qualifications systems require teachers to act as assessors (Brown, 2011; Brown, Lake, & Matters, 2011), despite their best intentions to prioritise formative assessment and feedback practices (Brown, 2004; Brown et al., 2012). In jurisdictions that prioritise learning rather than ranking, teachers seek to implement an assessment for learning pedagogy, which is good teaching but poor assessment for the reasons of error and ego that I have reviewed already (Brown, 2020). Assessment, instead of being learning oriented, becomes a convergent classroom approach eliciting responses that the teacher already knows (Torrance & Pryor, 1998). The Anglo-American trend of the last 30 years of using assessments to evaluate schools and teachers and privatise educational development has been borrowed as a policy in many nations (Lingard & Lewis, 2016), with some positive effects when the results align with local priorities (Teltemann & Klieme, 2016).

Across these contrasting environments, assessments take on quite different roles, functions, and effects. The differences in environments seem to devolve to different patterns: (a) conventional, summative, high-stakes public examination jurisdictions where the onus is on the learner, (b) more progressive, formative, learning-oriented assessment systems that attempt to balance student self-regulated learning and improved instruction with summative reporting and decision making, and (c) almost non-existent, despite advocacy (Boud & Associates, 2010; Tan, 2012; Taras, 2012), pure formative systems that let student choices about learning determine the purposes of assessment. In terms of global populations, the most common pattern is the high-stakes public examination approach (Brown, 2017c), but all higher education jurisdictions use assessment to make high-stakes judgments that impact student futures.

Application:

There is a continuing and ever-present need in any approach to assessment in higher education to pay attention to the quality requirements any assessment event and consider how it fits within the educational process. Institutions need to be aware of the presence of error and the impact of ego-related psychological factors in performance and scoring within their own specific environment and its requirements. The materials outlined here have to be appropriated according to the environment in which academics work. This review provides insights that should caution us to follow robust processes for ensuring high quality decision making.

Robustness comes from evaluating all assessments with appropriate psychometric methods and appropriate expert judgments; given how few academics are trained or expert in assessments and how poorly we teach psychometrics (Brown, 2017a) or deploy psychometric tools (Abbasnasab Sardareh et al., 2021), it is not surprising that assessment is done poorly. This is a matter which many institutions have glossed over, despite the need to demonstrate to all stakeholders that the decisions arising from their assessments are robustly reliable and valid. Improving academic staff capacity in assessment is necessary, not just because online assessment has become normative under the covid pandemic, but because the reputation of the institution depends on the defensibility of assessment-based decisions. Institutions cannot just place responsibility on academics to do this well without providing better resources and training in assessment design, marking and scoring, psychometric analysis, and feedback processes.

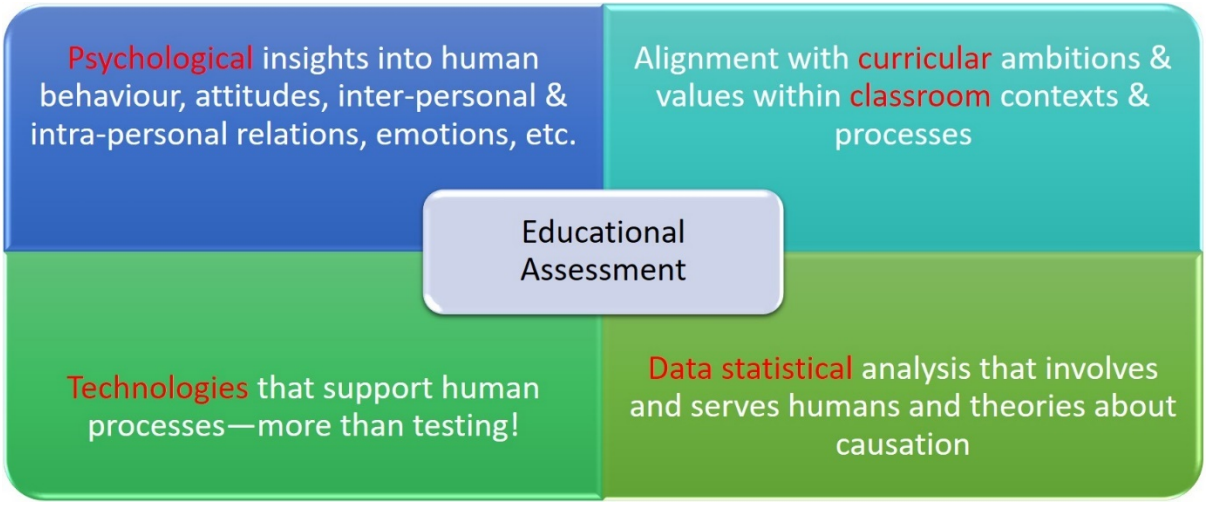
Key messages/considerations for advancing the field in the future:

In thinking about where educational assessment needs to go, it seems that the greater integration, rather than siloed isolation, is needed among the foundational disciplines of assessment (Figure 3). A greater understanding of the maladaptive and adaptive psychological factors impinging upon instructors, markers, and assessees needs to be

integrated with a greater attention to ensuring that we assess the valuable things that our students need now and into the beginning of life beyond higher education. At the same time, we need greater and more simplified integration of advanced psychometric and data analytic tools to ensure that scores, grades, or marks mean what we want them to mean (i.e., people with higher marks actually know or can do more than those with less and that such discrimination is greater than would occur by chance).

This call means paying attention to the quality of assessment events and processes that might invalidate decisions. Generally, systems need to administer multiple assessments, using multiple formats, and score with multiple judges. At the same time, students must not be overwhelmed with the number of assessments, but at the same time enough high-quality information has to be available to support decisions. Use of programmatic portfolios of performance over the lifespan of a degree seems to be a way to reduce the burden on a single assessment event and better align graduate attributes or competencies (van der Vleuten et al., 2012). Although developed in medical education, the emphasis on tracking assessment and feedback information in a portfolio which requires student reflections and recorded discussions with mentors seems to allow rich description of student development across a range of valued competencies (Schuwirth & van der Vleuten, 2019). Concentration upon graduate competencies or employability skills as an over-arching approach to reporting assessment results may well provide a mechanism for getting away from a near-sighted focus on test or examination scores (Brown et al., 2019).

Figure 3.



Integrated Assessment Framework

Technologies that integrate these functionalities are being rapidly deployed, though rarely at the level of simplicity that allow lay academics to exploit. One project we are working on exploits open-source item response theory libraries in R in a simple interface that allows removal of poor-quality items and the setting of grade standards, so that scores can be converted into institutionally-defined grades (Brown et al., 2021 in review). Other technological possibilities that can ease academic workload and improve learning include (a) clickers which allow students to actively participate in lectures and demonstrate mastery recall of key points (Camacho- Miñano & del Campo, 2016), (b) the prize-winning international success of PeerWise which allows students to create multiple-choice questions and take advantage of self-testing as a learning technique (Denny et al., 2009), and (c) some technology enabled feedback systems have shown promise to speed up quality feedback to learners (Munshi & Deneen, 2018). The generality of the last point reminds us that there are multiple technological resources seeking our financial investment and energy, but few achieve what we need in assessment. Keeping our eye on this integrated framework will allow us to develop policies and practices that ensure valid decision making about student learning and performance.

Suggested Readings .

1. Brown, G., & Harris, L. R. (Eds.). (2016). *Handbook of Human and Social Conditions in Assessment*. Routledge.

   This collection of 29 chapters brings together extensive research on how assessment policies, practices, and technologies impact both students and teachers in cross-cultural contexts.

2. Pitoniak, M., & Cook, L. (Eds.). (2022). *Educational Measurement* (5th ed.). National Council for Measurement in Education.

   The forthcoming 5th edition of the pre-eminent handbook of educational measurement provides extensive and authoritative insights into psychometric methods of evaluating all formal and informal assessment formats.

3. Drasgow, F. (Ed.). (2016). *Technology and Testing: Improving Educational and Psychological Measurement*. Routledge.

   This volume, published as part of the National Council for Measurement in Education book series, explains how various technologies are being developed, evaluated, and deployed to create high-quality tests.

4. Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.

   This volume is at the cutting-edge of review and theorisation about how digital assessments can be kept secure so that valid decisions are made about student learning.

References

Abbasnasab Sardareh, S., Brown, G. T. L., & Denny, P. (2021). Comparing four contemporary statistical software tools for introductory data science and statistics in the social sciences. *Teaching Statistics*, *43*(S1), S157-S172. https://doi.org/https://doi.org/10.1111/test.12274

Alder, K. (2002). *The measure of all things: The seven-year odyssey that transformed the world*. London: Abacus.

Ashenafi, M. M. (2015). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education*, *42*(2), 226-251. https://doi.org/10.1080/02602938.2015.1100711

Ashford-Rowe, K., Herrington, J., & Brown, C. (2013). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*, 205-222. https://doi.org/10.1080/02602938.2013.819566

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of Non-Invariance. *International Journal of Testing*, *16*(1), 71-93. https://doi.org/10.1080/15305058.2015.1064431

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122-147. https://doi.org/doi:10.1037/0003-066X.37.2.122

Biddle, D. (2005). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing* (2nd, Ed.). Ashgate.

Bishop, R., & Berryman, M. (2006). *Culture speaks: Cultural relationships & classroom learning*. Huia.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An international review*, *54*(2), 199-231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Boekaerts, M., & Niemivirta, M. (2000). Self-regulated learning: Finding a balance between learning goals and ego-protective goals. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 417-450). Academic 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). Routledge.

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

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, *18*(5), 529-549. https://doi.org/10.1007/BF00138746

Bourdieu, P., & Passeron, J. C. (1976). *Reproduction in education society and culture*. Sage Publications.

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). National Center for Education Statistics.

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

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). Ako Aotearoa.

Brown, G. T. L. (2010a). Assessment: Principles and practice. In R. H. Cantwell & J. J. Scevak (Eds.), *An academic life: A handbook for new academics* (pp. 35-44). ACER Press.

Brown, G. T. L. (2010b). 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. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45-70. https://doi.org/10.18296/am.0097

Brown, G. T. L. (2017a). Doctoral education in quantitative research methods: Some thoughts about preparing future scholars. *Frontiers in Applied Mathematics & Statistics*, *3*, 1-4. https://doi.org/doi: 10.3389/fams.2017.00025

Brown, G. T. L. (2017b). 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. (2017c). What we know we don't know about teacher education. In D. J. Clandinin & J. Husu (Eds.), *The SAGE International Handbook of Research on Teacher Education* (pp. 123-138). Sage.

Brown, G. T. L. (2018). *Assessment of Student Achievement*. Routledge.

Brown, G. T. L. (2020). Responding to Assessment for Learning: A pedagogical method, not assessment. *New Zealand Annual Review of Education*, *26*, 18-28. https://doi.org/10.26686/nzaroe.v26.6854

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

Brown, G. T. L., & Abdulnabi, H. (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., & Eklöf, H. (2018). Swedish student perceptions of achievement practices: The role of intelligence. *Intelligence*, *69*, 94--103. https://doi.org/10.1016/j.intell.2018.05.006

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation*, *6*(12), 68-91. http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/236

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). Sage.

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., Cooper-Thomas, H., Curtin, J., Dunham, A., Geertshuis, S., Knewstubb, B., Lees-Marshment, J., Lewis, N., Liu, Q., Kensington-Miller, B., Mobberley, P., Sturm, S., & Wass, R. (2019). *Teaching for transformation: The 4E's of teaching future-ready graduates.* HERDSA annual conference, Auckland, New Zealand. https://bit.ly/2Vfs8FL

Brown, G. T. L., Denny, P., San Jose, D. L. & Li, E. (2021, in review). *Setting standards with multiple-choice tests: A preliminary intended-user evaluation of SmartStandardSet*. Manuscript submitted for publication.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105-121. https://doi.org/10.1016/j.asw.2004.07.001

Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education*, *28*(7), 968-978. https://doi.org/10.1016/j.tate.2012.05.003

Brown, G. T. L., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, *40*(1), 66-90. https://doi.org/10.1080/1743727X.2015.1070823

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

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology*, *86*(4), 606-629. https://doi.org/10.1111/bjep.12126

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). Routledge.

Bulut, O. (in press). Beyond multiple-choice with digital assessments. *eLearn Magazine*. https://doi.org/10.1145/3472394

Camacho-Miñano, M.-d.-M., & del Campo, C. (2016). Useful interactive teaching tool for learning: clickers in higher education. *Interactive Learning Environments*, *24*(4), 706-723. https://doi.org/10.1080/10494820.2014.917108

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

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, *3*, 265-285. https://doi.org/10.1080/0969594960030302

Cropanzano, R., & Stein, J. H. (2009). Organizational Justice and Behavioral Ethics: Promises and Prospects. *Business Ethics Quarterly*, *19*(2), 193-233. https://doi.org/10.5840/beq200919211

Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.

Deneen, C. C., Brown, G. T. L., & Carless, D. (2018). Students' conceptions of eportfolios as assessment and technology. *Innovations in Education and Teaching International*, *55*(4), 487-496. https://doi.org/doi:10.1080/14703297.2017.1281752

Denny, P., Luxton-Reilly, A., & Simon, B. (2009, January). Quality of student contributed questions using PeerWise. *Conferences in Research and Practice in Information Technology*, *95*, 55-63, Eleventh Australasian Computing Education Conference (ACE2009), Wellington, New Zealand.

Drasgow, F. (Ed.). (2016). *Technology and Testing: Improving Educational and Psychological Measurement*. Routledge.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Ecclestone, K., & Pryor, J. (2003). 'Learning careers' or 'Assessment careers'? The impact of assessment systems on learning. *British Educational Research Journal*, *29*(4), 471-488. https://doi.org/10.1080/01411920301849

Elliott, E., & Dweck, C. S. (1988). Goals. *Journal of Personality and Social Psychology*, *54*(1), 5-12. https://doi.org/10.1037/0022-3514.54.1.5

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment: Introduction to the special issue. *Higher Education*, *22*, 201-204. https://doi.org/10.1007/BF00132287

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white'". *The Urban Review*, *18*, 176- 206. https://doi.org/10.1007/BF01112192

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining Authentic Classroom Assessment. *Practical Assessment, Research & Evaluation*, *17*(2). https://doi.org/10.7275/sxbs-0829

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Praeger.

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., & Peddie, R. (2003). School reports: "Praising with faint damns". *set: research information for teachers*(3), 4-9. https://doi.org/10.18296/set.0710

Jisc. (2020). *The future of assessment: Five principles, five targets for 2025*. Retrieved from Bristol, UK: https://www.jisc.ac.uk/reports/the-future-of-assessment

Jones, A. (1991). *At school I've got a chance. Culture/privilege: Pacific Islands and Pākehā girls at school.* Dunmore Press.

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

Konradt, U., Garbers, Y., Böge, M., Erdogan, B., & Bauer, T. N. (2017). Antecedents and Consequences of Fairness Perceptions in Personnel Selection: A 3-Year Longitudinal

Study. *Group & Organization Management*, *42*(1), 113-146. https://doi.org/10.1177/1059601115617665

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

LeVine, R. A., & White, M., I. (1986). *Human conditions: The cultural basis of educational developments.* Routledge & Kegan Paul.

Lindquist, E. F. (Ed.). (1951). *Educational Measurement*. American Council on Education.

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). Routledge.

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.), *The Handbook of Human and Social Conditions in Assessment* (pp. 169-185). Routledge.

Lovett, S., & Sinclair, L. (2005). *The socialisation of teachers into the culture of assessment*. New Zealand Council for Educational Research.

Mangels, J. A., Butterfield, B., Lamb, J., Good, C., & Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social Cognitive and Affective Neuroscience*, *1*(2), 75-86. https://doi.org/10.1093/scan/nsl013

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/doi:10.1016/j.tate.2017.02.021

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

Meyer, L. H., McClure, J., Walkey, F., Weir, K. F., & McKenzie, L. (2009). Secondary student motivation orientations and standards-based achievement outcomes. *British Journal of Educational Psychology*, *79*(2), 273-293. https://doi.org/10.1348/000709908X354591

Munshi, C., & Deneen, C. C. (2018). Technology-Enhanced Feedback. In J. K. Smith & A. A. Lipnevich (Eds.), *The Cambridge Handbook of Instructional Feedback* (pp. 335-356). Cambridge University Press. https://doi.org/10.1017/9781316832134.017

National Research Council. (2002). *Minority Students in Special and Gifted Education*. The National Academies Press. https://doi.org:10.17226/10128

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149-170. https://doi.org/10.1080/09695940701478321

Ngan, M. Y., Lee, J. C. K., & Brown, G. T. L. (2010). Hong Kong principals' perceptions on changes in evaluation and assessment policies: They're not for learning. *Asian Journal of Educational Research and Synergy*, *1*(36-46).

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). Routledge.

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

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210-225. https://doi.org/10.1007/BF01419938

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). Routledge.

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

Ravitch, D. (2013). *Reign of Error: The hoax of the privatization movement and the danger to America's public schools*. A. E. Knopf.

Robinson, V., Timperley, H., Ward, L., Tuioto, L., Stevenson, V. T., & Mitchell, S. (2004). *Strengthening Education in Mangere and Otara Evaluation: Final Evaluation Report* [Commissioned report to the Ministry of Education]. https://www.educationcounts.govt.nz/__data/assets/pdf_file/0006/9285/semo.pdf

Rogoff, B. (1991). The joint socialization of development by young children and adults. In P. Light, S. Sheldon, & M. Woodhead (Eds.), *Learning to Think: Child Development in Social Context 2* (pp. 67-96). Routledge.

Schuwirth, L., & van der Vleuten, C. (2019). Current Assessment in Medical Education: Programmatic Assessment. *Journal of Applied Testing Technology*, *20*(S2), 2-10. http://jattjournal.net/index.php/atp/article/view/143673

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). Rand McNally.

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

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *39*(7), 4-14. https://doi.org/10.3102/0013189X029007004

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76. https://doi.org/10.1016/j.asw.2013.04.001

Smith, J. K., & Lipnevich, A. A. (2018). Instructional Feedback: Analysis, Synthesis, and Extrapolation. In J. K. Smith & A. A. Lipnevich (Eds.), *Cambridge handbook of instructional feedback* (pp. 591-603). Cambridge University Press.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, *67*, 797-811. https://doi.org/10.1037//0022-3514.69.5.797

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

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). Routledge.

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

Tan, K. H. K. (2012). *Student self-assessment. Assessment, learning and empowerment*. Research Publishing Services.

Taras, M. (2010). Student self-assessment: Processes and consequences. *Teaching in Higher Education*, *15*(2), 199-209. https://doi.org/10.1080/13562511003620027

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). Routledge.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom.* Open University Press.

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*(3), 205-214. https://doi.org/10.3109/0142159X.2012.652239

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548-573. https://doi.org/10.1037/0033-295X.92.4.548

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Directorate of Human Resources Research and Evaluation, Department of National Defense. http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf