frontiers | Frontiers in Education

TYPE Conceptual Analysis
PUBLISHED 11 November 2022
DOI 10.3389/feduc.2022.1060633

OPEN ACCESS

EDITED BY
Lan Yang,
The Education University of Hong Kong,
Hong Kong SAR, China

REVIEWED BY
Bruno D. Zumbo,
University of British Columbia, Canada
Juliette Lyons-Thomas,
Educational Testing Service,
United States

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

SPECIALTY SECTION
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

RECEIVED 03 October 2022
ACCEPTED 25 October 2022
PUBLISHED 11 November 2022

CITATION
Brown GTL (2022) The past, present and
future of educational assessment: A
transdisciplinary perspective.
Front. Educ. 7:1060633.
doi: 10.3389/feduc.2022.1060633

The past, present and future of educational assessment: A transdisciplinary perspective

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

[1]Department of Applied Educational Sciences, Umeå Universitet, Umeå, Sweden, [2]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand

To see the horizon of educational assessment, a history of how assessment has been used and analysed from the earliest records, through the 20th century, and into contemporary times is deployed. Since paper-and-pencil assessments validity and integrity of candidate achievement has mattered. Assessments have relied on expert judgment. With the massification of education, formal group-administered testing was implemented for qualifications and selection. Statistical methods for scoring tests (classical test theory and item response theory) were developed. With personal computing, tests are delivered on-screen and through the web with adaptive scoring based on student performance. Tests give an ever-increasing verisimilitude of real-world processes, and analysts are creating understanding of the processes test-takers use. Unfortunately testing has neglected the complicating psychological, cultural, and contextual factors related to test-taker psychology. Computer testing neglects school curriculum and classroom contexts, where most education takes place and where insights are needed by both teachers and learners. Unfortunately, the complex and dynamic processes of classrooms are extremely difficult to model mathematically and so remain largely outside the algorithms of psychometrics. This means that technology, data, and psychometrics have become increasingly isolated from curriculum, classrooms, teaching, and the psychology of instruction and learning. While there may be some integration of these disciplines within computer-based testing, this is still a long step from where classroom assessment happens. For a long time, educational, social, and cultural psychology related to learning and instruction have been neglected in testing. We are now on the cusp of significant and substantial development in educational assessment as greater emphasis on the psychology of assessment is brought into the world of testing. Herein lies the future for our field: integration of psychological theory and research with statistics and technology to understand processes that work for learning, identify how well students have learned, and what further teaching and learning is needed. The future requires greater efforts by psychometricians, testers, data analysts, and technologists to develop solutions that work in the pressure of living classrooms and that support valid and reliable assessment.

KEYWORDS

assessment, testing, technology, psychometrics, psychology, curriculum, classroom

Introduction

In looking to the horizon of educational assessment, I would like to take a broad chronological view of where we have come from, where we are now, and what the horizons are. Educational assessment plays a vital role in the quality of student learning experiences, teacher instructional activities, and evaluation of curriculum, school quality, and system performance. Assessments act as a lever for both formative improvement of teaching and learning and summative accountability evaluation of teachers, schools, and administration. Because it is so powerful, a nuanced understanding of its history, current status, and future possibilities seems a useful exercise. In this overview I begin with a brief historical journey from assessments past through the last 3000 years and into the future that is already taking place in various locations and contexts.

The past

Early records of the Chinese Imperial examination system can be found dating some 2,500 to 3,000 years ago (China Civilisation Centre, 2007). That system was used to identify and reward talent wherever it could be found in the sprawling empire of China. Rather than rely solely on recommendations, bribery, or nepotism, it was designed to meritocratically locate students with high levels of literacy and memory competencies to operate the Emperor's bureaucracy of command and control of a massive population. To achieve those goals, the system implemented standardised tasks (e.g., completing an essay according to Confucian principles) under invigilated circumstances to ensure integrity and comparability of performances (Feng, 1995). The system had a graduated series of increasingly more complex and demanding tests until at the final examination no one could be awarded the highest grade because it was reserved for the Emperor alone. Part of the rationale for this extensive technology related to the consequences attached to selection; not only did successful candidates receive jobs with substantial economic benefits, but they were also recognised publicly on examination lists and by the right to wear specific colours or badges that signified the level of examination the candidate had passed. Unsurprisingly, given the immense prestige and possibility of social advancement through scholarship, there was an industry of preparing cheat materials (e.g., miniature books that replicated Confucian classics) and catching cheats (e.g., ranks of invigilators in high chairs overlooking desks at which candidates worked; Elman, 2013).

In contrast, as described by Encyclopedia Brittanica (2010a), European educational assessment grew out of the literary and oratorical remains of the Roman empire such as schools of grammarians and rhetoricians. At the same time, schools were formed in the various cathedrals, monasteries (especially, the Benedictine monasteries), and episcopal schools throughout Europe. Under Charlemagne, church priests were required to master Latin so that they could understand scripture correctly, leading to more advanced religious and academic training. As European society developed in the early Renaissance, schools were opened under the authority of a bishop or cathedral officer or even from secular guilds to those deemed sufficiently competent to teach. Students and teachers at these schools were given certain protection and rights to ensure safe travel and free thinking. European universities from the 1100s adopted many of the clerical practices of reading important texts and scholars evaluating the quality of learning by student performance in oral disputes, debates, and arguments relative to the judgement of higher ranked experts. The subsequent centuries added written tasks and performances to the oral disputes as a way of judging the quality of learning outcomes. Nonetheless, assessment was based, as the Chinese Imperial system, on the expertise and judgment of more senior scholars or bureaucrats.

These mechanisms were put in place to meet the needs of society or religion for literate and numerate bureaucrats, thinkers, and scholars. The resource of further education, or even basic education, was generally rationed and limited. Standardised assessments, even if that were only the protocol rather than the task or the scoring, were carried out to select candidates on a relatively meritocratic basis. Families and students engaged in these processes because educational success gave hope of escape from lives of poverty and hard labour. Consequently, assessment was fundamentally a summative judgement of the student's abilities, schooling was preparation for the final examination, and assessments during the schooling process were but mimicry of a final assessment.

With the expansion of schooling and higher education through the 1800s, more efficient methods were sought to the workload surrounding hearing memorized recitations (Encyclopedia Brittanica, 2010b). This led to the imposition of leaving examinations as an entry requirement to learned professions (e.g., being a teacher), the civil service, and university studies. As more and more students attended universities in the 1800s, more efficient ways collecting information were established, most especially the essay examination and the practice of answering in writing by oneself without aids. This tradition can still be seen in ordered rows of desks in examination halls as students complete written exam papers under scrutiny and time pressure.

The 20th century

By the early 1900s, however, it became apparent that the scoring of these important intellectual exercises was highly problematic. Markers did not agree with each other nor were they consistent within themselves across items or tasks and over time so that their scores varied for the same work. Consequently, early in the 20th century, multiple-choice question tests were developed so that there would be consistency in scoring and efficiency in administration (Croft and Beard, 2022). It is also worth noting that considerable cost and time efficiencies were obtained through using multiple-choice test methods. This aspect led, throughout the century, to increasingly massive use of standardised machine scoreable tests for university entrance, graduate school selection, and even school evaluation. The mechanism of scoring items dichotomously (i.e., right or wrong), within classical test theory statistical modelling, resulted in easy and familiar numbers (e.g., mean, standard deviation, reliability, and standard error of measurement; Clauser, 2022).

As the 20th century progressed, the concepts of validity have grown increasingly expansive, and the methods of validation have become increasingly complex and multi-faceted to ensure validity of scores and their interpretation (Zumbo and Chan, 2014). These included scale reliability, factor analysis, item response theory, equating, norming, and standard setting, among others (Kline, 2020). It is worth noting here that statistical methods for test score analysis grew out of the early stages of the discipline of psychology. As psychometric methods became increasingly complex, the world of educational testing began to look much more like the world of statistics. Indeed, Cronbach (1954) noted that the world of psychometrics (i.e., statistical measurement of psychological phenomena) was losing contact with the world of psychology which was the most likely user of psychometric method and research. Interestingly, the world of education makes extensive use of assessment, but few educators are adept at the statistical methods necessary to evaluate their own tests, let alone those from central authorities. Indeed, few teachers are taught statistical test analysis techniques, even fewer understand them, and almost none make use of them.

Of course, assessment is not just a scored task or set of questions. It is legitimately an attempt to operationalize a sample of a construct or content or curriculum domain. The challenge for assessment lies in the conundrum that the material that is easy to test and score tends to be the material that is the least demanding or valuable in any domain. Learning objectives for K-12 schooling, let alone higher education, expect students to go beyond remembering, recalling, regurgitating lists of terminology, facts, or pieces of data. While recall of data pieces is necessary for deep processing, recall of those details is not sufficient. Students need to exhibit complex thinking, problem-solving, creativity, and analysis and synthesis. Assessment of such skills is extremely complex and difficult to achieve.

However, with the need to demonstrate that teachers are effective and that schools are achieving society's goals and purposes it becomes easy to reduce the valued objectives of society to that which can be incorporated efficiently into a standardised test. Hence, in many societies the high-stakes test becomes the curriculum. If we could be sure that what was on the test is what society really wanted, this would not be such a bad thing; what Resnick and Resnick (1989) called measurement driven reform. However, research over extensive periods since the middle of the 20th century has shown that much of what we test does not add value to the learning of students (Nichols and Harris, 2016).

An important development in the middle of the 20th century was Scriven's (1967) work on developing the principles and philosophy of evaluation. A powerful aspect to evaluation that he identified was the distinction between formative evaluation taking place early enough in a process to make differences to the end points of the process and summative evaluation which determined the amount and quality or merit of what the process produced. The idea of formative evaluation was quickly adapted into education as a way of describing assessments that teachers used within classrooms to identify which children needed to be taught what material next (Bloom et al., 1971). This contrasted nicely with high-stakes end-of-unit, end-of-course, or end-of-year formal examinations that summatively judged the quality of student achievement and learning. While assessment as psychometrically validated tests and examinations historically focused on the summative experience, Scriven's formative evaluation led to using assessment processes early in the educational course of events to inform learners as to what they needed to learn and instructors as to what they needed to teach.

Nonetheless, since the late 1980s (largely thanks to Sadler, 1989) the distinction between summative and formative transmogrified from timing to one of type. Formative assessments began to be only those which were not formal tests but were rather informal interactions in classrooms. This perspective was extended by the UK Assessment Reform Group (2002) which promulgated basic principles of formative assessment around the world. Those classroom assessment practices focused much more on what could be seen as classroom teaching practices (Brown, 2013, 2019, 2020a). Instead of testing, teachers interacted with students on-the-fly, in-the-moment of the classroom through questions and feedback that aimed to help students move towards the intended learning outcomes established at the beginning of lessons or courses. Thus, assessment for learning has become a child-friendly approach (Stobart, 2006) to involving learners in their learning and developing rich meaningful outcomes without the onerous pressure of testing. Much of the power of this approach was that it came as an alternative to the national curriculum of England and Wales that incorporated high-stakes standardised assessment tasks of children at ages 7, 9, 11, and 14 (i.e., Key Stages 1 to 4; Wherrett, 2004).

In line with increasing access to schooling worldwide throughout the 20th century, there is concern that success on high-consequence, summative tests simply reinforced pre-existing social status and hierarchy (Bourdieu, 1974). This position argues tests are not neutral but rather tools of elitism (Gipps, 1994). Unfortunately, when assessments have significant consequences, much higher proportions of disadvantaged students (e.g., minority students, new speakers of the language-medium of assessment, special needs students, those with reading difficulties, etc.) do not experience such benefits (Brown, 2008). This was a factor in the development of using assessment high-quality formative assessment to accelerate the learning progression of disadvantaged students. Nonetheless, differences in group outcomes do not always mean tests are the problem; group score differences can point out that there is sociocultural bias in the provision of educational resources in the school system (Stobart, 2005). This would be rationale for system monitoring assessments, such as Hong Kong's Territory Wide System Assessment,[1] the United States' National Assessment of Educational Progress,[2] or Australia's National Assessment Program Literacy and Numeracy.[3] The challenge is how to monitor a system without blaming those who have been let down by it.

Key Stage tests were put in place, not only to evaluate student learning, but also to assure the public that teachers and schools were achieving important goals of education. This use of assessment put focus on accountability, not for the student, but for the school and teacher (Nichols and Harris, 2016). The decision to use tests of student learning to evaluate schools and teachers was mimicked, especially in the United States, in various state accountability tests, the No Child Left Behind legislation, and even such innovative programs of assessment as Race to the Top and PARCC. It should be noted that the use of standardised tests to evaluate teachers and schools is truly a global phenomenon, not restricted to the UK and the USA (Lingard and Lewis, 2016). In this context, testing became a summative evaluation of teachers and school leaders to demonstrate school effectiveness and meet accountability requirements.

The current situation is that assessment is perceived quite differently by experts in different disciplines. Psychometricians tend to define assessment in terms of statistical modelling of test scores. Psychologists use assessments for diagnostic description of client strengths or needs. Within schooling, leaders tend to perceive assessment as jurisdiction or state-mandated school accountability testing, while teachers focus on assessment as interactive, on-the-fly experiences with their students, and parents (Buckendahl, 2016; Harris and Brown, 2016) understand assessment as test scores and grades. The world of psychology has become separated from the worlds of classroom teaching, curriculum, psychometrics and statistics, and assessment technologies.

This brief history bringing us into early 21st century shows that educational assessment is informed by multiple disciplines which often fail to talk with or even to each other. Statistical analysis of testing has become separated from psychology and education, psychology is separated from curriculum, teaching is separated from testing, and testing is separated from learning. Hence, we enter the present with many important facets that inform effective use of educational assessment siloed from one another.

Now and next

Currently the world of educational statistics has become engrossed in the large-scale data available through online testing and online learning behaviours. The world of computational psychometrics seeks to move educational testing statistics into the dynamic analysis of big data with machine learning and artificial intelligence algorithms potentially creating a black box of sophisticated statistical models (e.g., neural networks) which learners, teachers, administrators, and citizens cannot understand (von Davier et al., 2019). The introduction of computing technologies means that automation of item generation (Gierl and Lai, 2016) and scoring of performances (Shin et al., 2021) is possible, along with customisation of test content according to test-taker performance (Linden and Glas, 2000). The Covid-19 pandemic has rapidly inserted online and distance testing as a commonplace practice with concerns raised about how technology is used to assure the integrity of student performance (Dawson, 2021).

The ecology of the classroom is not the same as that of a computerised test. This is especially notable when the consequence of a test (regardless of medium) has little relevance to a student (Wise and Smith, 2016). Performance on international large-scale assessments (e.g., PISA, TIMSS) may matter to government officials (Teltemann and Klieme, 2016) but these tests have little value for individual learners. Nonetheless, governmental responses to PISA or TIMSS results may create policies and initiatives that have trickle-down effect on schools and students (Zumbo and Forer, 2011). Consequently, depending on the educational and cultural environment, test-taking motivation on tests that have consequences for the state can be similar to a test with personal consequence in East Asia (Zhao et al., 2020), but much lower in a western democracy (Zhao et al., 2022). Hence, without surety that in any educational test learners are giving full effort (Thorndike, 1924), the information generated by psychometric analysis is likely to be invalid. Fortunately, under computer testing conditions, it is now possible to monitor reduced or wavering effort during an actual test event and provide support to such a student through a supervising proctor (Wise, 2019), though this feature is not widely prevalent.

Online or remote teaching, learning, and assessment have become a reality for many teachers and students, especially in light of our educational responses to the Covid-19 pandemic. Clearly, some families appreciate this because their children can progress rapidly, unencumbered by the teacher or classmates. For such families, continuing with digital schooling would be seen as a positive future. However, reliance on a computer interface as the sole means of assessment or teaching may dehumanise the very human experience of learning and teaching. As Asimov (1954) described in his short story of a future world in which children are taught individually by machines, Margie imagined what it must have been like to go to school with other children:

> Margie …was thinking about the old schools they had when her grandfather's grandfather was a little boy. All the kids from the whole neighborhood came, laughing and shouting in the schoolyard, sitting together in the schoolroom, going home together at the end of the day. They learned the same

[1] https://www.hkeaa.edu.hk/en/sa_tsa/tsa/

[2] https://www.nationsreportcard.gov/

[3] https://nap.edu.au/

> things so they could help one another on the homework and talk about it.
>
> And the teachers were people. . . .
>
> The mechanical teacher was flashing on the screen: "When we add the fractions ½ and ¼ -"
>
> Margie was thinking about how the kids must have loved it in the old days. She was thinking about the fun they had.

As Brown (2020b) has argued the option of a de-schooled society through computer-based teaching, learning, and assessment is deeply unattractive on the grounds that it is likely to be socially unjust. The human experience of schooling matters to the development of humans. We learn through instruction (Bloom, 1976), culturally located experiences (Cole et al., 1971), inter-personal interaction with peers and adults (Vygotsky, 1978; Rogoff, 1991), and biogenetic factors (Inhelder and Piaget, 1958). Schooling gives us access to environments in which these multiple processes contribute to the kinds of citizens we want. Hence, we need confidence in the power of shared schooling to do more than increase the speed by which children acquire knowledge and learning; it helps us be more human.

This dilemma echoes the tension between *in vitro* and *in vivo* biological research. Within the controlled environment of a test tube (vitro) organisms do not necessarily behave the same way as they do when released into the complexity of human biology (Autoimmunity Research Foundation, 2012). This analogy has been applied to educational assessment (Zumbo, 2015) indicating that how students perform in a computer-mediated test may not have validity for how students perform in classroom interactions or in-person environments.

The complexity of human psychology is captured in Hattie's (2004) ROPE model which posits that the various aspects of human motivation, belief, strategy, and values interact as threads spun into a rope. This means it is hard to analytically separate the various components and identify aspects that individually explain learning outcomes. Indeed, Marsh et al. (2006) showed that of the many self-concept and control beliefs used to predict performance on the PISA tests, almost all variables have relations to achievement less than $r = 0.35$. Instead, interactions among motivation, beliefs about learning, intelligence, assessment, the self, and attitudes with and toward others, subjects, and behaviours all matter to performance. Aspects that create growth-oriented pathways (Boekaerts and Niemivirta, 2000) and strategies include *inter alia* mastery goals (Deci and Ryan, 2000), deep learning (Biggs et al., 2001) beliefs, malleable intelligence (Duckworth et al., 2011) beliefs, improvement-oriented beliefs about assessment (Brown, 2011), internal, controllable attributes (Weiner, 1985), effort (Wise and DeMars, 2005), avoiding dishonesty (Murdock et al., 2016), trusting one's peers (Panadero, 2016), and realism in evaluating one's own work (Brown and Harris, 2014). All these adaptive aspects of learning stand in contrast to deactivating and maladaptive beliefs, strategies, and attitudes that serve to protect the ego and undermine learning. What this tells us that psychological research matters to understanding the results of assessment and that no one single psychological construct is sufficient to explain very much of the variance in student achievement. However, it seems we are as yet unable to identify which specific processes matter most to better performance for all students across the ability spectrum, given that almost all the constructs that have been reported in educational psychology seem to have a positive contribution to better performance. Here is the challenge for educational psychology within an assessment setting —which constructs are most important and effectual before, during, and after any assessment process (Mcmillan, 2016) and how should they be operationalised.

A current enthusiasm is to use 'big data' from computer-based assessments to examine in more detail how students carry out the process of responding to tasks. Many large-scale testing programs through computer testing collect, utilize, and report on test-taker engagement as part of their process data collection (e.g., the United States National Assessment of Educational Progress[4]). These test systems provide data about what options were clicked on, in what order, what pages were viewed, and the timings of these actions. Several challenges to using big data in educational assessment exist. First, computerised assessments need to capture the processes and products we care about. That means we need a clear theoretical model of the underlying cognitive mechanisms or processes that generate the process data itself (Zumbo et al., in press). Second, we need to be reminded that data do not explain themselves; theory and insight about process are needed to understand data (Pearl and Mackenzie, 2018). Examination of log files can give some insight into effective vs. ineffective strategies, once the data were analysed using theory to create a model of how a problem should be done (Greiff et al., 2015). Access to data logs that show effort and persistence on a difficult task can reveal that, despite failure to successfully resolve a problem, such persistence is related to overall performance (Lundgren and Eklöf, 2020). But data by itself will not tell us how and why students are successful and what instruction might need to do to encourage students to use the scientific method of manipulating one variable at a time or not giving up quickly.

Psychometric analyses of assessments can only statistically model item difficulty, item discrimination, and item chance parameters to estimate person ability (Embretson and Reise, 2000). None of the other psychological features of how learners relate to themselves and their environment are included in score estimation. In real classroom contexts, teachers make their best efforts to account for individual motivation, affect, and cognition

[4] https://www.nationsreportcard.gov/process_data/

to provide appropriate instruction, feedback, support, and questioning. However, the nature of these factors varies across time (cohorts), locations (cultures and societies), policy priorities for schooling and assessment, and family values (Brown and Harris, 2009). This means that what constitutes a useful assessment to inform instruction in a classroom context (i.e., identify to the teacher who needs to be taught what next) needs to constantly evolve and be incredibly sensitive to individual and contextual factors. This is difficult if we keep psychology, curriculum, psychometrics, and technology in separate silos. It seems highly desirable that these different disciplines interact, but it is not guaranteed that the technology for psychometric testing developments will cross-pollinate with classroom contexts where teachers have to relate to and monitor student learning across all important curricular domains.

It is common to treat what happens in the minds and emotions of students when they are assessed as a kind of 'black box' implying that the processes are opaque or unknowable. This is an approach I have taken previously in examining what students do when asked to self-assess (Yan and Brown, 2017). However, the meaning of a black box is quite different in engineering. In aeronautics, the essential constructs related to flight (e.g., engine power, aileron settings, pitch and yaw positions, etc.) are known very deeply, otherwise flight would not happen. The black box in an airplane records the values of those important variables and the only thing unknown (i.e., black) is what the values were at the point of interest. If we are to continue to use this metaphor as a way of understanding what happens when students are assessed or assess, then we need to agree on what the essential constructs are that underlie learning and achievement. Our current situation seems to be satisfied with everything is correlated and everything matters. It may be that data science will help us sort through the chaff for the wheat provided we design and implement sensors appropriate to the constructs we consider hypothetically most important. It may be that measuring timing of mouse clicks and eye tracking do connect to important underlying mechanisms, but at this stage data science in testing seems largely a case of crunch the 'easy to get' numbers and hope that the data mean something.

Conclusion

To address this concern, we need to develop for education's sake, assessments that have strong alignment with curricular ambitions and values and which have applicability to classroom contexts and processes (Bennett, 2018). This will mean technology that supports what humans must do in schooling rather than replace them with teaching/testing machines. Fortunately, some examples of assessment technology for learning do exist. One supportive technology is PeerWise (Denny et al., 2008; Hancock et al., 2018) in which students create course related multiple-choice questions and use them as a self-testing learning strategy. A school-based technology is the e-asTTle computer assessment system that produces a suite of diagnostic reports to support teachers' planning and teaching in response to what the system indicated students need to be taught (Hattie and Brown, 2008; Brown and Hattie, 2012; Brown et al., 2018). What these technologies do is support rather than supplant the work that teachers and learners need to do to know what they need to study or teach and to monitor their progress. Most importantly they are well-connected to what students must learn and what teachers are teaching. Other detailed work uses organised learning models or dynamic learning maps to mark out routes for learners and teachers using cognitive and curriculum insights with psychometric tools for measuring status and progress (Kingston et al., 2022). The work done by Wise (2019) shows that it is possible in a computer assisted testing environment to monitor student effort based on their speed of responding and give prompts that support greater effort and less speed.

Assessment needs to exploit more deeply the insights educational psychology has given us into human behavior, attitudes, inter- and intra-personal relations, emotions, and so on. This was called for some 20 years ago (National Research Council, 2001) but the underlying disciplines that inform this integration seem to have grown away from each other. Nonetheless, the examples given above suggest that the gaps can be closed. But assessments still do not seem to consider and respond to these psychological determinants of achievement. Teachers have the capability of integrating curriculum, testing, psychology, and data at a superficial level but with some considerable margin of error (Meissel et al., 2017). To overcome their own error, teachers need technologies that support them in making useful and accurate interpretations of what students need to be taught next that work with them in the classroom. As Bennett (2018) pointed out more technology will happen, but perhaps not more tests on computers. This is the assessment that will help teachers rather than replace them and give us hope for a better future.

Author contributions

GB wrote this manuscript and is solely responsible for its content.

Funding

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

Acknowledgments

A previous version of this paper was presented as a keynote address to the 2019 biennial meeting of the European Association for Research in Learning and Instruction, with the title *Products,*

Processes, Psychology, and Technology: Quo Vadis Educational Assessment?

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Asimov, I. (1954). Oh the fun they had. *Fantasy Sci. Fiction* 6, 125–127.

Assessment Reform Group (2002). *Assessment for learning: 10 principles Research-based Principles to Guide Classroom Practice* Cambridge: Assessment Reform Group.

Autoimmunity Research Foundation. (2012). Differences between in vitro, in vivo, and in silico studies [online]. The Marshall Protocol Knowledge Base. Available at: http://mpkb.org/home/patients/assessing_literature/in_vitro_studies (Accessed November 12, 2015).

Bennett, R. E. (2018). Educational assessment: what to watch in a rapidly changing world. *Educ. Meas. Issues Pract.* 37, 7–15. doi: 10.1111/emip.12231

Biggs, J., Kember, D., and Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *Br. J. Educ. Psychol.* 71, 133–149. doi: 10.1348/000709901158433

Bloom, B. S. (1976). *Human Characteristics and School Learning*. New York: McGraw-Hill.

Bloom, B., Hastings, J., and Madaus, G. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York:McGraw Hill.

Boekaerts, M., and Niemivirta, M. (2000). "Self-regulated learning: finding a balance between learning goals and ego-protective goals," in *Handbook of Self-regulation*. eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (San Diego, CA: Academic Press).

Bourdieu, P. (1974). "The school as a conservative force: scholastic and cultural inequalities," in *Contemporary Research in the Sociology of Education*. ed. J. Eggleston (London: Methuen).

Brown, G. T. L. (2008). *Conceptions of Assessment: Understanding what Assessment Means to Teachers and Students*. New York: Nova Science Publishers.

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' conceptions of assessment inventory. *Educ. Psychol.* 31, 731–748. doi: 10.1080/01443410.2011.599836

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

Brown, G. T. L. (2019). Is assessment for learning really assessment? *Front. Educ.* 4:64. doi: 10.3389/feduc.2019.00064

Brown, G. T. L. (2020a). Responding to assessment for learning: a pedagogical method, not assessment. *N. Z. Annu. Rev. Educ.* 26, 18–28. doi: 10.26686/nzaroe.v26.6854

Brown, G. T. L. (2020b). Schooling beyond COVID-19: an unevenly distributed future. *Front. Educ.* 5:82. doi: 10.3389/feduc.2020.00082

Brown, G. T. L., and Harris, L. R. (2009). Unintended consequences of using tests to improve learning: how improvement-oriented resources heighten conceptions of assessment as school accountability. *J. MultiDisciplinary Eval.* 6, 68–91.

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

Brown, G. T. L., O'leary, T. M., and Hattie, J. A. C. (2018). "Effective reporting for formative assessment: the asTTle case example," in *Score Reporting: Research and Applications*. ed. D. Zapata-Rivera (New York: Routledge).

Brown, G. T., and Hattie, J. (2012). "The benefits of regular standardized assessment in childhood education: guiding improved instruction and learning," in *Contemporary Educational Debates in Childhood Education and Development*. eds. S. Suggate and E. Reese (New York: Routledge).

Buckendahl, C. W. (2016). "Public perceptions about assessment in education," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

China Civilisation Centre (2007). *China: Five Thousand Years of History and Civilization*. Hong Kong: City University of Hong Kong Press.

Clauser, B. E. (2022). "A history of classical test theory," in *The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice*. eds. B. E. Clauser and M. B. Bunch (New York: Routledge).

Cole, M., Gay, J., Glick, J., and Sharp, D. (1971). *The Cultural Context of Learning and Thinking: An Exploration in Experimental Anthropology*. New York: Basic Books.

Croft, M., and Beard, J. J. (2022). "Development and evolution of the SAT and ACT," in *The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice*. eds. B. E. Clauser and M. B. Bunch (New York: Routledge).

Cronbach, L. J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika* 19, 263–270. doi: 10.1007/BF02289226

Dawson, P. (2021). *Defending Assessment Security in a Digital World: Preventing e-cheating and Supporting Academic Integrity in Higher Education*. London: Routledge.

Deci, E. L., and Ryan, R. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78.

Denny, P., Hamer, J., Luxton-Reilly, A., and Purchase, H. PeerWise: students sharing their multiple choice questions. ICER '08: Proceedings of the Fourth international Workshop on Computing Education Research; September6–7 (2008). Sydney, Australia, 51-58.

Duckworth, A. L., Quinn, P. D., and Tsukayama, E. (2011). What no child left behind leaves behind: the roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *J. Educ. Psychol.* 104, 439–451. doi: 10.1037/a0026280

Elman, B. A. (2013). *Civil Examinations and Meritocracy in Late IMPERIAL China*. Cambridge: Harvard University Press.

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah: LEA.

Encyclopedia Brittanica (2010a). Europe in the middle ages: the background of early Christian education. Encyclopedia Britannica.

Encyclopedia Brittanica (2010b). Western education in the 19th century. Encyclopedia Britannica.

Feng, Y. (1995). From the imperial examination to the national college entrance examination: the dynamics of political centralism in China's educational enterprise. *J. Contemp. China* 4, 28–56. doi: 10.1080/10670569508724213

Gierl, M. J., and Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educ. Meas. Issues Pract.* 35, 6–20. doi: 10.1111/emip.12129

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer Press.

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Hancock, D., Hare, N., Denny, P., and Denyer, G. (2018). Improving large class performance and engagement through student-generated question banks. *Biochem. Mol. Biol. Educ.* 46, 306–317. doi: 10.1002/bmb.21119

Harris, L. R., and Brown, G. T. L. (2016). "Assessment and parents," in *Encyclopedia of Educational Philosophy And theory*. ed. M. A. Peters (Springer: Singapore).

Hattie, J. Models of self-concept that are neither top-down or bottom-up: the rope model of self-concept. 3rd International Biennial Self Research Conference; July, (2004). Berlin, DE.

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

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York; Basic Books

Kingston, N. M., Alonzo, A. C., Long, H., and Swinburne Romine, R. (2022). Editorial: the use of organized learning models in assessment. *Front. Education* 7:446. doi: 10.3389/feduc.2022.1009446

Kline, R. B. (2020). "Psychometrics," in *SAGE Research Methods Foundations*. eds. P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug and R. A. Williams (London: Sage).

Linden, W. J. V. D., and Glas, G. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. London: Kluwer Academic Publishers.

Lingard, B., and Lewis, S. (2016). "Globalization of the Anglo-American approach to top-down, test-based educational accountability," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

Lundgren, E., and Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educ. Res. Eval.* 26, 275–301. doi: 10.1080/13803611.2021.1963940

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., and Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *Int. J. Test.* 6, 311–360. doi: 10.1207/s15327574ijt0604_1

Mcmillan, J. H. (2016). "Section discussion: student perceptions of assessment," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

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

Murdock, T. B., Stephens, J. M., and Groteweil, M. M. (2016). "Student dishonesty in the face of assessment: who, why, and what we can do about it," in *Handbook of Human and Social Conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* The National Academies Press.

Nichols, S. L., and Harris, L. R. (2016). "Accountability assessment's effects on teachers and schools," in *Handbook of human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

Panadero, E. (2016). "Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

Pearl, J., and Mackenzie, D. (2018). *The Book of why: The New Science of Cause and Effect*. New York: Hachette Book Group.

Resnick, L. B., and Resnick, D. P. (1989). *Assessing the Thinking Curriculum: New Tools for Educational Reform*. Washington, DC: National Commission on Testing and Public Policy.

Rogoff, B. (1991). "The joint socialization of development by young children and adults," in *Learning to Think: Child Development in Social Context 2*. eds. P. Light, S. Sheldon and M. Woodhead (London: Routledge).

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18, 119–144. doi: 10.1007/BF00117714

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

Shin, J., Guo, Q., and Gierl, M. J. (2021). "Automated essay scoring using deep learning algorithms," in *Handbook of Research on Modern Educational Technologies, Applications, and Management*. ed. D. B. A. M. Khosrow-Pour (Hershey, PA, USA: IGI Global).

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assess. Educ. Principles Policy Pract.* 12, 275–287. doi: 10.1080/09695940500337249

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

Teltemann, J., and Klieme, E. (2016). "The impact of international testing projects on policy and practice," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

Thorndike, E. L. (1924). Measurement of intelligence. *Psychol. Rev.* 31, 219–252. doi: 10.1037/h0073975

Von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T., and Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Front. Educ.* 4:69. doi: 10.3389/feduc.2019.00069

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA:Harvard University Press.

Weiner, B. (1985). An Attributional theory of achievement motivation and emotion. *Psychol. Rev.* 92, 548–573. doi: 10.1037/0033-295X.92.4.548

Wherrett, S. (2004). The SATS story. The Guardian, 24 August.

Wise, S. L. (2019). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Educ. Inq.* 10, 21–33. doi: 10.1080/20004508.2018.1490127

Wise, S. L., and Demars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educ. Assess.* 10, 1–17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., and Smith, L. F. (2016). "The validity of assessment when students don't give good effort," in *Handbook of Human and Social Conditions in Assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge).

Yan, Z., and Brown, G. T. L. (2017). A cyclical self-assessment process: towards a model of how students engage in self-assessment. *Assess. Eval. High. Educ.* 42, 1247–1262. doi: 10.1080/02602938.2016.1260091

Zhao, A., Brown, G. T. L., and Meissel, K. (2020). Manipulating the consequences of tests: how Shanghai teens react to different consequences. *Educ. Res. Eval.* 26, 221–251. doi: 10.1080/13803611.2021.1963938

Zhao, A., Brown, G. T. L., and Meissel, K. (2022). New Zealand students' test-taking motivation: an experimental study examining the effects of stakes. *Assess. Educ.* 29, 1–25. doi: 10.1080/0969594X.2022.2101043

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.

Zumbo, B. D., and Chan, E. K. H. (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. Cham, CH: Springer Press.

Zumbo, B. D., and Forer, B. (2011). "Testing and measurement from a multilevel view: psychometrics and validation," in *High Stakes Testing in Education-Science and Practice in K-12 Settings*. eds. J. A. Bovaird, K. F. Geisinger and C. W. Buckendahl (Washington: American Psychological Association Press).

Zumbo, B. D., Maddox, B., and Care, N. M. (in press). Process and product in computer-based assessments: clearing the ground for a holistic validity framework. *Eur. J. Psychol. Assess.*