frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 18 September 2024
DOI 10.3389/feduc.2024.1488012

OPEN ACCESS

EDITED AND REVIEWED BY
Raona Williams,
Ministry of Education, United Arab Emirates

*CORRESPONDENCE
Gavin T. L. Brown
✉ gt.brown@auckland.ac.nz

RECEIVED 29 August 2024
ACCEPTED 06 September 2024
PUBLISHED 18 September 2024

CITATION
Brown GTL (2024) Editorial: Insights in assessment, testing, and applied measurement: 2022. *Front. Educ.* 9:1488012. doi: 10.3389/feduc.2024.1488012

Editorial: Insights in assessment, testing, and applied measurement: 2022

Gavin T. L. Brown*

Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand

KEYWORDS

testing, psychometrics, educational assessment, evaluation, methods

Editorial on the Research Topic
Insights in assessment, testing, and applied measurement: 2022

Introduction

This Research Topic focused on new insights, novel developments, current challenges, recent advances, and future perspectives in the field of assessment in education. The goal was to shed light on the progress made in the past decade in the assessment, testing and applied measurement field and on its future challenges to provide a thorough overview of the state of the art of the assessment, testing and applied measurement field. Measurement, assessment, testing, and various classroom or testing protocols matter to quality and justice. These processes are used to inform decision making but as we enter the third decade of the twenty-first century there is increasing complexity in the world in which assessment functions. That means what we already know may not be a good basis for future action, policy, or practice. Consequently, we solicited brief, forward-looking contributions from Assessment, Testing and Applied Measurement editorial board members that either described the state of the art or highlighted changes needed to move the field forward. We expected these authors, based on their contribution to the journal through editing and reviewing manuscripts, let alone their own research agendas, to identify the greatest challenges in their sub-disciplines, and how to address those challenges.

This article Research Topic will inspire, inform, and provide direction and guidance to researchers in the field. From 2022 to 2024, a total of 20 manuscripts were added to the Research Topic. Seven of the papers involved students, nine focused on teachers, and four spoke to concerns of researchers and policymakers. Students and teachers at various levels of the K-12 compulsory school systems were the focus of 12 papers and six papers focused on students at various levels of tertiary or higher education. As an aside, my thanks go out to the many reviewers and editors who helped the authors create good papers. The quality of this RT depended on those folk.

Insights

Two of the papers addressed to researchers were highly technical expositions related to Lawshe's content validity index (Jeldres et al.) and error variance inflation in measurement models (Metsämuuronen), both of which should benefit psychometric researchers needing authoritative sources on those methods.

Similarly, two papers provided review or overview perspectives. Brown, Kannan et al. provide a discursive set of opinions and perspectives about how test developers can better communicate test results to teachers, administrators, and other educational stakeholders. While potentially somewhat repetitive, the voices of five different experts, with varied approaches and contexts, give strong suggestions for future research. Pastore provides a systematic review of the literature on teacher assessment literacy for the most recent 10-year period (2013–2022). Pastore shows that the field has wide variation in how this core classroom teacher competence is defined, understood, and studied; a clear example of the perpetual problem of "jingle-jangle" in educational and psychological research. Nonetheless, Pastore reports that there are foundational components which are contingent upon contextual factors, reinforcing results from previous reviews.

Two other studies highlighted aspects of teacher assessment competence. Kissi et al. demonstrated the weaknesses Ghanaian teachers had in terms of creating multiple-choice test questions, using a test of item quality and an analysis of actual test forms created by the teachers. Their multimethod study showed, despite well- and long-established guidelines for writing good test questions, teachers could not recognize or create consistently high-quality objectively scored test items. Leukel et al. examined how teachers of gymnastics form quality judgements of student performance compared to more expert trainers. Their study found judgment accuracy, agreement on ratings, and agreement about the temporal structuring of tasks was significantly lower for teachers compared to trainers. Expertise in any domain being evaluated leads to better judgements and grading. Given that teacher assessment literacy is a very broad multifaceted competency, it is highly likely that teacher-made assessments, judgements, or feedback will remain problematic for a long time to come.

Five papers reported scale development or validation studies, drawing on data from Mexico, Sweden, China, USA, and Iran. The studies used complex statistical methods, including multi-group confirmatory factor analysis invariance testing (Henríquez et al.), content analysis of test items (Rosenlund), exploratory factor analysis, hierarchical regression, and multilevel modeling (Lu et al.), WLSMV estimation of longitudinal item factor analysis with invariance testing (Ding et al.), and confirmatory factor analysis and structural equation modeling (Brown, Andersson et al.).

A wide variety of Research Topics have been captured by these studies. Rosenlund examined the epistemic cognition requirements of large-scale tests of history in Sweden, finding an over-emphasis on objective dimensions of historical knowledge, challenging the design of future tests to better evaluate all epistemic requirements of the subject. Henríquez et al. evaluated an inventory with Mexican students for student evaluation of higher education teaching in the social sciences. They reported invariance and good model fit for a three-factor model (i.e., course organization, teaching quality, and evaluation and feedback) of teachers' performance. Lu et al. tested the cross-cultural reliability and validity of a scale concerning educator cognitive sensitivity with a sample of Chinese early childhood educators, concluding that the validation evidence was weak and necessitated further work. Brown, Andersson et al. tested a previously published measure of teacher conceptions of feedback in Sweden and modified it by proposing some of the items constituted a factor of teacher feedback practices. This model had good fit and showed that endorsement of feedback for improvement and that students may ignore feedback both contributed to the feedback practices teachers claimed to make. These studies identify and support further use of measures for practice, research, and possibly even with teacher professional development.

Bridgeman et al.'s analysis of the relationship between test scores for entry into graduate higher education (i.e., Graduate Record Examination, GRE) and doctoral degree completion used multilevel analysis to show that greater persistence was associated with higher verbal and analytical writing scores and inversely with quantitative scores. Despite the odds ratio values being close to 1.00 (grand average = 1.03), the authors recommend keeping GRE scores in the decision matrix, a recommendation that should be taken cautiously.

Innovations

The world of test validity and academic integrity is being threatened by uncontrolled use of AI or LLM technologies, so we need to have insights that might lead to effective use of such technologies in actual teaching and learning practices, let alone its potential to validly create or reliably score high-stakes, large scale assessments. Justice in society is ensured when mean score differences at individual or group levels are supported by well-designed assessments and systems that consider differential opportunity to learn. These are important messages for policymakers and politicians who have responsibility for the design of education systems.

Innovative views of the future of assessment included a paper on how AI can be used in formative assessment (Hopfenbeck et al.) and how AI can be used to design assessments that do not create oppressive outcomes for minority students (Sparks et al.). Hopfenbeck et al. identify ways that AI might be used in classroom contexts (e.g., automated essay scoring, generating feedback, and generating learner profiles and subsequent automated tutoring). However, they point out teachers lack skills to exploit these innovative possibilities and there is considerable work needed to turn the potential of AI into actual formative practices. Sparks et al. have conceived a framework for how AI can be used to develop personalized classroom assessment. Their framework adapts assessment processes to provide "care" for learners before, during, and after testing. The aim is to ensure that contextual information about learners, including their personal characteristics and ways of behaving, are incorporated into the design and administration of assessment. These papers offer visions of how AI can be used to improve the quality of classroom assessments, but this still remains a major challenge for systems.

In terms of new testing or assessment protocols, Kafipour and Khoshnood demonstrated that dynamic assessment in which the instructor assesses student language performance by asking questions and providing hints or prompts had a positive impact on Field Dependent EFL learners in Iran. Field dependence is a cognitive style in which the learner focuses on the overall meaning and the whole field, exhibits more relational behaviors, and needs

more external reinforcements to stay motivated. This alignment makes sense and raises interesting challenges for those of us who rely on traditional assessment processes. Perhaps, AI machines can be programmed to interact with language learners in assessments and reduce workload on teachers?

Remesal and Estrada present a small-scale study of Spanish teacher educators who used an innovative synchronous self-assessment strategy during written exam situations (i.e., during the exam, students select the tasks or questions they will answer and they choose a weighted grading scheme for their successful answers). The four instructors were individually interviewed after the examination and they claimed marking was less tedious because students did not all do the same tasks and that the different weighting choices provided clues to teachers about student competence, potentially informing more effective instruction.

Instead of the traditional focus on determining cognitive difficulty for mainstream subjects, Ehninger et al. focused on predicting the cognitive difficulty of items in a test of music-related argumentation (i.e., MARKO). They found among German high school students that the strongest predictor of harder test questions was "reference to musical attributes," "cross-sentence argumentation," and "dialogical argumentation" features. The study provides validation evidence for the MARKO test and may provide a model for testing of other creative and/or performing arts.

Greater use of peer feedback is advocated, especially in higher education, on the assumption that this helps both feedback recipient and provider. However, lack of psychological safety in the process will lead to faulty communication. Senden et al. created a brief student training program to increase psychological safety and trust in peer feedback, but their experiment with Belgian higher education students in acrobatic sports didactics failed to find any statistically significant effect for safety or grade improvement. Nonetheless, researchers and instructors might want to inspect the treatment design to identify ways in which their own work might improve results.

Xue et al. used artificial neural network (ANN) analysis to predict academic performance in English listening and speaking as a Foreign Language among Chinese university students. Despite a very small sample size ($n = 62$), the data driven ANN found that overall performance seemed to depend on academic performance on the Chinese college entrance examination English test (gao kao), average scores of all peers' assessment covering English abilities, class participation, cooperation and competitiveness, and learning attitude and perseverance, standardized teacher ratings and student self-assessment. Clearly, the results are greatly limited by sample size, but it is encouraging that the ANN system worked with such small numbers, perhaps because there were so many variables per student.

Unsurprisingly, Chauliac et al. studied five different approaches to determining if survey responses are "careless." They reported that notable proportions of Flemish adolescent students (age 15–17) exhibited careless responding when completing self-report surveys (rate ranged from 12 to 31% depending on method). Carelessness clearly mattered to the quality of data. Interestingly, the method of determining carelessness matters, because few participants would be eliminated by two or more methods. Hence, the study provides new options for researchers as to how they might determine whether participants were attentive or not.

Conclusion

I commend this set of papers as a useful contribution to the field. They provide both insights to current methods or findings and innovations concerning the future of assessment, testing, and applied measurement.

Author contributions

GB: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note