PROCEEDINGS OF THE 46TH ANNUAL MEETING OF THE NORTH AMERICAN CHAPTER OF THE INTERNATIONAL GROUP FOR THE PSYCHOLOGY OF MATHEMATICS EDUCATION

Envisioning the Future of Mathematics Education in Uncertain Times

EDITED BY

Karl W. Kosko

Joanne Caniglia

Scott A. Courtney

Maryam Zolfaghari

Grace A. Morris

Kent State University

SYNTHESIZING BIAS AND FAIRNESS EVIDENCE FOR THE PSM-CAT

Jonathan D. Bostic
BGSU
bosticj@bgsu.edu

Gabriel Matney
BGSU
gmatney@bgsu.edu

Toni May
Drexel Univ.
Tas365@drexel.edu

Kristin Koskey
Drexel Univ.
Kk3436@drexel.edu

Gregory Stone
Clarity Assessment
gregory@metriks.com

Timothy Folger
BGSU
tdfolge@bgsu.edu

Keywords: Assessment

Assessment is a part of teaching and learning, and recent efforts seek to improve assessment practices within K-12 education (Harris et al., 2023). There has also been a substantive shift from paper-and-pencil assessments to technology-delivered assessments (Thompson, 2017). Fairness and bias considerations are an important aspect of the assessment validation process (AERA et al., 2014; Herman & Cook, 2022). Exploring fairness and bias issues related to technology-delivered mathematics tests is necessary. This study's purpose is to synthesize bias and fairness validity evidence related to a technology-delivered, computer-adaptive mathematics problem-solving test called the PSM-CAT. The PSM-CAT is composed of standards-aligned, mathematics word problems. It is designed for students in grades 6-8 (age 11-14) and is intended to be formative in nature (see Bostic et al., 2024 for more information).

Potential respondents as well as possible test administrators and users can provide useful data regarding issues of test bias and fairness (Lane, 2014). Our team conducted 1-1 and small-group interviews with students and adults from a broad sample over two years. We purposefully and representatively interviewed (i) students and (ii) teachers, administrators, or curriculum specialists, and (iii) STEM Education faculty and professionals representing different (a) geographic regions, (b) school communities, (c) individuals with and without learning disabilities, (d) multilinguals and native English speakers, (e) genders, and (f) students representing BIPOC and White students. Interviews were audio recorded and transcribed. Data were qualitatively analyzed using a two-cycle approach to generate themes, with checks and balances throughout the analysis to promote trustworthiness (Miles et al., 2014). We report our findings from student-data (potential respondents) and adult-data (potential test users) separately.

One theme emerged from students' data: Items' contexts, language, and content were *broadly accessible to peers*; nearly all students did not perceive bias within items. We will present quotations in our poster and summarize the data as showing that 125 of 128 students (98%) reported items as fair, content-appropriate, and free from bias. A second theme came from adults: Readability, contexts, and standards-alignment was *fair and appropriate for diverse learners*. An implication from this two-year study is opening our test development scholarship on bias and fairness for scrutiny, which may be a call for other test developers to publicly share their findings. Bias and fairness are a cornerstone for a validity argument and should be explored to promote better, shorter tests and concomitantly, more time for teaching and learning (Bostic, 2023).

Acknowledgments

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF 2100988l; 2101026). Anything expressed by the authors do not necessarily reflect the views of the National Science Foundation.

References

Bostic, J. (2023). Engaging hearts and minds in assessment research. *School Science and Mathematics Journal*, *123*(6), 217–219. https://doi.org/10.1111/ssm.12621

Bostic, J., May, T., Matney, G., Koskey, K., Stone, G., & Folger, T. (2024, March). *Computer adaptive mathematical problem-solving measure: A brief validation report. In D. Kombe & A. Wheeler (Eds.), Proceedings of the 51st annual meeting of the* Research Council on Mathematics Learning (pp. 102–110). Columbia, SC.

Black, P., & Wiliam, D. (1998*). Inside the black box: Raising standards through classroom assessment.* King's College London.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational measurement: Issues and practice, 22*(4), 5–12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Harris, C., Wiebe, E., Grover, S., & Pellegrino, J. (Eds.) (2023). *Classroom-Based STEM assessment: Contemporary issues and perspectives. Community for Advancing Discovery Research in Education* (CADRE). Education Development Center, Inc.

Herman, J., & Cook, L. (2022). Broadening the reach of the fairness standards. In J. Jonson & K. Geisinger (Eds.), *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 standards* (pp. 33–59). American Educational Research Association.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. 10.7334/psicothema2013.258

Miles, M., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data Analysis: A methods sourcebook* (3rd ed.). Sage.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268– 280. https://doi.org/10.1080/0969594X.2016.1141168

Thompson, G. (2017). Computer adaptive testing, big data, and algorithmic approaches to education. *British Journal of Sociology of Education, 38*(6), 827–840. https://doi.org/10.1080/01425692.2016.1158640