ADAPTATION OF THE DELPHI TECHNIQUE IN THE DEVELOPMENT OF ASSESSMENTS OF PROBLEM-SOLVING IN COMPUTER ADAPTIVE TESTING ENVIRONMENTS (DEAP-CAT)

K. Koskey[1], D. Bright[1], K. Struloeff[1], T. Sondergeld[1], G. Stone[2], J. Bostic[3], G. Matney[3]

[1]Drexel University (UNITED STATES)
[2]University of Toledo (UNITED STATES)
[3]Bowling Green State University (UNITED STATES)

Abstract

The Standards for educational and psychological assessment[1] specify assessment developers establish five types of validity evidence. Relevant to this paper is consequential validity evidence that identifies the potential negative impact of testing or bias. Standard 3.1 of *The Standards*[1] on fairness in testing states that "those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant sub-groups in the intended populations" (p. 63). The Delphi technique is a systematic research method used in instrument development to identify sources potentially leading to item bias against one or more subgroups[2]. A diverse panel of two or more experts review item content through an iterative process until saturation and consensus are reached among experts as established through some criterion. Research has illustrated this technique applied *after* detecting differential item functioning*,* but not *before* administering items in the field. This paper is a methodological illustration of an adaptation of the Delphi technique applied in the item construction phase of assessment development. The Delphi technique was used as part of the earlier phases of a larger five-year study initiated in August 2021 and funded by the National Science Foundation (Award no. 2101026, 2100988) to develop and test new problem-solving measures (PSM)[3,4] for U.S.A. grades 6-8 in a computer adaptive testing environment. As part of an iterative design-based research methodology[5], how the Delphi technique was integrated into the initial phase of item writing process is outlined. A description of the process and partial results from a three-person item bias panel reviewing a set of 45 PSM items are outlined to illustrate the technique. IBP members rated the use of the Delphi technique for identifying sources potentially leading to bias and process to facilitate their review tasks as "*effective*" to "*very effective."*

Keywords: Assessment, problem-solving, item bias, Delphi technique, item bias panel, consequential validity

1 INTRODUCTION

The Standards for educational and psychological assessment were developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education[1]. Five types of validity evidence are specified in *The Standards*: test content, response processes, internal structure, relationship to other variables, and consequential/bias. Relevant to this paper is consequential validity evidence that identifies the potential negative impact of testing or bias. Standard 3.1 of *The Standards*[1] on fairness in testing states that "those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant sub-groups in the intended populations" (p. 63). Three types of bias include construct, method, and item bias[6]. Testing for differential item functioning (DIF) is a standard analysis adopted to detect item bias against a subgroup[6]. Example subgroups include gender, race/ethnic group, socioeconomic status, native language, and disability. DIF is when "equally able test takers differ in their probabilities answering a test item correctly as a function of group membership"[1] (p. 51). DIF indicates systematic error as

compared to real mean group differences[7]. Items exhibiting significant DIF are removed or reviewed for sources leading to bias to determine modifications to retain and further field test an item.

An emergent systematic research method for examining items identified as having significant DIF is the Delphi technique. Expert panel members review item content through an iterative process[6], and experts independently evaluate items for potential sources leading to DIF. Researchers compile experts' responses to rate their level of agreement with the anonymously grouped responses. This process continues until saturation and consensus are reached among experts as established through some criterion or combination of set of criteria (e.g., median agreement rating, item quartile range, and/or percent agreement). The technique allows researchers to "identify, learn, and share the ideas of experts by searching for agreement among experts"[6] (p. 451). Research has illustrated this technique applied *after* DIF is detected, but not *before* administering items in the field.

In this study, the Delphi technique was applied in the initial item construction phase of assessment development as part of a five-year study initiated in August 2021 and funded by the National Science Foundation (Award no. 2101026, 2100988) entitled "The Development of Assessments of Problem-Solving in Computer Adaptive Testing Environments" (DEAP-CAT). This research develops and tests new problem-solving measures (PSM) for U.S.A. grades 6-8 in a CAT form. Problem solving has been a priority within K-12 mathematics education for over four decades[8,9,10,11]. Broadly defined, problem solving is fostered through "mathematical tasks that have potential to provide intellectual challenges for enhancing students' mathematical understanding and development"[11] (para. 1). The importance of problem solving is reflected in the Common Core State Standards for Mathematics Initiative (CCSSI)[12], which was adopted in some form by 41 states in the U.S.A. In prior research[3,4], PSMs were developed within the context of CCSSI for grades 3-5 and 6-8. The current study expands the scope of the use and score interpretation of PSM, in part, by constructing PSM 6-8 in a CAT environment. CAT offers multiple advantages to non-adaptive tests: a) *efficient assessment* – fewer items are needed compared to non-adaptive tests to measure proficiency level using an item selection algorithm; b) *increased testing fairness* – items are selected to match different ability levels; c) *real-time reporting* – students and teachers can receive score interpretations in minutes rather than weeks; and d) *more precise measures of abilities* – students complete items targeted at their individual ability levels, rather than a broad set of items spanning ability levels[13,14,15]. Illustrated next is the integration of the Delphi technique into the DEAP-CAT item writing process part of an iterative design-based research (DBR) methodology[5]. Full results are being prepared for consideration for publication in another manuscript and thus are not reported here.

2 METHODOLOGY

The overall item writing cycle spans over the first two years of the larger 5-year study. Goals were set for developing batches of items forwarded to multiple panels (external expert content, psychometric experts, and item bias experts) for review. Items are revised by the item writing team based on feedback from all internal and external panels prior to progressing to quantitative field testing. As illustrated in Figure 1, a 10-week timeline was established for item writing, review, and refinement to be employed in fall and spring to develop 720 new PSM items for the CAT environment (180 items per batch consisting of 60 items for each grade level). Yildirum and Büyüköztürk's[2] application of the Delphi technique was adopted for its advantages of providing capacity to "identify, learn, and share the ideas of experts by searching for agreement among experts"[6] (p. 451) but modified to fit within the constraints of the set timeline. Five experts external to the DEAP-CAT project were recruited to serve as the item bias panel (IBP) members. The larger panel was split into two panels (Panel A and Panel B) made up of three members varying in gender and ethnicity. One member served on both panels. Each panel consisted of individuals with varying experience and expertise in mathematics education, experience teaching in K-12, and inclusive practices. Both panels were to review two batches of items in fall and spring for the first two years of the grant (4 batches annually). Each batch consisted of about 45 items.

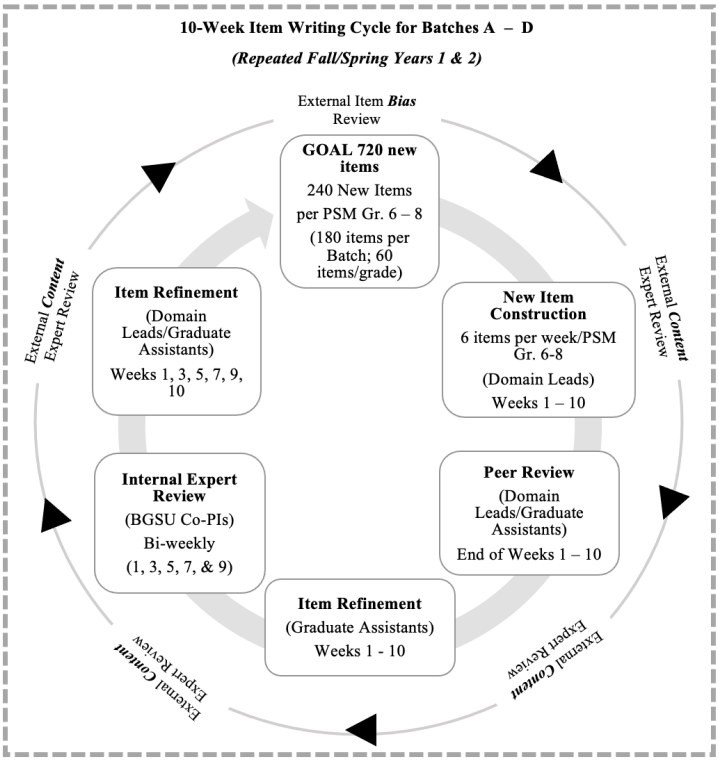

Figure 1. 10-week DEAP-CAT item writing, review, and refinement cycle.

Prior to reviewing items for potential sources of bias, IBP members attended a one-hour virtual orientation lead by the first three authors. The orientation involved:

- Introducing the DEAP-CAT project;
- Outlining the iterative item development cycle and where item bias panel review is situated in the iterative process;
- Defining “item bias” and subgroups that item bias traditionally impacts;
- Describing the intended target student population of the PSM;
- Engaging in an exercise to familiarize the reviewers with the task (independently reviewing four example items followed by a whole-group de-brief); and
- Walking through the logistics of accessing items, reporting their evaluations, and timeline for review.

After completing the orientation, panel reviewers were sent an electronic folder of problem-solving items shared using Google Drive to begin the independent review process. Outlined next is the process and a description of the results based on Panel A’s review of the first batch of items.

3 INITIAL INDEPENDENT REVIEW

3.1.1 Process

IBP members were provided 8 days (Friday to the following Friday) to complete their initial review of 45 items. This timeline provided the IBP a weekend and week to complete their first round of review as some members preferred reviewing items over the weekend. The folder contained a list of the 45 items for review by item identification number, problem-solving domain, and a short item descriptor (e.g., “Building a Fence”) to help easily identify an item. Each item for review with an example solution was provided in a single word document. Reviewers completed an online survey using Qualtrics to report on his or her evaluation of each item. For each item, IBP members were asked to: a) report whether any sources potentially leading to item

bias were identified, b) describe the source of bias, and c) classify the subgroups(s) potentially impacted by the bias (see Figure 2). Which subgroups were potentially impacted was purposefully set as allowing the IBP to "select all that apply" given that some sources of bias intersect across subgroups. As Lyons and colleagues[16] explained in their recently published, *A Call to Action: Confronting Inequity in Assessment*, evaluating for biases cross-sectionally provides for acknowledging and identifying "within-group intersectional effects such as socioeconomic status among Black females" (p. 19).

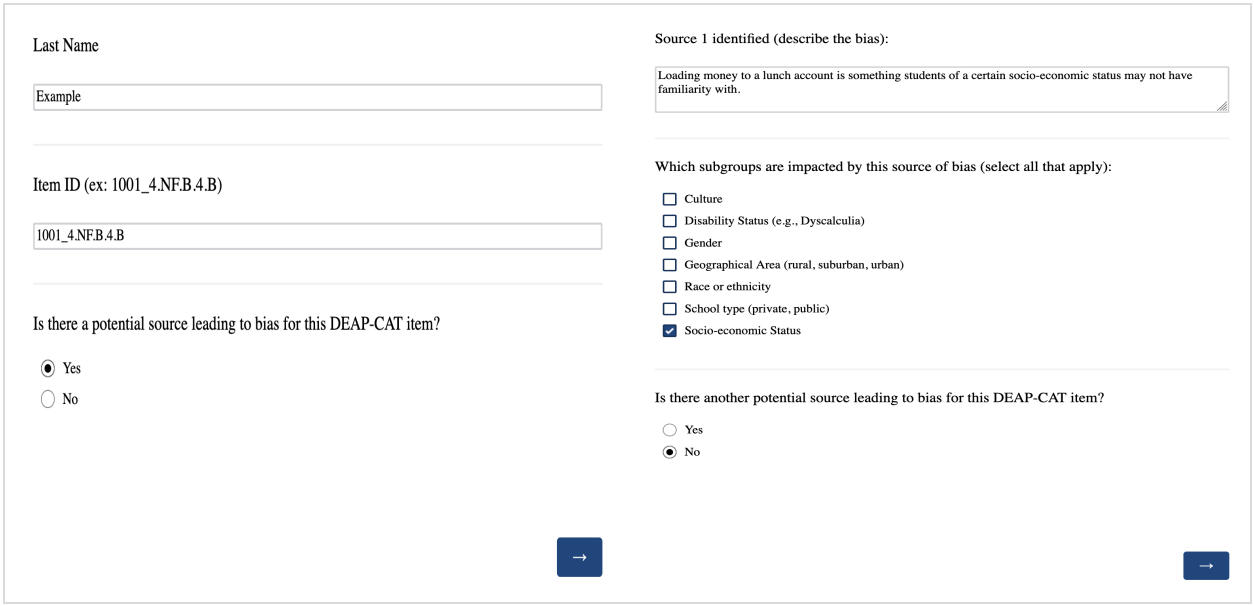

Figure 2. Online survey form for reporting independent initial review of each item.

At the end of this process, each IPB member is asked to provide a holistic rating for the set of items. For the holistic rating, participants were asked, "Reflecting on all 45 of the items you have reviewed, to what degree do you believe that these items are collectively accessible to students in the target population with diverse experiences and identities?" We described the target population as 6-8 grade students attending American school system (either private or public). Panel reviewers rated the degree of accessibility of the set of items as a whole using a 4-point Likert-type scale: 1= *Not accessible to the target student population*, 2= *Accessible to some of the target student population,* 3 = *Accessible to most of the target student population*, and 4 = *Accessible to all of the target student population*. The holistic rating scale aimed to indicate the cultural responsiveness of the items as a whole and whether the items had a sufficient range in context (e.g., not over reliant on sports-related examples). After the item review step was complete, the researchers compiled the reviewers' responses for each item and identified the items potentially leading to item bias. As reported by the three IBP members, initial independent review took between 72 to 405 minutes to complete, indicating a wide range in time to complete task.

3.1.2 Description of Independent Review Results

The item-level and aggregated results are being prepared for consideration for publication in another manuscript and thus are not reported here. But, to provide an example, one item showed students an image of a hand-drawn quilt that included complementary and parallel lines. The lines created triangles that were demarcated by various colors. Students were asked to give the angle measure for the triangular angle marked X in the quilt. Two IBP members indicated that this item contains sources that may potentially lead to bias. One IBP member identified cultural bias in the question and stated that "*Quilt-making is not a universal concept that all students would know,"* which may skew their approach to the question. Another member denoted that the question contained bias that impacted the Disability subgroup. Specifically, this question may impact the ability of students with colorblindness to differentiate the colors and find the X. It was suggested that item developers "*label each color section of the quilt so these students can easily access this question*" and understand which angle the X is supposed to be measuring. This example shows

the importance of having multiple IBP members with diverse expertise and backgrounds because they provide different insights and are able to grasp how this same question can impact different groups of students. Individuals have a set of identifies that shape their perspective and provide the research team with a well-rounded understanding of how these items operate in practice.

4 REACHING CONSENSUS

4.1.1 Consensus Process

After all IBP members submitted their initial independent reviews, two team members who did not participate in the IBP reviewed the identified potential biases and compiled the results. Descriptions of sources of bias were merged when repetitive. For example, if two IBP members identified different potential biases in an item, both would be presented. If two members identified similar bias, but with different wording in their descriptions of the bias, the descriptions were merged into a single source to present to IBP members. One research team member completed this task with the second research team member cross-checking. Any disagreements were discussed to reach consensus across the two researchers on the final list of potential sources to be provided to IBP members for a second review.

A second online survey was subsequently created to be shared with the IPB members to reach a consensus on identified biases. This survey was grouped into three sections. In the first section, items identified by IBP members as having one or more sources potentially leading to bias were presented (see Figure 3).

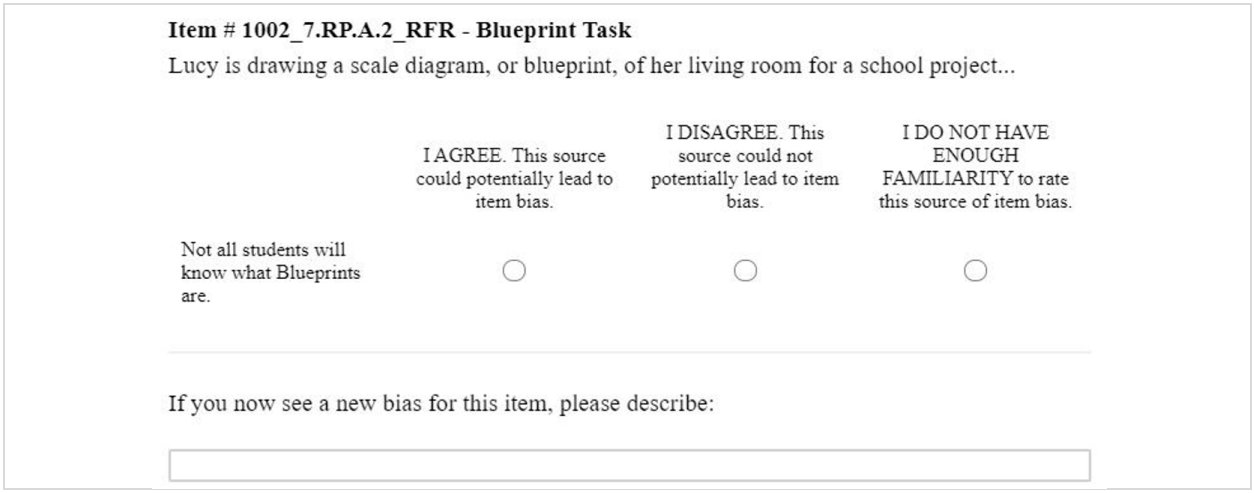

Figure 3. Survey asking IBP members to indicate agreement/agreement with each source identified during the consensus reaching step.

Consistent with the Delphi technique, the sources identified were not linked with any IBP member's name to maintain anonymity during the process. IBP members were asked to determine their level of agreement or disagreement with the identified potential bias selecting among 1= *I agree: This source could potentially lead to item bias,* 2 = *I disagree: This source could not potentially lead to item bias*, or 3 = *I do not have enough familiarity to rate this source of item bias*. IBP members were also provided a space to identify new potential biases that they may see during an additional review of the items.

The second section of the survey provided all remaining items identified as having no sources potentially leading to bias observed. IPB members had space to respond if a new source of bias was identified (Figure 4). No researchers identified new potential bias for these items in the process.

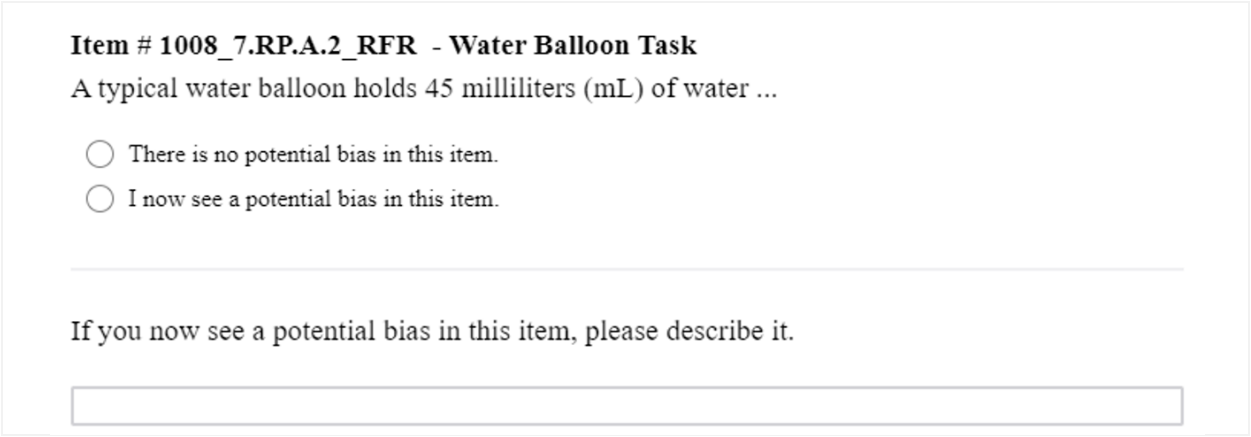

Figure 4. Survey task asking IBP to confirm/disconfirm a potential source of bias was not observed for an item.

In the second section of the survey and similar to when conducting their initial review, IBP members were also asked to provide a holistic rating of the accessibility of the items as a whole.

Section three of the survey was designed to gain insight into the IBP reviewers' experience for methodological reflection and process improvements. This section consisted of five questions. The first two questions asked IBP members to rate and explain their rating related to the extent the 1) method was ineffective/effective for identifying sources potentially leading to bias and 2) logistics of the process were for facilitating their completion of the task. The next two questions asked each member to report the minutes it took to complete the 3) initial independent review and 4) reaching consensus review. A fifth open-ended question asked whether there was anything else IBP members wanted to share about the method, process, or time to complete the task.

4.1.2 Description of Consensus Results

Whether the three IBP members reached consensus on the sources identified for an item was analyzed. The results are being prepared for publication in another manuscript; thus, only the process is described here. No established standardized criteria exist for this step in the Delphi technique. However, when consensus is not achieved based on a set criterion by the researchers, the IBP is asked to conduct another level of review until saturation and consensus are reached. In the current research, for practical purposes, any sources identified by any panel member needed to be addressed by the item writing team before moving an item forward for quantitative field testing. Also, given that grants fund research for specified timeframes, the item bias review needed to stay within a certain timeframe to progress items to the next field-testing phase. With this goal in mind, three categories were established to report the IPB consensus results to the item writing team for each source by item: 1) Retain: No sources of bias identified, 2) Review: 1 member identified this source of bias, and 3) Revise, Replace, or Remove: > 50% members identified this source of bias (at least 2 members in this study). For research purposes, degree of consensus was computed across each source by item.

An example of an item flagged for review through this process was an item referring to "kayaks." All three IBP members agreed that not all students would understand what a kayak was, especially those that may be in different geographic regions or have different socio-economic backgrounds.

4 METHODOLOGICAL REFLECTIONS

As described, after the consensus reaching process, the IBP members were asked to reflect on the methodology of the item bias review process. They were asked to rate to what extent this method was ineffective/effective for identifying sources potentially leading to item bias. A 5-point Likert-type scale was provided: 1 = *not effective*, 2 = *somewhat effective*, 3 = *effective*, 4 = *very effective*, 5 = *extremely effective*. Two IBP members rated the method as *effective*, while one member rated the method *extremely effective*.

When asked to explain the rating assigned, a member elaborated "This process allowed for the further examination of how educators frame and utilize the terminology in designing math questions which allowed for a critical examination of how bias can impact the questioning design process." This response affirms that this method of evaluating item bias is effective and should be continued for future iterations of item review. Moreover, we know that the effectiveness of this review process enabled the panelists to complete this task with minimal technical difficulties and provide thoughtful considerations for each item.

Subsequently, IBP members were asked to rate to what extent were the logistics of the process (Google folder with items, a survey tool for responses) ineffective/effective for facilitating your completion of the task using the same 5-point Likert-type scale. Two IBP members rated the process was *effective,* while one member indicated that it was *extremely effective*. All three members agreed that the process was well-organized and the platform (Google Drive) was accessible and familiar. These reflections will inform our future practices and aid us in refining the item bias processes for the next cohort of panelists over the course of this project.

In this iterative process, the IBP members identified items that contained unnecessarily challenging words/language, sports information, and niche cultural references. These items containing sources of bias often favored male students from traditionally American backgrounds. This employed Delphi technique was utilized before administering the items into the field and thus provided us with an opportunity to rectify these issues before students attempted the work. The advantage of this method is its ability to gain a greater perspective and understand the nuance of why these questions cause misconceptions for students. This paper offers future researchers' insight into how the Delphi technique can be adapted and employed in the early phases of item development prior to quantitative field testing.

ACKNOWLEDGEMENTS

This research was supported by a grant funded by the National Science Foundation – DRL – Discovery Research K-12 [Award no: 2101026 and 2100988] awarded to the first, fourth through seventh authors.

REFERENCES

[1] American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), "Standards for educational and psychological testing," Washington, DC: American Educational Research Association, 2014.

[2] H. Yildirim and S. Büyüköztürk, "Using the Delphi Technique and focus-group interviews to determine item bias on the mathematics section of the level determination exam for 201," *Educational Sciences: Theory & Practice*, vol. 18, pp. 447-470, 2018.

[3] J.D. Bostic, T.A. Sondergeld, T. Folger, and L. Kruse, "PSM7 and PSM8: Validating two problem-solving measures," *Journal of Applied Measurement,* vol. 18, pp. 1-12, 2017.

[4] J.D. Bostic and T.A. Sondergeld, "Measuring sixth-grade students' problem-solving: Validating an instrument addressing the mathematics common core," *School Science and Mathematics Journal,* vol. 115, pp. 281-291, 2015.

[5] J. Middleton, S. Gorard, C. Taylor, and B. Bannan-Ritlandm, The "compleat" design experiment. In A. Kelly, R. Lesh, & J. Baek (Eds.), "Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics teaching and learning," pp. 21-46, New York: Routledge, 2008.

[6] D. Boer, K. Hanke, and J. He, "On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests," *Journal of Cross-Cultural Psychology,* vol. 49, pp. 713-734, 2018.

[7] G. Camilli and L.A. Shepard, "Methods for identifying biased test items," Thousand Oaks, CA: Sage, 1994.

[8] National Council of Teachers of Mathematics (NCTM). "An agenda for action," Reston, VA: NCTM, 1980.

[9] National Council of Teachers of Mathematics (NCTM). "Mathematics assessment: A practical handbook for grades 6-8," Reston, VA: NCTM, 2005.

[10] National Council of Teachers of Mathematics (NCTM). “Principles to actions: ensuring mathematical success for all,” Reston, VA: NCTM, 2014.
[11] National Council of Teachers of Mathematics (NCTM), “Problem solving,” NCTM, 2020, Retrieved https://www.nctm.org/Research-and-Advocacy/Research-Brief-and-Clips/Problem-Solving NCTM
[12] Common Core State Standards Initiative, “Common core standards for mathematics,” CCSSI, 2010, Retrieved from http://www.corestandards.org/Math/
[13] R.C. Gershon and B. Bergstrom, “Understanding Rasch measurement: Computer adaptive testing,” *Journal of Applied Measurement,* vol. 6, pp. 109-127, 2005.
[14] R.L. Luecht and S. Sireci, “A review of models for computer-based testing (Report NO. 2011-2012)”, New York, NY: The College Board, 2011.
[15] J-J. Vie, J-J., F. Popineau, E. Bruillard, and Y. Bourda. A review of recent advances in adaptive assessment. In A. Pena-Ayala’s (Ed.), “Learning Analytics, Fundaments, Applications, and Trends: A view of the current state of the art to enhance e-learning” (Studies in Systems, Decision and Control, vol. 94, pp. 113-142, Cham, Switzerland: Springer International Publishing, 2017.
[16] S. Lyons, M. Johnson, and B.F Hinds, “A call to action: Confronting inequity in assessment,” 2021, Retrieved from https://www.lyonsassessmentconsulting.com/assets/files/Lyons-JohnsonHinds_CalltoAction.pdf