Methods in Psychology 8 (2023) 100117

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.sciencedirect.com/journal/methods-in-psychology

Flip it: An exploratory (versus explanatory) sequential mixed methods design using Delphi and differential item functioning to evaluate item bias

Kristin L.K. Koskey [a,*], Toni A. May [a], Yiyun "Kate" Fan [a], Dara Bright [a], Gregory Stone [b], Gabriel Matney [c], Jonathan D. Bostic [c]

[a] Drexel University, The Methods Lab, Department of Teaching, Learning, & Curriculum, School of Education, 3401 Market Street, Philadelphia, PA, 19104-3871, United States
[b] The University of Toledo, Research and Measurement, Department of Educational Studies, Judith Herb College of Education, Gillham Hall, Mail Stop 914, United States
[c] Bowling Green State University, College of Education & Human Development, 444 Education Building, Bowling Green, OH, 43403, United States

ARTICLE INFO

Keywords:
Exploratory sequential
Delphi
DIF
Concordance
Item bias
Joint display
Instrument development

ABSTRACT

The Delphi method has been adapted to inform item refinements in educational and psychological assessment development. An explanatory sequential mixed methods design using Delphi is a common approach to gain experts' insight into why items might have exhibited differential item functioning (DIF) for a sub-group, indicating *potential* item bias. Use of Delphi *before* quantitative field testing to screen for potential sources leading to item bias is lacking in the literature. An exploratory sequential design is illustrated as an additional approach using a Delphi technique in Phase I and Rasch DIF analyses in Phase II. We introduce the 2 × 2 Concordance Integration Typology as a systematic way to examine agreement and disagreement across the qualitative and quantitative findings using a concordance joint display table. A worked example from the development of the Problem-Solving Measures Grades 6–8 Computer Adaptive Tests supported using an exploratory sequential design to inform item refinement. The 2 × 2 Concordance Integration Typology (a) crystallized instances where additional refinements were potentially needed and (b) provided for evaluating the distribution of bias across the set of items as a whole. Implications are discussed for advancing data integration techniques and using mixed methods to improve instrument development.

1. Objective

The purpose of this methodological illustration is to advance methods used in assessment construction and validation through demonstrating a new technique for integrating findings from the Delphi method ("Delphi"; Dalkey and Helmer, 1963; Helmer, 1967) and Rasch (1960, 1980) differential item functioning (DIF) analysis using an exploratory sequential mixed methods design (Creswell and Plano Clark, 2018). This research resituates Delphi as a mixed method grounded in *dialectical pluralism* described as a meta-paradigm that "concurrently and equally value [s] multiple perspectives and paradigms" (Johnson, 2017, p. 159). In doing so, this research aims to advance the use of mixed methods to study potential item bias when constructing and validating instruments, for which examples are limited (Gómez-Benito et al., 2018). A worked example supports the use of an *explor*atory sequential approach compared to the more commonly applied *explan*atory sequential mixed methods design. We introduce the 2 × 2 Concordance Integration Typology as a systematic way to examine agreement or disagreement across Delphi and DIF findings to inform next steps in item construction. As a result, this work advances methodological boundaries through (a) re-aligning Delphi with a more modern dialectical pluralism lens and (b) integrating common techniques (Delphi and DIF) in a novel way for exploring item bias. A worked example is used from the development of the Problem-Solving Measures (Bostic and Sondergeld, 2015, Bostic et al., 2017) Grades 6–8 Computer Adaptive Tests (PSM 6–8 CAT). Four research questions were addressed across a two-phase design followed by integration of findings from the Delphi study (Phase I) and DIF study (Phase II).

1. What sources of bias were identified by experts for the PSM6-8 CAT prototype items as potentially leading to item bias? (Phase I)

* Corresponding author.
E-mail addresses: kk3436@drexel.edu (K.L.K. Koskey), tas365@drexel.edu (T.A. May), yf366@drexel.edu (Y.". Fan), dnb66@drexel.edu (D. Bright), gregory.stone@utoledo.edu (G. Stone), gmatney@bgsu.edu (G. Matney), bosticj@bgsu.edu (J.D. Bostic).

https://doi.org/10.1016/j.metip.2023.100117
Received 29 August 2022; Received in revised form 26 January 2023; Accepted 24 February 2023
Available online 1 March 2023

2. To what extent did the expert item bias panel members agree on the sources of bias identified as potentially leading to item bias among the PSM6-8 CAT prototype items? (Phase I)
3. Did PSM6-8 CAT prototype items exhibit statistically significant DIF? (Phase II)
4. In what ways were the item revisions based on Delphi findings supported or not supported by DIF results to inform further PSM6-8 CAT prototype item refinement? (Integration)

1.1. Defining the Delphi method

Delphi is a consensus-building technique originally developed by the RAND Corporation in the 1950s to collect experts' opinions to aid in government decision-making and predict future events (Dalkey and Helmer, 1963; Linstone and Turoff, 1975). "Expert" is defined within the context of the phenomenon or issue under study. The premise is to "identify, learn, and share the ideas of experts by searching for agreement among experts" (Yildirium and Büyüköztürk, 2018, p. 451). Although often referred to as Delphi in general, this method has been classified into four types, each serving different purposes (Paré et al., 2013): Classical Delphi (consensus building), Policy Delphi (identifying differentiated views on social and political issues), Decision Delphi (determining support for decisions), and Ranking-type Delphi (ranking issues to inform action steps or strategic planning). Regardless of the form adopted, Delphi moves beyond a typical questionnaire about "what is" to explore "what could/should be" (Paré et al., 2013, p. 213).

Irrespective of the Delphi study type, purposive sampling is used to recruit individuals with specific expertise or meeting certain inclusion criteria to participate as a panel (Yousuf, 2007). For example, if the topic relates to social issues within a local community, participants could be a diverse group of community stakeholders. As another example, key decision-makers from one organization could serve as a panel if the topic is decision-making within a particular entity. As a final example, when seeking consensus on items representing a construct for a newly developed instrument, researchers and practitioners with expertise in that area, as well as individuals from the target population might be recruited as participants. Expert panels thus can share homogenous characteristics or be a diverse group depending on the context and purpose of the Delphi study. While Zunder and Islam (2011) recommended a minimum of 10 experts, no specific sample size criteria are adopted in the literature as Delphi "does not depend on statistical power but rather on group dynamics to arrive at a consensus among experts" (Cafiso et al., 2013, p. 256).

While often described as primarily a qualitative method (e.g., Brady, 2015; Sekayi and Kennedy, 2017), Delphi often involves integrating quantitative and qualitative data (Brady, 2015). Initial expert responses are collected on the topic or issue using an open-ended questionnaire. Expert opinions are analyzed thematically and aggregated by researchers to subsequently present to the same expert panel to review the synthesized results for indicating their level of agreement or disagreement. Multiple rounds are conducted until reaching a consensus that represents the collective expert opinion. Experts can modify their responses in each round. Modifications might occur after being exposed to other experts' perspectives or to clarify an opinion. Someone external to the panel, often a researcher, facilitates the process and responses are anonymous to other experts.

Analyses are embedded after each round of data collection to determine when saturation and consensus are reached. Variations exist in how experts indicate their agreement (e.g., open-ended response, rated on an agreement scale) and how consensus is determined (e.g., median agreement ratings, percent agreements, interquartile ranges, and non-parametric statistical tests such as Kendall's *W*). Despite these variations, the core components of Delphi are "purposive sampling, emergent design, anonymous and structured communication between participants, and thematic analysis" (Brady, 2015, p. 3). In their systematic review of 42 applications of Delphi, Paré et al. (2013) found that the rigor of the method is commonly upheld through purposeful sampling, anonymity of experts, an external facilitator, a stopping rule, and controlled iterative feedback to experts.

1.2. Advantages and limitations of Delphi

Delphi is described as appropriate to use as a method when: (a) a research problem calls for subjective collective feedback for which analytical techniques lack; (b) diverse backgrounds, experience, or expertise are needed to inform a problem; (c) frequent face-to-face panel meetings are not feasible or resources are limited; (d) external facilitation of communication is needed to maintain anonymity on a polarized or sensitive topic; and (e) it is necessary to create a space for individuals to share their voice without anyone dominating the group dynamic (Yousuf, 2007).

Potential limitations associated with Delphi include but are not limited to: (a) consensus reached representing a compromised or middle-of-the road opinion; (b) time commitment required by experts to complete the process; and (c) opinions representing a select group (Yousuf, 2007). An additional challenge implementing a Delphi study is maintaining the meaning of experts' opinions when synthesizing or conducting thematic analysis throughout the iterative process (Yousuf, 2007).

1.3. Applications of Delphi

Since its inception, Delphi has been used in a wide range of domains such as business, education, engineering, health care, information technology, policy, public administration, and physical sciences (Brady, 2015; Paré et al., 2013; Yousuf, 2007). More specific applications include using Delphi to: (a) forecast trends and conduct cost-benefit analysis (see for review Green, 2014); (b) gain an understanding of lived experiences (e.g., Sekayi and Kennedy, 2017); (c) engage in participatory action research to study an organizational system (e.g., Fletcher et al., 2014); (d) identify priorities and gather opinions to guide initiatives (see for review Paré et al., 2013) or public planning (Cafiso et al., 2013); and (e) establish content validity evidence in behavioral (Wessels et al., 2022), diagnostic (e.g., Bannatyne et al., 2018), or educational (e.g., Yildirim and Büyüköztürk, 2018) assessment development.

1.4. Uses of Delphi in instrument development

Ways in which Delphi has been used specific to instrument development is synthesized in Table 1. Other qualitative methods (e.g., cognitive interviewing) are included in Table 1 to situate Delphi as an alternative approach used for similar research purposes in instrument development. Example studies are cited to illustrate the pattern we observed in our review of the literature regarding the timing and purpose of Delphi or other qualitative method. Three ways the current methodological illustration advances and fills a gap in existing applications of Delphi in the psychological and educational instrument construction process can be drawn from Table 1 and are outlined next.

1.4.1. Advancement of Delphi to examine item bias

Delphi is more commonly used as a technique to collect expert opinions related to the phenomenon under study to inform item development (see Table 1). For instance, Bannatyne et al. (2018) applied the method to reach consensus on symptoms of eating disorder during pregnancy to identify item content of a new measure. As another example, Delphi is used to facilitate reaching a consensus on the quality of newly constructed items to make decisions on item refinement or removal. In these studies, experts are asked to iteratively review items to reach consensus on the (a) alignment with and representativeness of an item set as a whole with the construct intended to be measured; (b)

Table 1
Delphi and common qualitative methods used in instrument development.

Timing of Delphi or Qualitative Strand in the Research Design				Purpose			Sources of Bias Examined Using Delphi		Example Citations
Mono	Simultaneous	Phase I	Phase II	Inform item content	Confirm or explain bias	Explore item bias	Pre-defined	Open	
Delphi				X			N/A		Bannatyne et al. (2018) Bauer et al. (2019) Bonnot et al. (2022) Wessels et al. (2022)
		Delphi		X			N/A		Al Zoubi et al. (2018) Dutt et al. (2019)
	Delphi				X		X	X	Karnati (2021)
			Delphi		X		X		Yildirim and Büyüköztürk (2018)
	Ethnographic interviews				X		X		Maddox et al. (2015)
			Cognitive interviews		X			X	Benítez and Padilla (2014)
			Think aloud interviews		X		X		Ercikan et al. (2010)
		Delphi				X	X	X	Current illustration

Notes. Mono (single method study). Simultaneous (collected at the same time as quantitative data). Phase I (implemented in Phase I followed by a quantitative Phase II). Phase II (implemented in Phase II and proceeded by quantitative Phase I). Pre-defined (examined sources of bias between specified sub-groups often defined by prior significant DIF results). Open (examined emergent sources of bias beyond for pre-defined sub-groups).

clarity of wording; (c) appropriateness of the rating scale; and/or (d) how items discriminate from other existing instruments (e.g., Bauer et al., 2019; Bonnot et al., 2022; Dutt et al., 2019; Wessels et al., 2022).

When used to examine for measurement invariance, common is to use Delphi or other qualitative methods to confirm or explain statistically significant DIF results (e.g., Karnati, 2021; Yildirim & Büyüköztürk, 2018). Delphi has been implemented in combination with DIF where a sample is drawn from the target population to test for potential item bias against a subgroup (e.g., Yildirium and Büyüköztürk, 2018). DIF results when "equally able test takers differ in their probabilities to answer a test item correctly as a function of group membership" (AERA et al., 2014, p. 51). As such, DIF indicates systematic error as compared to real mean group differences (Camilli and Shepard, 1994). While significant DIF does not equate to item bias, it can be indicative of *potential* bias and thereby important to investigate (Osterlind and Everson, 2009). DIF is examined to understand measurement variance (i.e., inequivalence), particularly possible sources of construct-irrelevant variance (Boer et al., 2018). As defined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), construct-irrelevant variance is "variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation" (p. 217).

Three types of potential biases influencing invariance are construct, method, and item bias (Boer et al., 2018, van de Vijver and Tanzer, 2014). Relevant to this paper is item bias, an unfavorable impact on a particular sub-group. This can cumulate to result in test bias occurring "when a decision, grounded to some degree by the scores yielded from a test, is unfair or has a perceived disparate impact on one group" (Osterlind and Everson, 2009, p. 4). Thus, item bias negatively influences consequential validity evidence, the interpretation and intended use of test scores (AERA et al., 2014). While Delphi has been used to examine measurement invariance (e.g., Karnati, 2021; Yildirim & Büyüköztürk, 2017), applications such as the current proposed methodological illustration are less common than its use for informing item content and clarity.

1.4.2. Use of Delphi to explore item bias beyond for only pre-defined groups

Studies using Delphi or other methods to examine DIF have advanced understanding of sources by demographic groups (e.g., gender, ethnical/racial, and school type). A limitation of this research is that only items exhibiting statistically significant DIF by pre-defined sub-groups are further investigated to explain why DIF resulted to inform item refinement. Cognitive interviewing (e.g., Benítez and Padilla, 2014), think aloud interviewing (e.g., Ercikan et al., 2010), focus groups, (e.g., Yildirium and Büyüköztürk, 2018), and Delphi (e.g., Yildirium and Büyüköztürk, 2018) have been used as follow-up methods. In adopting this explanatory sequential approach, sources of biases are narrowly described within a demographic-group framework in both phases of the study. As a result, potential biases related to other variables and compounding biases against crossing aspects of groups (e.g., girls from under-resourced schools) are left unexplored (Lyons et al., 2021).

Also, items not presented to panel members based on DIF results might have potential sources of biases that can be minimized. However, because they did not yield statistically significant DIF, these items are often not examined further. One exception is a study conducted by Karnati (2021) examining all items for both DIF and potential sources of bias identified by an expert panel through Delphi. However, only bias for pre-defined groups were examined including for gender and English language learner status. The methodological illustration presented in this paper proposes to fill this gap through use of an exploratory sequential mixed methods design (Creswell and Plano Clark, 2018).

1.4.3. Advancement of a systematic data integration technique

Systematic approaches to determining DIF and evaluating for consensus in follow-up Delphi studies conducted in effort to understand the DIF detected are used in existing applications. Despite these advances, to our knowledge, a systematic approach to integrating the joint results has not been developed to date. Delphi is often implemented or outcomes are reported as a stand-alone method (e.g., Bannatyne et al., 2018; Bauer et al., 2019; Green, 2014; Fletcher and Marchildon, 2014). When comparing qualitative descriptors of bias to DIF results, data integration was limited to merging through discussion (e.g., Maddox et al., 2015).

One exception to the latter limitation was Yildirium and Büyüköztürk's (2018) study comparing Delphi and DIF results. Although the specific mixed methods data integration technique was not explicitly explained by the researchers, qualitative sources of bias described by experts were reported using a joint display table. However, only results for items exhibiting DIF were jointly reviewed with Delphi results and for pre-defined groups (i.e., gender and school type). A second exception

to the latter limitation was a recent dissertation conducted by Karnati (2021). Although the mixed methods data integration technique was also not explicitly described in this study, Karnati (2021) presented a joint display table to report expert panel consensus and DIF results for the total item set. While joint display tables were implemented in both of these two studies, they were not used to determine concordance/discordance across data sources. The current methodological illustration is distinguished from these prior design approaches by presenting a systematic approach for joint review of Delphi and DIF results. Our developed typology advances the limited examples of joint displays for comparing the concordance/discordance of Delphi and DIF results through a formal framework to inform next steps in minimizing sources of item bias as part of instrument development.

1.5. Current methodological illustration

Mixed methods are used to examine potential item bias (e.g., Benítez & Padilla, 2016; Gadermann et al., 2011; Maddox et al., 2015), but examples continue to be scarce (Gómez-Benito et al., 2018). Delphi has been implemented in combination with DIF analyses as a systematic technique for collecting data to explain DIF results (see Fig. 1a). Lacking is an approach for integrating Delphi *before* quantitative testing of items (see Fig. 1b).

Flipping the purpose from initially *explain*ing to first *explor*ing potential sources of bias to guide item writing could minimize bias *before* exposing the target population to items and further advise item writing and refinement in advance of participants completing items. Also lacking is integration of Delphi findings and DIF results to inform the item refinement process. A new integration typology for connecting data from the two methods is proposed in this paper.

2. Resituating Delphi under dialectic pluralism

We adopted a dialectic pluralistic (Johnson, 2012, 2017) lens to guide the development of a more integrated approach to examining item bias. Delphi was initially grounded in pragmatism (Dalkey and Helmer, 1963) described as the "what methods work" paradigm (Creswell and Plano Clark, 2018). It was associated with a pragmatic lens due to the utilization of qualitative methods to collect expert opinions and quantitative methods for determining consensus. Since the original inception of Delphi, however, research paradigms have advanced to include dialectic pluralism "as a metaparadigm (that is, a paradigm that dialogues with multiple paradigms)" (Stefurak et al., 2016, p. 345). Dialectic pluralism is explained as engaging in the following six activities during the research process for which, we maintain, Delphi's alignment is apparent:

> (a) Dialectically listen, carefully and thoughtfully, to different paradigms, disciplines, theories, and stakeholder and citizen perspectives; (b) combine important ideas from competing paradigms and values into a new workable whole for each research study or program evaluation; (c) explicitly state and "pack" the approach with stakeholders' and researchers' epistemological and socio-political values to guide the research (including the valued ends and one hopes for the valued means for getting there); (d) conduct the research ethically; (e) facilitate dissemination and use of research findings (locally and more broadly); and (f) continually, formatively evaluate and improve the outcomes of the research-an-use process … In short, [dialectic pluralism] is a change theory, and it requires listening, understanding, learning, and acting. (Johnson, 2012, p. 752)

Core components of Delphi are consistent with the main ideologies of dialectic pluralism: "negotiation, conflict management, and group process for dealing with differences" (Johnson, 2017, p. 158). Delphi involves listening to a diverse panel of experts or stakeholders, gaining an understanding a phenomenon through iterative consensus building, learning from the thematic analysis results, and acting on those results to inform a change. Thus, we propose resituating Delphi in this paradigm to move beyond its initial grounding in pragmatism.

Adopting this lens, in general, was appropriate for the current methodological illustration. Over the past decade, mixed methods research has been increasingly associated with dialectic pluralism (Johnson et al., 2014; Stefurak et al., 2016). In addition, mixed methods are called for in instrument construction and validation studies to use multiple quantitative and qualitative data sources to iteratively inform instrument refinements and provide for a robust validity argument (Koskey et al., 2018; Luyt, 2012; Onwuegbuzie et al., 2010).

3. Methods

3.1. Study context

A worked example was drawn from Year 1 of a 5-year grant supporting the development of the PSM6-8 CAT. These items are open word-problem type tasks based on real-world scenarios. The *Marshmallow Treats Tasks* is an example from PSM7.

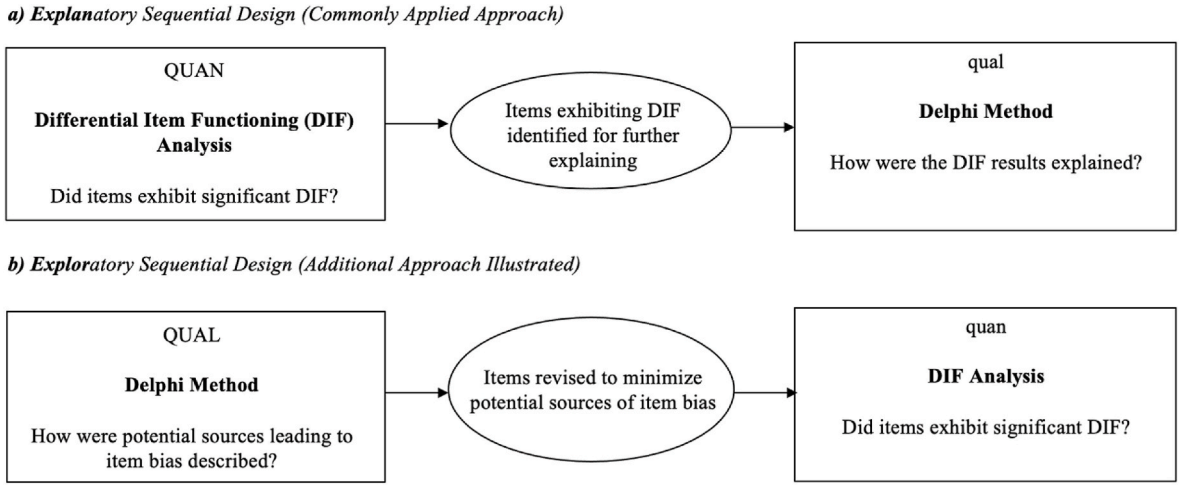

Fig. 1. Common sequential mixed method research designs used to evaluate for potential item bias. Illustration adapted from Creswell and Plano Clark (2019, p. 68). QUAN/quan and QUAL/qual indicate higher/lower priority status in addressing the research questions. Variations in priority status assigned for both designs can be supported (Morgan, 1998; Teddlie and Tashakkori, 2009). Arrow (→) indicates sequential.

Rosalinda plans to make marshmallow treats to share with her class. The recipe requires 2 $^{1}/_{8}$ cups of marshmallows, 3 ¼ cups of rice cereal, and ¼ cup of butter. It serves eight people. How many cups of rice cereal will she need if she must make treats for 28 people?

Because writing PSM items involves drawing from real-world settings, efforts are needed to minimize biases reflected in item content potentially shaped by item writers' own experiences.

An iterative Design-Based Research methodology (Middleton et al., 2008) approach to instrument construction and validation was adopted in the larger study. This current work is related to initial item writing construction phases. Content expert, psychometrician, and item bias review were built into the iterative item writing cycle. While this worked example is in the context of the development of a mathematics assessment, the proposed design and integration typology can transfer in general to educational and psychological instrument construction and validation studies examining item bias.

3.2. Study constraints

Constraints included that results had to be communicated to the item writing team within a three-week period to inform item refinement. Delphi was identified as a systematic yet efficient method for item bias review. However, Delphi post hoc for explanatory purposes dominated the literature. As a result, the goal was to develop an approach to integrate Delphi and DIF analyses to conduct a rapid evaluation of potential item bias before exposing items to the target student population for further qualitative and quantitative field testing.

3.3. Proposed exploratory sequential approach

Fig. 2 illustrates the proposed two-phase exploratory sequential mixed methods research design (Creswell and Plano Clark, 2018) nested in the larger study. The four research questions addressed are presented by phase. Overall, potential sources of item biases were explored in an open-ended approach using the Delphi method (Phase I) and quantitative DIF analyses (Phase II) followed by integration of findings from the two phases.

Two research questions bounded the Delphi study (Phase I). Research question one (RQ1) addressed experts' opinions related to potential sources of bias in the PSM-CAT 6–8 prototype items, while research question 2 (RQ2) determined the extent they reached consensus on those sources of bias. Findings from Phase I were communicated to the item writing team to guide item refinements as part of the item writing cycle. Also, these findings were used to inform strategies shared with item writers in the subsequent year's refresher item writing training related to minimizing potential sources of item bias. Following item refinements by content experts, items were advanced and tested for DIF (Phase II), which addressed research question three (RQ3) related to evaluating for significant DIF with quantitative methods. Research question four (RQ4) addressed integration of the results from both phases to inform further item refinement. More specifically, integration examined in what ways the DIF (Phase II) results supported that item revisions made were effective or that additional refinements may be needed prior to further field testing. Jointly analyzing the results provided for crystallization, a "deepened complex, thoroughly partial, understanding" (Richardson, 2000, p. 934) of potential biases associated with items.

3.4. Data collection and analyses

3.4.1. Delphi study (Phase I)

The Delphi study consisted of two item bias panels each made up of three experts purposefully sampled to represent expertise in diversity and mathematics assessment on each panel, as well as a diverse demographic make-up by gender and ethnic/racial identity. Engagement of a diverse expert panel is consistent with a core ideology of dialectic

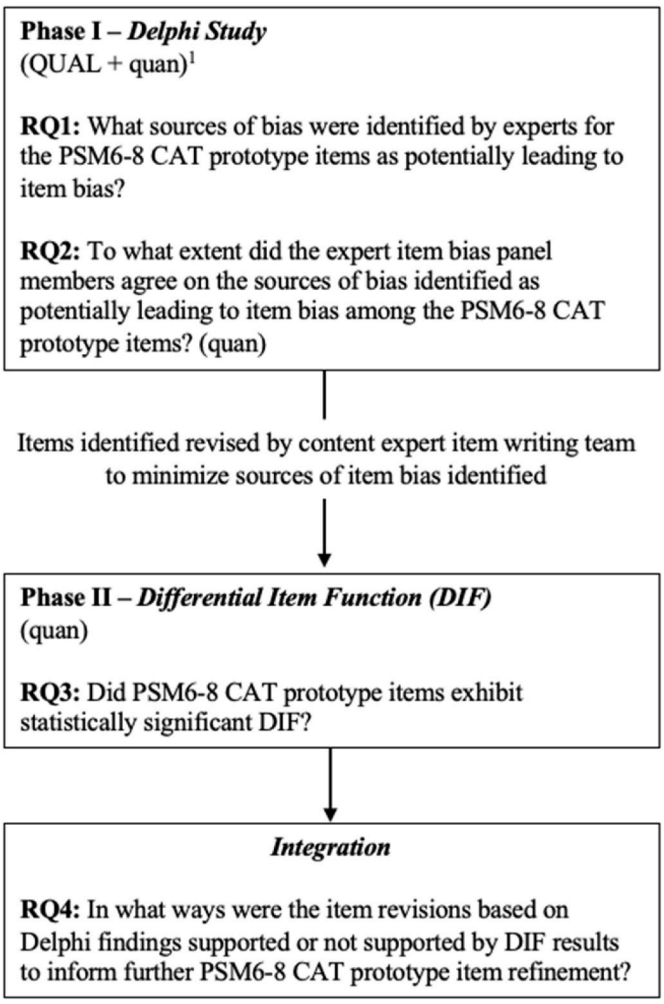

Fig. 2. Exploratory sequential mixed methods design using the Delphi method and Rasch differential item functioning analysis to evaluate for potential item bias. [1]Uppercase/lowercase indicates priority status as higher/lower (Morse, 1991, 2003). RQ = Research question.

pluralism to "pack" the research approach to reflect multiple stakeholders' values (Johnson, 2012). Each item bias panel reviewed 180 items split into two sets (averaging 45 items) for a total of 360 items with each set undergoing the Delphi method. Our decision to create two item bias panels was in effort to present a task feasible for experts to complete two iterations within the given timeframe. The Delphi method utilized is detailed in a prior work (Koskey et al., 2021).

3.4.1.1. Round 1 of the Delphi study. Experts completed an online survey asking them to: (a) describe any potential sources of item bias; (b) identify the type(s) of source: cultural, disability status (e.g., dyscalculia), gender, geographical area (suburban, urban, or rural), race/ethnicity, school type (private, public), socio-economic status, or other; and (c) assign a holistic rating of the level of accessibility of the items to the target population (not accessible, accessible to some, accessible to most, accessible to all) and explain the primary reason for their rating. After collecting experts' opinions on each item, two researchers collaboratively conducted a rapid thematic analysis (Saldaña, 2016) to identify common or unique potential biases identified by the panel by item. Biases coded as common were synthesized into a single bias statement, while unique biases were retained as separate bias statements.

3.4.1.2. Round 2 of the Delphi study. Experts were sent a second survey

that synthesized the biases (based on the thematic findings) and asked them to complete three core tasks: (a) rate whether they agreed/disagreed "this source could potentially lead to item bias; " (b) describe any modifications to existing biases and any additional biases emerging in their second review; and (c) rate the level of items accessibility to the target population. Round 2 served as a way to collect any new emergent biases, and it was a strategy for member-checking the synthesized bias statements. Throughout the Delphi study, experts' verbatim wording from their written responses was retained to the highest degree possible when stating the biases so to maintain the meaning of the experts' opinions. It was important to adopt experts' language when possible to maintain the meaning since additional iterations to validate their views was not possible.

3.4.1.3. Item refinements based on Delphi results. As the Delphi process was completed for each set by each panel, the results were iteratively reported using three feedback buckets: (a) items with consensus reached that no bias was identified; (b) items with consensus reached that bias was identified; and (c) items with one expert identifying bias. Consensus was defined as at least two of three experts agreeing a source could potentially lead to item bias. The item writing leads subsequently used the results to guide item refinements. Using results to inform a change is consistent with both the purposes of a Delphi study (Yousuf, 2007). Also, findings were used at a local level (to guide item refinements to PSM-CAT 6–8 items) and broadly (to advance methods for minimizing measurement variance), as Johnson (2012) described as a core ideology of dialectic pluralism. Further consistent with dialectic pluralism is the iterative and reciprocal use of the outcomes of the Delphi study to improve our research on the assessment development process, which reflects to "continually, formatively evaluate and improve the outcomes of the research-and-use process" (Johnson, 2012, p. 752).

3.4.2. DIF study (Phase II)

A series of 18–21 item tests were created and administered to college students in mathematics education programs. The purpose of the initial pilot testing was to uncover potential fundamental interpretive and performance issues prior to testing with middle school students to avoid test fatigue and for time constraints in classrooms. DIF (Phase II) was tested analyses applying the dichotomous Rasch model, a probabilistic measurement model (Bond and Fox, 2007). Using WINSTEPS© Version 4.4.6 (Linacre, 2019), the dichotomous Rasch model was implemented as items on the PSM6-8 CAT were scored as either correct or incorrect. Person measures and item difficulties were transformed to log-odd units (i.e., logits).

DIF contrast sizes were computed to indicate the magnitude of differences in item difficulty between sub-groups. DIF contrast sizes $\geq$ 0.50 logits were moderate to large magnitude in differences, indicative of potential item bias (Linacre, 2021). Sub-groups recommended to test for DIF in educational assessment include gender, geographical location, socio-economic status, and racial/ethnic groups (AERA et al., 2014). For the purposes of this methodological illustration, we use a sub-set of exemplar items tested for DIF by geographical location (suburban vs. urban/rural). Urban and rural were combined to form one group as students from K-12 schools in these geographic areas have been found to have similar academic performance and are viewed as traditionally underserved in STEM education (Harris and Hodges, 2018).

3.4.3. Integration

Our first cycle of analysis was exploring in what ways the two data sources might be crystallized (Ellingson, 2009, 2014; Rasch, 1960, 1980) to identify next steps in item development. Delphi and DIF results were organized in a joint display table, a common data integration technique used in mixed methods to compare qualitative and quantitative data sources (Plano Clark, 2019). Guetterman et al. (2015) have also generally described this table as *comparing results display* and noted its appropriateness for use in test validation. Each item was treated as a unit of analysis with Phase I and Phase II results as columns for the tested samples. Data for each item included synthesized bias statements (Phase I), DIF contrast sizes (Phase II), and identification of which group was negatively impacted (Phase II).

Four typologies emerged in our collective examination of the joint display table. As such, this resulted in a concordance table often used in mixed methods research studies (e.g., Creamer and Reeping, 2020) and previously familiar in Delphi studies (e.g., Cafiso et al., 2013). Table 2 displays the developed 2 $\times$ 2 Concordance Integration Typology for evaluating in what way DIF (Phase II) supported item refinement/non-refinement based on Delphi (Phase I) results. Typologies are mutually exclusive (i.e., independence of observation) in that each item can only be classified in one typology.

Concordance Type I is defined as when an item revision was made in response to the Delphi (Phase I) results that the panel reached consensus the item exhibited potential bias and DIF (Phase II) was not detected. In this case, this joint result provides evidence the item revision as effective at minimizing item bias for geographical location. *Concordance Type II* is defined as when no item revision was necessary based on the Delphi (Phase I) results and DIF (Phase II) was not detected. This joint result supports bias was minimized.

Discordance Type I is defined as when an item revision as made in response to the Delphi (Phase I) results that the panel reached consensus the item exhibited potential bias, but DIF (Phase II) was detected after item revision. *Discordance Type II* is defined as when no item revision was necessary based on the Delphi (Phase II) results, but DIF (Phase II) was detected. Discordance joint results provide evidence potential item bias may persist and items may need further review and/or refinement (see Table 2).

In a second cycle of analysis, each item was classified applying the 2 $\times$ 2 Concordance Integration Typology. The joint display table was expanded to reflect the typology classification affiliated with each item, providing for further crystallization of potential item bias.

4. Results

4.1. Delphi study results (Phase I)

Full results of the Delphi study of 360 items including the thematic findings are reported in another paper (Fan et al., 2022). For the purpose of this methodological illustration, we present results for a sub-sample of 30 PSM6-8 CAT prototype items cycling through the item development phase of the larger study. We focus on results related to students self-reported geographical area. Out of the 30 items, the panels reached consensus that a total of 15 items (50.00%) exhibited potential sources of bias with 9 items (30.00%) specifically identified as geographical type bias. As one example, panel members reached consensus that an item referencing a "ride sharing app" in the problem-solving task exhibited a potential source of geographical bias. Two panel members identified this source in Round 1 of the Delphi study. All three panel members agreed on the synthesized bias statement the item exhibits potential biased specifically against rural students in that, "Very few rural students would have experience with a ride sharing app."

Table 2
2 $\times$ 2 Concordance Integration Typology for Connecting Delphi and DIF Results.

Post Delphi (Phase I)	DIF (Phase II) Results	
	No significant DIF	Significant DIF
Item revision	Concordance Type I	Discordance Type I
No item revision	Concordance Type II	Discordance Type II
Integration	↓	↓
	Item revision effective. Supports bias minimized.	Evidence potential bias persists. Item may need further review.

4.3. DIF analysis results (Phase II)

Among the 30 items, a total of 11 items (36.67%) showed statistically significant DIF contrast sizes $\geq$ 0.50 logits, indicating moderate to large magnitude in systematic differences between geographical groups. DIF contrast sizes ranged from −2.84 to 1.93. A total of six items (20.00%) favored students who self-reported living in a suburban area, while five items (16.67%) favored students who self-reported living in a rural or urban area. These results indicated that when DIF was detected, it was relatively equally balanced between the two sub-groups across the 30 items. Items exhibiting statistically significant DIF were subsequently communicated to the item writing team for further review.

4.4. Integration

The frequency and percentage of items classified in each Concordance Integration Typology follows: Concordance Type I (Item revised and no DIF) – 9 items (30.00%) Concordance Type II (Item not revised and no DIF) – 10 items (33.33%), Discordance Type I (Item revised but DIF) – 6 items (20.00%), and Discordance Type II (Item not revised but DIF) – 5 items (16.67%). Based on these results, a total of 19 items (63.33%) exhibited concordance, while a total of 11 items (36.67%) classified as discordance and potentially needing additional review. For illustration purposes, Table 3 reports on results of four exemplar items using a concordance joint display classifying a typology as present "X".

5. Discussion and conclusion

Integrated results can be used to identify next steps at the item level and assessment level. In our study, points of concordance provided supportive evidence for continuing to move an item through the larger item development cycle and field-testing phase. Points of discordance highlighted the importance of evaluating both data sources to draw meta-inferences regarding the next steps in item development. Meta-inferences are conclusions "generated by integrating the inferences obtained from the qualitative and quantitative strands" (Teddlie and Tashakkori, 2009, p. 338). Discordance can be examined through multiple lenses. Example potential next steps at the item level to consider include.

1. Retain and monitor an item.
2. Revise (or further revise) an item.
3. Conduct further expert review.
4. Conduct cognitive interviews with target population.
5. Remove an item based on crystallization of findings.

Which step to engage in depends on critical review of the joint results, as well as the item content. For example, item 3 in Table 2 originally included context of calculating a "tip" for a restaurant bill. Panel members reached consensus that, "A student who is not financially able to afford to go out to eat may not know what a tip is. Tipping is cultural as well (more Western). In some cultures, tip is simply included in the cost of the bill." After communicating the results to the item writing team lead, the item context was revised. Despite the item revisions, statistically significant DIF was detected in favor of students identified as from a suburban area (*Discordance Type I*). This discordance may indicate that the revision of "tip" to "coupon" and "restaurant" to "store" did not address the potential source leading to item bias. These results can be communicated to the item writing team that the item context may need further revision.

Table 3
Joint display of example findings using the 2 × 2 concordance integration typology.

Item	Post Delphi (Phase I) Results	DIF (Phase II) Results		Joint Results			
		DIF Contrast	Group Favoring	Concordance		Discordance	
				Type I	Type II	Type I	Type II
1	Item revised	0.18^{NS}	–	X			
2	No item revision	0.06^{NS}	–		X		
3	Item revised	−1.02*	Suburban			X	
4	No item revision	−1.48*	Rural/ Urban				X

Note. ^{NS} = Not statistically significant. * = Statistically significant ($p \leq 0.05$). X = Type present.

Results at the assessment level can be used in two ways. First, to evaluate the overall balance of bias across the assessment as a whole. For instance, examining whether there is an even distribution of items exhibiting potential bias against students from suburban or urban/rural areas. Evaluating the balance of DIF at the assessment level is appropriate given that sources of bias cannot be fully eliminated (Osterlind and Everson, 2009), but an assessment with minimized or balanced DIF across sub-groups is fairer than a largely unbalanced test in terms of DIF. For the 30 item sub-sample used in the current illustration, as a whole approximately the same percentage of items were biased against suburban (16.67%) and rural/urban (20.00%) students. Second, results can be utilized to inform continuous item writer training in an effort to minimize potential sources in future item construction. In our study, we shared with item writers' common types of biases identified and strategies for minimizing those biases grounded in both literature and our specific findings.

5.1. Advancement of mixed methods

Minimizing item bias is important to increase fairness in testing (AERA et al., 2014) and, in turn, appropriate consequential uses of results such as in making cross-cultural comparisons (Boer et al., 2018). Examples integrating DIF with qualitative data sources exist but are still few (Gómez-Benito et al., 2018), and they dominantly use an explanatory sequential design. The current work fills a gap in the literature by illustrating the advantages of integrating Delphi and DIF using an exploratory sequential design in the item construction phase of instrument development and validation. In this example, DIF was not only an indicator of potential sources leading to item bias, but also whether item revisions based on Phase I Delphi findings were effective. Intended test takers' differing backgrounds "… must be considered throughout all stages of development … so that barriers to fair assessment can be reduced" (AERA, 2014, p. 50). Thus, this exploratory approach and the 2 × 2 Concordance Integration Typology developed demonstrated the flexibility in the use of the Delphi method as part of the feedback loop in the item writing cycle of assessment development.

Additionally, modeled after concordance joint displays (Cafiso et al., 2013; Guetterman et al., 2015), the typology developed through this research contributes to data integration techniques in mixed methods by presenting an approach to jointly evaluating Delphi and DIF results. Table 2 provides a template for future studies applying this framework. The illustrated design and typology are transferable to the development process in general for educational and psychological assessments. Framed within a dialectic pluralism paradigm (Johnson, 2017), drawing meta-inferences based on both data sources (Delphi and DIF results) reflects a more holistic evaluation of bias from multiple perspectives during the item writing cycle.

5.2. Conclusions and future research

The findings from this study contributes to advancing integration strategies in mixed methods, thus responding to the "integration challenge" described in mixed methods research (Fetters and Freshwater,

2015). Joint display tables are commonly used in convergent designs examining convergence or divergence across data sources (Bustamante, 2019). The current paper contributes to mixed methods research by extending the use of joint displays in sequential designs. A 2 × 2 Concordance Integration Typology can be applied adopting the definitions used in this study or modified to fit other study purposes and questions comparing qualitative and quantitative data sources for concordance. In doing so, we advocate for future research to combine the dialectical pluralism perspective with Delphi.

We identified lessons learned as researchers applying an exploratory sequential mixed methods design to integrate Delphi and DIF results. First, a diverse panel was essential for the Delphi study (Phase I) to increase the likelihood that underlying sources of bias will be identified that might not be detected by DIF but warrant item refinement. While there were instances where members identified a common source of bias in the first iteration of review, there were many occasions where individual experts identified biases not first recognized by other experts. Second, engaging in the Delphi study multiple times seemed to influence panel members' detection of sources of bias. They were exposed to multiple critical perspectives over time as inherent in using the Delphi method. We observed the evolution of the types of sources of biases described by experts with their increased exposure to the opinions of experts from different backgrounds. These first two lessons learned illuminated the importance in engaging multiple perspectives in the research process, consistent with dialectic pluralism (Johnson, 2012). Future research might focus on a systematic mixed methods study related to whether engaging in a Delphi study being exposed to others' critical perspectives influences how experts see (re-see) potential sources of bias in subsequent reviews.

Third, conducting the Delphi study in Phase I as compared to Phase II casted a broader net for identifying sources of item bias than allowing the DIF results to drive the initial phase of the study using pre-defined categories to explore (e.g., gender, racial/ethnicity). As a parallel, the benefits were similar to when applying emergent coding as compared to *a priori* coding (Saldaña, 2016) in qualitative research. Fourth, and finally, the joint review of Delphi and DIF results informed actions such that only considering one data source would have potentially altered decisions about how an item progressed through the item writing cycle.

DIF analysis is advocated for as a "routine part" in developing instruments (Martinková et al., 2017, p. 1). We firmly support the engagement of diverse critical perspectives in screening for sources of item bias be routine *before* exposing participants to items on a larger scale. The Delphi technique provides a systematic approach to that end. Researchers are encouraged to consider the value of incorporating a Delphi study in the early stages of research given the added information not understood from quantitative approaches alone.

5.3. Limitations

A limitation of this research was that only two iterations were feasible to conduct as part of the Delphi study. Member-checking and verbatim language from experts were used to address this limitation. Another limitation is that only one demographic variable was evaluated at a time utilizing DIF analysis. Biases identified by experts in this study emerged as bias against a single group (e.g., girls, students living in urban spaces), which could be due to Delphi directions not explicitly asking panel members to reflect on biases crossing over variables. As a result, intersectional biases (girls living in urban spaces) were not quantitatively tested. Future research should explore any emergent qualitatively described cross-intersectional biases using advanced modern statistical techniques available (see for review, Boer et al., 2018).

Credit author statement

Kristin L. K. Koskey: Funding acquisition; Conceptualization; Methodology; Data collection; Formal analysis; Lead writing of the original and revised manuscript. Toni A. May: Funding acquisition; Conceptualization Methodology; Data collection; Formal analysis; Contributing to writing the original and revised manuscript. Yiyun "Kate" Fan: Data collection; Formal analysis; Contributed to writing the original and revised manuscript. Dara Bright: Formal analysis; Contributed to writing the original and revised draft. Gabriel Matney: Funding acquisition; Conceptualization; Data Collection; Review of the original and revised manuscript. Jonathan D. Bostic: Funding acquisition; Conceptualization; Review of the original and revised manuscript. Gregory Stone: Funding acquisition; Formal analysis; Review of the original and revised manuscript.

Funding

This work was supported by the National Science Foundation [#2100988, #2101026]. Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Toni A. May (Sondergeld) (Grant PI), Jonathan D. Bostic (Grant PI), Kristin Koskey (Corresponding author, Grant Co-PI), Gabriel Matney (Co-Author, Grant Co-PI), and Gregory Stone (Co-Author, Grant CO-PI) reports financial support was provided by National Science Foundation.

Data availability

The data that has been used is confidential.

References

Al Zoubi, F.A., Mayo, N., Rochette, A., Thomas, A., 2018. Applying modern measurement approaches to constructs relevant to evidence-based practice among Canadian physical and occupational therapists. Implement. Sci. 13 (1), 152 doi:10.1186-018-0844-4.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), 2014. Standards For Educational and Psychological Testing. AERA.

Bannatyne, A.J., Hughes, R., Stapleton, P., Watt, B., MacKenzie-Shalders, K., 2018. Signs and symptoms of disordered eating in pregnancy: a Delphi consensus study. BMC Pregnancy Childbirth 18 (262), 1–16. https://doi.org/10.1186/s12884-018-1849-3.

Bauer, S.M., Fusté, A., Andrés, A., Saldaña, C., 2019. The barcelona orthorexia scale (BOS): development process using the Delphi method. Eating and Weight Disorders – Studies on Anorexia, Bulimia and Obesity 24, 247–255. https://doi.org/10.1007/s40519-018-0556-4.

Benítez, I., Padilla, J.-L., 2014. Analysis of nonequivalent assessments across different lingusitic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. J. Mixed Method. Res. 8 (1), 52–68. https://doi.org/10.1177/1558689813488245.

Boer, D., Hanke, K., He, J., 2018. On detecting systematic measurement error in cross-cultural research: a review and critical reflection on equivalence and invariance tests. J. Cross Cult. Psychol. 49 (5), 713–734 doi:1177/0022022117749042.

Bond, T.G., Fox, C.M., 2007. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed. Lawrence Erlbaum Associates.

Bonnot, O., Insua, J.L., Walterfang, M., Torres, J.V., Kolb, S.A., 2022. Development of a suspicion index for secondary schizophrenia using the Delphi method. Aust. N. Z. J. Psychiatr. 56 (5), 500–509. https://doi.org/10.1177/00048674211025715.

Bostic, J.D., Sondergeld, T.A., 2015. Measuring sixth-grade students' problem-solving: Validating an instrument addressing the mathematics common core. School Sci. Math. J. 115 (6), 281–291.

Bostic, J.D., Sondergeld, T.A., Folger, T., Kruse, L., 2017. PSM7 and PSM8: Validating two problem-solving measures. J. Appl. Measur. 18 (2), 1–12.

Brady, S.R., 2015. Utilizing and adapting the Delphi Methods for use in qualitative research. Int. J. Qual. Methods 1–6. https://doi.org/10.1177/1609406915621381.

Bustamante, C., 2019. TPACK and Teachers of Spanish: development of a theory-based joint display in a mixed methods research case study. J. Mix. Methods Res. *13* (2), 163–178. https://doi.org/10.1177/1558689817712119.

Cafiso, S., Graziano, A.D., Pappalardo, G., 2013. Using the Delphi Method to evaluate opinions of public transport managers on bus safety. Saf. Sci. *57*, 254–263.

Camilli, G., Shepard, L.A., 1994. Methods For Identifying Biased Test Items. Sage.

Creamer, E.G., Reeping, D., 2020. Advancing mixed methods in psychological research. Methods in Psychology 3. https://doi.org/10.1016/j.metip.2020.100035.
Creswell, J.W., Plano Clark, V.L., 2018. Designing And Conducting Mixed Methods Research, 3rd ed. Sage.
Dalkey, N., Helmer, O., 1963. An experimental application of the Delphi methods to the use of experts. Manag. Sci. 9, 458–467. https://doi.org/10.1287/mnsc.9.3.458.
Dutt, A., Tan, M., Alagumalai, S., Nair, R., 2019. Development and validation. Of the ability in behavior assessment and interventions for teachers using Delphi technique and Rasch analysis. J. Autism Dev. Disord. 49, 1976–1987. https://doi.org/10.1007/s10803-019-03887-4.
Ellingson, L.L., 2009. Engaging Crystallization in Qualitative Research. Sage.
Ellingson, L.L., 2014. The truth must dazzle gradually": enriching relationships research using a crystallization framework. J. Soc. Pers. Relat. 31 (4), 442–450. https://doi.org/10.1177/0265407514523553.
Ercikan, K., Arim, R., Law, D., 2010. Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. Educational Measurement 29 (2), 24–35.
Fan, Y.K., Koskey, K.L.K., Bright, D., May, T.A., Matney, G., Bostic, J., 2022. Exploring sources of bias to improve the universal design of assessments of mathematical problem-solving skills [Manuscript submitted for publication]. School of Education, Drexel University.
Fetters, M.D., Freshwater, D., 2015. The 1 + 1 = 3 integration challenge. J. Mix. Methods Res. 9 (2), 115–117. https://doi.org/10.1177/1558689815581222.
Fletcher, A.J., Marchildon, G.P., 2014. Using the Delphi Method for qualitative, participatory action research in health leadership. Int. J. Qual. Methods 13 (1), 1018. https://doi.org/10.1177/160940691401300101.
Gadermann, A.M., Guhn, M., Zumbo, B.D., 2011. Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adaption for children: a focus on cognitive processes. Soc. Indicat. Res. 100, 37–60. https://doi.org/10.1007/s11205-010-9603-x.
Green, R.A., 2014. The Delphi Technique in Educational Research. SAGE Open, pp. 1–8. https://doi.org/10.1177/2158244014529773.
Gómez-Benito, J., Sireci, S., Padilla, J.-L., Dolores Hidalgo, M.D., Benítez, I., 2018. Differential item functioning: beyond validity evidence based on internal structure. Psicothema 30 (1), 104–109. https://doi.org/10.7334/psicothema2017.183.
Guetterman, T.C., Fetters, M.D., Creswell, J.W., 2015. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann. Fam. Med. 13, 554–561.
Harris, R.S., Hodges, C.B., 2018. STEM education in rural schools: implications for untapped potential. National Youth-At-Risk Journal 3 (1), 1–12. https://doi.org/10.20429/nyarj.2018.030102.
Helmer, P., 1967. Systematic Use of Expert Opinions (Report NO. P-3721). The Rand Corporation.
Johnson, R.B., 2012. Dialectical pluralism and mixed methods research. Am. Behav. Sci. *56*, 751–754. https://doi.org/10.1177/0002764212442494.
Johnson, R.B., Onwuegbuzie, A.J., Tucker, S.A., Icenogle, M.L., 2014. Conducting mixed methods research: using dialectical pluralism and social psychological strategies. In: Leavy, P. (Ed.), The Oxford Handbook of Qualitative Research. Oxford University Press, pp. 557–578.
Johnson, R.B., 2017. Dialectical pluralism: a metaparadigm whose time has come. J. Mix. Methods Res. *11* (2), 156–173. https://doi.org/10.1177/1558689815607692.
Karnati, C.S., 2021. Differential Item Functioning in a Teacher-Created Benchmark Mathematics Assessment. ProQuest Dissertation Publishing.
Koskey, K.L.K., Bright, D., Struloeff, K., Sondergeld, T., Stone, G., Bostic, J., Matney, G., 2021. Delphi technique in the development of assessments of problem-solving in computer adaptive testing environments (DEAP-CAT). Proc. Int. Conf. Educ. Res. Innovat. 9299–9306. https://doi.org/10.21125/iceri.2021.2142.
Koskey, K.L.K., Sondergeld, T.A., Stewart, V.C., Pugh, K.J., 2018. Applying the mixed methods Instrument Development and Construct Validation Process: The Transformative Experience Questionnaire. J. Mixed Method. Res. 12 (1), 95–122. https://doi.org/10.1177/1558689816633310.
Linacre, J.M., 2019. WINSTEPS® [Computer software]. Beaverton, Oregon, Version 4.4.5. https://www.winsteps.com/winsteps.htm.
Linacre, J.M., 2021. DIF - DPF - Bias - Interactions. Winsteps Online Manual. https://www.winsteps.com/winman/difconcepts.htm.
Linstone, H.A., Turoff, M., 1975. The Delphi method: techniques and applications. J. Market. Res. *18* (3) https://doi.org/10.2307/3150755.
Luyt, R., 2012. A framework for mixing methods in quantitative instrument development, validation, and revision: a case study. J. Mix. Methods Res. *6* (4), 294–316. https://doi.org/10.1177/1558689811427912.
Lyons, S., Johnson, M., Hinds, B.F., 2021. *A Call to Action: Confronting Inequity in Assessment.* Lyons Assessment Consulting. https://www.lyonsassessmentconsulting.com/assets/files/Lyons-JohnsonHinds_CalltoAction.pdf.
Maddox, B., Zumbo, B.D., Tay-Lim, B., Qu, D., 2015. An anthropologist among the psychometricians: assessment events, ethnography, and Differential Item Functioning in the Mongolian Gobi. Int. J. Test. *15* (4), 291–309. https://doi.org/10.1080/15305058.2015.1017103.
Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E.A., McFarland, J.L., Price, R.M., 2017. Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. Life Sci. *16* (2), 1–13.
Middleton, J., Gorard, S., Taylor, C., Bannan-Ritland, B., 2008. The "compleat" design experiment. In: Kelly, A., Lesh, R., Baek, J. (Eds.), Handbook Of Design Research Methods in Education: Innovations In Science, Technology, Engineering, and Mathematics Teaching and Learning. Routledge, pp. 21–46.
Morgan, D.L., 1998. Practical strategies for combining qualitative and quantitative methods: applications to health research. Qual. Health Res. *8*, 362–376. https://doi.org/10.1177/104973239800800307.
Morse, J.M., 1991. Approaches to qualitative-quantitative methodological triangulation. Nurs. Res. 40, 120–123.
Morse, J.M., 2003. Principles of mixed methods and multimethod research design. In: Tashakkori, A., Teddlie (Eds.), Handbook of Mixed Methods in Social & Behavioral Research. Sage, pp. 189–208.
Onwuegbuzie, A.J., Bustamante, R.M., Nelson, J.A., 2010. Mixed research as a tool for developing quantitative instruments. J. Mix. Methods Res. 4 (1), 56–78. https://doi.org/10.1177/1558689809355805.
Osterlind, S.J., Everson, H.T., 2009. Differential Item Functioning – Series: Quantitative Applications in the Social Sciences, 2nd ed. Sage.
Paré, G., Cameron, A.-F., Poba-Nzaou, P., Templier, M., 2013. A systematic assessment of rigor in information systems ranking-type Delphi studies. Inf. Manag. 50, 207–217. https://doi.org/10.1016/j.im.2013.03.003.
Plano Clark, V.L., 2019. Meaningful integration within mixed methods studies: identifying why, what, when and how. Contemp. Educ. Psychol. 57, 106–111 doi: 10.016/j.cedpsy.2019.01.007.
Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks Paedagogiske Institut, Copenhagen.
Rasch, G., 1980. Probabilistic Models for Some Intelligence and Attainment Tests, Expanded ed. University of Chicago Press.
Richardson, L., 2000. Writing: A method of inquiry. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of qualitative research, 2nd ed. Sage, pp. 923–943.
Saldaña, J., 2016. The Coding Manual for Qualitative Researchers, 3rd ed. Sage.
Sekayi, D., Kennedy, A., 2017. Qualitative Delphi method: a four round process with a worked example. Qual. Rep. 22 (10), 2755–2763. https://doi.org/10.46743/2160-3715/2017.2974.
Stefurak, T., Johnson, R.B., Shatto, E., 2016. Mixed methods and dialectical pluralism. In: Jason, L.A., Glenwick, D.S. (Eds.), Handbook Of Methodological Approaches to Community-Based Research: Qualitative, Quantitative, and Mixed Methods. Oxford University Press, pp. 345–354.
Teddlie, C., Tashakkori, A., 2009. Foundations Of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences. Sage.
van de Vijver, F.J.R., Tanzer, N.K., 2014. Bias and equivalence in cross-culture assessment: an overview. Eur. Rev. Appl. Psychol. 54, 119–135. https://doi.org/10.1016/j.erap.2003.12.004.
Wessels, M.D., Paap, M.C.S., van der Putten, A.A.J., 2022. The content validity of the Behavioral Appraisal Scales in people with profound intellectual and multiple disabilities: a Delphi study. J. Pol. Pract. Intellect. Disabil. 19, 86–101. https://doi.org/10.1111/jppi.12409.
Yildirim, H., Büyüköztürk, S., 2018. Using the Delphi Technique and focus-group interviews to determine item bias on the mathematics section of the level determination exam for 201. Educ. Sci. Theor. Pract. 18 (2), 447–470. https://doi.org/10.12738/estp.2018.2.0317.
Yousuf, M.I., 2007. Using experts' opinions through Delphi technique. Practical Assess. Res. Eval. 12 (4) https://doi.org/10.7275/rrph-t210.
Zunder, T.H., Islam, D.M.Z., 2011. E-logistics systems applications for service users and providers. Transport. Res. Rec. 2238 (1), 50–60. https://doi.org/10.3141/2238-07.