frontiers | Frontiers in Education

TYPE Brief Research Report
PUBLISHED 06 June 2025
DOI 10.3389/feduc.2025.1595648

OPEN ACCESS

EDITED BY
Yousef Wardat,
Yarmouk University, Jordan

REVIEWED BY
Serkan Boyraz,
Aksaray University, Türkiye
Jericho Yu Baybayan,
Notre Dame University, Philippines

*CORRESPONDENCE
Michalis P. Michaelides
✉ michaelides.michalis@ucy.ac.cy
Evi Konstantinidou
✉ econst02@ucy.ac.cy

RECEIVED 18 March 2025
ACCEPTED 12 May 2025
PUBLISHED 06 June 2025

CITATION
Konstantinidou E and Michaelides MP (2025) Assessment mode and inconsistent responding on a mixed-worded scale: evidence from TIMSS 2019 across grades and countries. *Front. Educ.* 10:1595648. doi: 10.3389/feduc.2025.1595648

Assessment mode and inconsistent responding on a mixed-worded scale: evidence from TIMSS 2019 across grades and countries

Evi Konstantinidou[1,2]* and Michalis P. Michaelides[1]*

[1]Department of Psychology, University of Cyprus, Nicosia, Cyprus, [2]Department of Education, University of Nicosia, Nicosia, Cyprus

Understanding factors that affect inconsistent responding in mixed-worded scales is crucial for ensuring the validity of survey score outcomes. This study investigated whether the assessment mode, defined as participating either in a digital or a paper-based achievement test, was associated with the prevalence of inconsistent responding on a mixed-worded scale on a questionnaire administered immediately after. Data were used from 4th- and 8th-grade students from 16 countries participating in the 2019 Trends in International Mathematics and Science Study (TIMSS). The self-reported mixed-worded scale measured self-concept in mathematics and was administered in paper format to all students. The study employed the mean absolute difference (MAD) and factor mixture analysis (FMA) methods to identify inconsistent respondents. Although cross-cultural variation was observed, the 4th-graders who had taken the computer-based assessment prior to the questionnaire engaged in a slightly higher frequency of inconsistent responses. Among 8th-graders, the difference was smaller and, in several country samples, reversed. Larger prevalence of inconsistent responding was found with the FMA approach. The study emphasizes the need for further research on inconsistent responding across different assessment modes and contexts, suggesting implications for survey methodology and international studies.

KEYWORDS

assessment mode, inconsistent responding, mixed-worded scale, TIMSS, cross-cultural comparison

1 Introduction

Mixed-worded scales comprise items which are positively or negatively phrased, measuring one construct (or multiple constructs in the case of multidimensional scales). This reversal in wording is intended to prompt participants to respond more thoughtfully, thereby improving the psychometric properties of instruments for constructs that are not directly observable (Podsakoff et al., 2003). However, students may occasionally treat positively and negatively worded items as if they had the same valence and consequently respond inconsistently across items (Steinmann et al., 2022b), e.g., by simultaneously agreeing with both positively and negatively worded items despite their opposite meaning. This inconsistency harms psychometric properties of a scale, like score reliability, and factor structure (Schmitt and Stuits, 1985; Steedle et al., 2019; Woods, 2006).

Two approaches proposed in the literature to identify inconsistent respondents based on their responses are: the mean absolute difference (MAD; Steedle et al., 2019), which gauges the agreement between responses to positively and negatively worded items for each respondent and the constrained factor mixture analysis (FMA; Steinmann et al., 2022b), which originates from factor analytic methods (Lubke and Muthén, 2005). Empirical studies have found that the prevalence of the inconsistent responding varies across countries (Steinmann et al., 2022a) and is higher in younger than older students (Steinmann et al., 2022b, 2024). This inconsistency may stem from a lack of reading and/or cognitive skills required to process mixed wording (Baumgartner et al., 2018; Steinmann et al., 2022b; Swain et al., 2008). Alternatively, the carelessness explanation suggests that distracted or disengaged respondents read scales superficially, failing to notice and respond accurately to reverse wording (Schmitt and Stuits, 1985; Steedle et al., 2019; Steinmann et al., 2022b). Another possibility is acquiescence response bias, which refers to the tendency of respondents to agree with items regardless of their content (Bentler et al., 1971).

Advancements in digital technologies have facilitated the gradual transition from paper-based to computer-based administration modes for national and international large-scale assessment programs. The primary concern during this transition has been to ensure the comparability of scores across modes (Buerger et al., 2019). A mode effect refers to any difference in performance that can be attributed to the administration mode. Changing the administration mode for a test can introduce sources of construct-irrelevant variance, potentially compromising fairness and the validity of score interpretations (Lynch, 2022). Even between different types of online platforms, the quality of data provided by respondents could vary significantly (Douglas et al., 2023).

Magraw-Mickelson et al. (2022) found minimal differences in careless responding between digital and paper/pencil survey modes. Johnson (2005) compared personality inventory data collected via a web-based platform to independently gathered paper-and-pencil data; online data were of lower quality, characterized by higher long string indices, more missing, but not more inconsistent responses. In achievement contexts, empirical evidence has shown that test scores in computerized tests tend to be significantly lower than in paper-based ones (Jerrim et al., 2018; Wagner et al., 2022). A topic that has not been addressed is whether mode-induced performance variations in large-scale testing programs impact subsequent self-reporting in questionnaires.

The aim of this study was to investigate whether participation in a computer-based vs. a paper-based assessment was associated with differential inconsistent responding on a subsequently administered mixed-worded scale on a paper-based questionnaire. Secondary data were obtained from the 2019 Trends in International Mathematics and Science Study (TIMSS) and the "self-concept in mathematics" self-report scale. Sixteen countries participated in the electronic TIMSS (eTIMSS) assessment and administered a paper version of TIMSS to an additional student sample. The MAD method was applied to detect inconsistent responding in samples from 16 countries. A second research question addressed variations in inconsistent responding in 4th- and 8th-grade student samples. Finally, FMA, a different detection method was implemented for cross-validation purposes.

Understanding and documenting such response tendencies can inform stakeholders, such as assessment developers, researchers, and policymakers, who are concerned with the validity of data obtained from large-scale, low-stakes studies with self-report instruments. Investigations on the prevalence of inconsistent responding are relevant to the quality of the data collected and may ultimately support the development of more robust measurement procedures. Moreover, assessment programs like TIMSS, which are administered internationally, aim to provide comparable data across countries, modes and age groups. Comparisons of response tendencies across various groups provide evidence about the validity of inferences on group differences. Finally, the examination of outcomes from two detection methods is particularly relevant for psychometric researchers and assessment developers working with large-scale educational surveys, as it provides information on how to identify and interpret inconsistencies in self-report data.

2 Method

2.1 Experimental design and sampling

Sixteen countries participating in the computerized TIMSS 2019 assessment also administered a paper-based version to different samples of 4th- and 8th-grade students, as a part of the Bridge study (Fishbein et al., 2021). The eTIMSS study refers to the digital format of the TIMSS assessment, an innovative development aiming to modernize TIMSS by incorporating elements reflecting digital educational environments and capturing richer information about students' problem-solving processes (Cotter et al., 2020). To ensure longitudinal score comparability with the paper-based format implemented in prior cycles of the program, TIMSS 2019 conducted the Bridge Study designed to assess and mitigate any mode effects that might result from transitioning from paper to digital administration (von Davier et al., 2020).

For eTIMSS a two-stage random sampling process of selecting schools and intact classes was used (see country sample sizes on Table 1). An additional 1,500 students per cohort received the Bridge paper-based booklets; they were selected from about a third of the schools from the full eTIMSS sample, hence they can be considered randomly equivalent (von Davier et al., 2020). Taking advantage of this allocation, this secondary data analysis study can be considered a posttest-only experimental design with two groups. Inspection of the database revealed that the students in the Bridge sample received a smaller number of items in the paper-based achievement tests than their eTIMSS counterparts prior to the administration of the student questionnaire. Irrespective of the test version they received, all students responded to a paper-based questionnaire (Mullis and Fishbein, 2020).

2.2 Measures

The mathematics self-concept scale was administered in the paper-based student questionnaire in both grades following the achievement tests. The scale consisted of four positively and five

TABLE 1 Prevalence of inconsistent respondents (%) on the paper-based self-concept scale per assessment mode and grade using MAD and FMA.

Country	eTIMSS sample sizes		Mean absolute difference (MAD) method				Factor mixture analysis (FMA) method			
			4th-grade		8th-grade		4th-grade		8th-grade	
	4th-grade	8th-grade	Bridge	eTIMSS	Bridge	eTIMSS	Bridge	eTIMSS	Bridge	eTIMSS
Chile	4,775	4,697	9.77	10.60	2.24	4.43	20.08	22.56	10.39	13.50
Chinese Taipei	4,295	5,610	5.99	7.00	2.56	3.09	19.16	23.45	7.89	12.40
England	3,872	3,858	2.97	2.78	1.66	3.24	6.93	7.05	6.94	9.88
Georgia	4,316	3,789	6.00	6.24	5.79	5.30	27.06	27.30	15.78	20.08
Hong Kong	3,386	3,730	7.78	8.32	3.46	5.08	22.73	23.77	15.68	23.23
Hungary	5,227	5,219	2.97	3.65	1.46	3.00	11.43	14.38	12.87	12.71
Italy	4,269	4,138	2.22	3.13	1.73	1.64	20.53	13.19	4.52	8.44
Republic of Korea	4,448	4,409	2.23	1.52	2.70	2.33	7.07	8.22	4.78	8.19
Lithuania	4,265	4,366	1.33	0.86	0.55	0.78	9.94	11.61	3.79	10.89
Norway	4,527	5,215	2.30	2.01	2.70	2.08	6.07	9.35	7.21	6.69
Qatar	5,646	4,436	11.73	14.02	7.61	11.26	27.91	31.90	20.35	28.83
Russian Federation	4,596	4,456	2.48	3.45	1.67	1.61	10.17	11.52	9.10	6.50
Singapore	6,839	5,546	3.24	3.67	1.46	2.22	10.40	12.12	8.74	7.08
Sweden	4,535	4,565	1.19	1.72	2.28	1.91	15.75	5.92	7.26	8.16
United Arab Emirates	29,515	25,539	8.12	10.14	8.33	9.24	29.60	41.36	18.20	24.71
United States	10,029	9,944	5.51	6.53	4.25	3.65	11.27	13.91	9.75	12.22
Total sample	104,540	99,517	4.38	5.12	3.13	3.05	10.19	12.45	10.57	11.36

Results weighted by total weights (TOTWGT).

negatively worded items, such as "*I usually do well in mathematics*" and "*Mathematics is harder for me than any other subject.*" Responses were given on a 4-point scale, with options 1 = "Agree a lot," 2 = "Agree a little," 3 = "Disagree a little," and 4 = "Disagree a lot." The reliability of the scale scores across the 16 countries was consistently high in the eTIMSS samples, with Cronbach's alphas ranging from 0.82 to 0.93 for 8th-grade and from 0.79 to 0.88 for 4th-grade (Yin and Fishbein, 2020).

2.3 Statistical analysis

Databases for each country and grade and the eTIMSS and Bridge samples were downloaded from the TIMSS 2019 International Database and prepared with the IDB Analyser. To identify inconsistent respondents, the MAD method was applied by reverse-coding the mean score of the positively worded items and subtracting it from the mean score of the negatively worded items; if the absolute difference was ≥1.75 the student response was classified as inconsistent (Steinmann et al., 2022a). The FMA method was implemented in Mplus 8.5 (Muthén and Muthén, 1998–2017) to classify respondents in one of two latent classes by constraining the item factor loadings of positively and negatively worded items (Steinmann et al., 2022b). For the consistent class the loadings were of opposite sign for the two wording types, while for the inconsistent one they were specified to be of the same sign suggesting similar responses irrespective of item directionality. Total student weights (TOTWGT) were applied in the analysis. Syntax files for the MAD and FMA analyses are available on the Open Science Framework.

3 Results

With the MAD approach, the overall proportion of inconsistent respondents was higher among 4th-grade students who completed the questionnaire after the computer-based assessment (5.12%), compared to those who completed it after the paper-based assessment (4.38%). These findings are summarized in Table 1. For 8th-grade students, the overall proportion was very similar among students in the eTIMSS context (3.05%) and the paper-based Bridge (3.13%). In most of the 16 countries (12 for grade four, 9 for grade eight), students who participated in the digital assessment exhibited higher percentages of inconsistent responding on the subsequently administered scale compared to their peers in the paper-based context. Figure 1 illustrates the proportion of inconsistent respondents across countries, grade levels, and assessment modes with the MAD approach in decreasing percentage in the eTIMSS samples. In both grades, eTIMSS samples in orange colors are in most cases higher than Bridge samples in green colors.

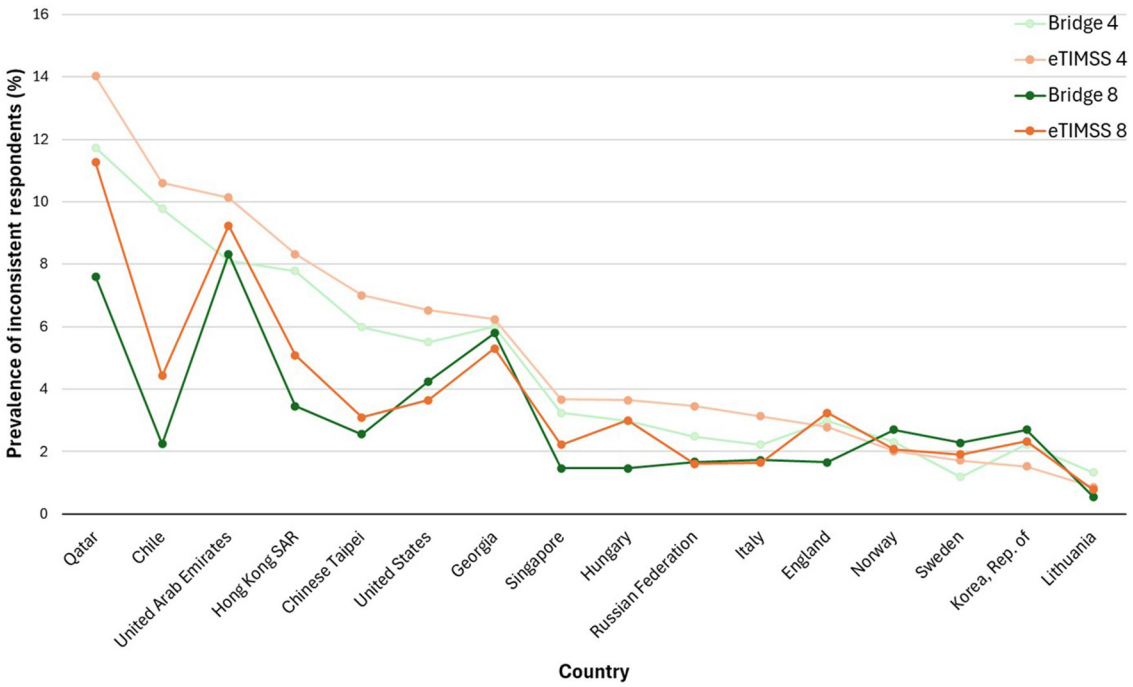

FIGURE 1
Prevalence of inconsistent respondents by country, grade (4, 8), and condition (eTIMSS, Bridge) using mean absolute difference.

For the second research question, the MAD approach revealed that, across countries, the percentage of inconsistent respondents was higher in 4th- than in 8th-grade, regardless of the assessment mode that preceded the questionnaire (lighter colors for younger students are higher compared to solid colors for older students on Figure 1). This pattern was noticeable in nearly all countries, but with cross-national variation. In Korea, Norway, and Sweden the difference was reversed and very small. England and the United Arab Emirates had mixed results in age comparisons across the two conditions.

Regardless of grade, the FMA identification approach revealed higher proportions of inconsistent respondents in the eTIMSS samples, as can be seen on Table 1 and Figure 2. Fewer countries were found to have lower proportions of inconsistent respondents after the computer-based assessment: two in grade four (Italy and Sweden) and four in grade eight (Hungary, Norway, Russia, and Singapore). FMA findings generally aligned with MAD trends across grades, with higher prevalence in the younger samples. Overall, the FMA classified more than twice as many students as inconsistent compared to the MAD approach in 4th-grade, with even larger discrepancies observed in 8th grade. In some countries, the percentage of inconsistent respondents identified by the FMA was substantial, exceeding 20%.

4 Discussion

Responses to reverse-worded items of a scale that are inconsistent introduce noise in the data, lower reliability and misrepresent the dimensionality of scales (Schmitt and Stuits, 1985; Steedle et al., 2019). This experimental study examined whether participation in a digital vs. paper-based TIMSS achievement test affected the percentage of inconsistent respondents on the mixed-worded self-concept in mathematics scale, which was subsequently administered on paper. The design was a posttest-only experiment with randomly equivalent groups and large student samples, supporting the internal validity of the comparison. The Bridge study was a large-scale investigation of effects of the transition to eTIMSS on achievement results (von Davier et al., 2020). Our focus shifted to the TIMSS paper-based student questionnaire and compared the prevalence of inconsistent responding on a self-reported contextual scale.

The problem of inconsistent responding in programs that are low-stakes, due to no personal consequences on respondents, can be influenced by factors such as the context in which the assessment is administered. The trend emerging from multiple samples analyzed in the current study was that inconsistent response rates were more pronounced among students who had completed the digital assessment. This pattern was evident among 4th-graders in most country samples but was less clear among 8th-graders. The difference in inconsistent responding between the two experimental conditions was bigger in 4th- than in 8th-grade. The prevalence of inconsistent responding was higher in grade four than in grade eight, in agreement with previous studies which found more inconsistent respondents in younger student samples (Steinmann et al., 2022b, 2024). Higher proficiency in verbal and cognitive skills, as well as more familiarity with computerized assessment conditions may explain this age difference.

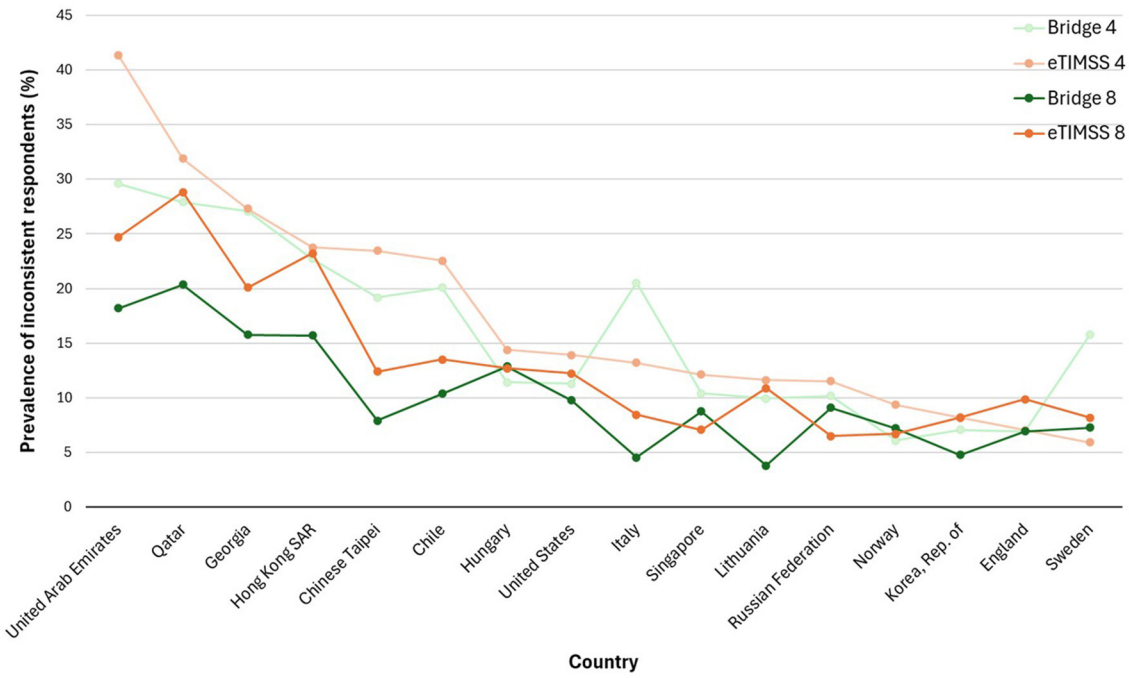

FIGURE 2
Prevalence of inconsistent respondents by country, grade (4, 8), and condition (eTIMSS, Bridge) using factor mixture analysis.

Meade and Craig (2012) describe various strategies for detecting careless responding, including (a) inserting special items into the survey design and (b) applying *post hoc* techniques such as consistency indices or response time analysis. In this study, secondary data were analyzed with no existing design elements or indices that would have allowed for identifying inconsistent respondents. Hence, two *post hoc* techniques were implemented due to their complementary features: MAD provides a simple, computationally efficient measure of inconsistency, albeit with an arbitrary threshold (Ulitzsch et al., 2022), while the model-based FMA leverages the flexibility of mixture models to account for latent heterogeneity in response patterns and generate classes of respondents. Despite larger prevalences yielded by the FMA approach, the general conclusions about the experimental conditions and age differences remained similar. The larger prevalence rates found with the FMA approach underscore the importance of carefully selecting and justifying analytical methods when investigating response quality in self-report data.

5 Conclusion and recommendations

This study investigated inconsistent responding to a mixed-worded self-concept scale following participation in either digital or paper-based achievement assessments, using TIMSS 2019 data from 16 countries. Results showed higher inconsistency after the digital assessment, particularly among younger students. The comparison of two analytical methods, MAD and FMA, revealed similar trends, though FMA identified more inconsistent respondents. These findings highlight the need for caution when interpreting self-report data across different administration contexts and age groups.

The contribution of this empirical study concerns evidence about inconsistent responding on a scale across two assessment contexts and two age groups from an international program. The unique opportunity afforded by the Bridge study with two large and randomly equivalent conditions in each of the 16 countries allowed for an experimental comparison of response behavior following participation in either a digital or paper-based achievement test. A threat to the validity of the design might be that the achievement tests preceding the questionnaire were not identical in length, and this may have had an impact on subsequent response behavior. Considerable heterogeneity was observed in many representative country samples. The slightly higher inconsistent responding among students exposed to a computerized test and among younger participants have implications on the quality of data from low-stakes assessments, as well as from similar contexts such as online surveys, teaching evaluations, and research questionnaires.

It is recommended that assessment designers and researchers remain aware of the potential impact of assessment task sequencing and context on respondent attentiveness, especially in low-stakes environments. The cognitive demands of mixed-worded items should be carefully evaluated, particularly for younger students who may be more susceptible to contextual influences. Researchers should

employ multiple detection methods to assess response quality, and policymakers should account for potential context- and age-related biases when interpreting questionnaire data in large-scale assessments.

While this study provides empirical evidence on inconsistent responding across different assessment administration conditions, countries, and grade levels, further research is needed to uncover the mechanisms driving such patterns. Future studies could investigate whether exposure to computer- vs. paper-based tasks influences engagement and motivation in follow-up tasks; and whether individual differences such as digital or cognitive skills and age moderate these effects. An alternative experimental design would involve the manipulation of the survey mode, where participants respond to computer- or paper-based questionnaires to examine variations in response inconsistency. Applying similar analyses in other large-scale assessments could also assess the generalizability of these findings. Finally, the divergence of the proportions of inconsistencies between the two methods highlights the need for further methodological research comparing detection methods, including their assumptions and thresholds.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: TIMSS data website https://timss2019.org/international-database/. Files with analysis code can be found at https://osf.io/r93jg/?view_only=e0efd99ee60040d489ff36a74c131c43.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

EK: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. MM: Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Funding

The authors declare that financial support was received for the research and/or publication of this article. The authors would like to acknowledge funding support to Michalis Michaelides by the University of Cyprus (Internal Funding Program, project title: Prevalence of inconsistent responding in mixed-worded scales and characteristics of inconsistent respondents in surveys, IR-MIX).

Conflict of interest

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

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Minor language editing was made using OpenAI's ChatGPT. The author(s) take full responsibility for the content of the manuscript.

Publisher's note

References

Baumgartner, H., Weijters, B., and Pieters, R. (2018). Misresponse to survey questions: a conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *J. Mark. Res.* 55, 869–883. doi: 10.1177/0022243718811848

Bentler, P. M., Jackson, D. N., and Messick, S. (1971). Identification of content and style: a two-dimensional interpretation of acquiescence. *Psychol. Bull.* 76, 186–204. doi: 10.1037/h0031474

Buerger, S., Kroehne, U., Koehler, C., and Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Stud. Educ. Eval.* 62, 1–9. doi: 10.1016/j.stueduc.2019.04.005

Cotter, K. E., Centurino, V. A. S., and Mullis, I. V. S. (2020). "Developing the TIMSS 2019 mathematics and science achievement instruments," in *Methods and Procedures: TIMSS 2019 Technical Report*, eds. M. O. Martin, M. von Davier, and I. V. S. Mullis (Boston, MA: TIMSS and PIRLS International Study Center, Boston College), 1.1–1.12. Available online at: https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch1-developing-achievement-instruments.pdf (accessed April 22, 2025).

Douglas, B. D., Ewell, P. J., and Brauer, M. (2023). Data quality in online human-subjects research: comparisons between MTurk, Prolific, CloudResearch, qualtrics, and SONA. *PLoS ONE* 18:e0279720. doi: 10.1371/journal.pone.0279720

Fishbein, B., Foy, P., and Yin, L. (2021). *TIMSS 2019 User Guide for the International Database*. Boston, MA: Boston College, TIMSS and PIRLS International Study Center. Available online at: https://timssandpirls.bc.edu/timss2019/international-database/ (accessed April 22, 2025).

Jerrim, J., Micklewright, J., Heine, J.-H., and Salzer, C. andMcKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxf. Rev. Educ.* 44, 476–493. doi: 10.1080/03054985.2018.1430025

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Pers.* 39, 103–129. doi: 10.1016/j.jrp.2004.09.009

Lubke, G. H., and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychol. Methods* 10, 21–39. doi: 10.1037/1082-989X.10.1.21

Lynch, S. (2022). Adapting paper-based tests for computer administration: lessons learned from 30 years of mode effects studies in education. *Pract. Assess. Res. Eval.* 27:22. doi: 10.7275/pare.1317

Magraw-Mickelson, Z., Wang, H. H., and Gollwitzer, M. (2022). Survey mode and data quality: careless responding across three modes in cross-cultural contexts. *Int. J. Test.* 22, 121–153. doi: 10.1080/15305058.2021.2019747

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Mullis, I. V. S., and Fishbein, B. (2020). "Updating the TIMSS 2019 instruments for describing the contexts for student learning," in *Methods and Procedures: TIMSS 2019 Technical Report*, eds. M. O. Martin, M. von Davier, and I. V. S. Mullis (Boston, MA: TIMSS and PIRLS International Study Center, Boston College), 2.1–2.9. Available online at: https://timssandpirls.bc.edu/timss2019/methods/chapter-2.html (accessed April 22, 2025).

Muthén, L. K., and Muthén, B. O. (1998–2017). *Mplus User's Guide*, 8th Edn. Los Angeles, CA: Muthén and Muthén.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Schmitt, N., and Stuits, D. M. (1985). Factors defined by negatively keyed items: the result of careless respondents? *Appl. Psychol. Meas.* 9, 367–373. doi: 10.1177/014662168500900405

Steedle, J. T., Hong, M., and Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educ. Meas. Issues Pract.* 38, 101–111. doi: 10.1111/emip.12256

Steinmann, I., Chen, J., and Braeken, J. (2024). Who responds inconsistently to mixed-worded scales? Differences by achievement, age group, and gender. *Assess. Educ. Princ. Policy Pract.* 31, 5–31. doi: 10.1080/0969594X.2024.2318554

Steinmann, I., Sánchez, D., van Laar, S., and Braeken, J. (2022a). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. *Assess. Educ. Princ. Policy Pract.* 29, 5–26. doi: 10.1080/0969594X.2021.2005302

Steinmann, I., Strietholt, R., and Braeken, J. (2022b). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychol. Methods* 27, 667–702. doi: 10.1037/met0000392

Swain, S. D., Weathers, D., and Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed likert items. *J. Mark. Res.* 45, 116–131. doi: 10.1509/jmkr.45.1.116

Ulitzsch, E. Yildirim-Erbasli, S. N., Gorgun, G., Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *Br. J. Math. Stat. Psychol.* 75, 668–698. doi: 10.1111/bmsp.12272

von Davier, M., Foy, P., Martin, M. O., and Mullis, I. V. (2020). "Examining eTIMSS country differences between eTIMSS data and bridge data," in *Methods and Procedures: TIMSS 2019 Technical Report*, eds. M. O. Martin, M. von Davier, and I. V. S. Mullis (Boston, MA: TIMSS and PIRLS International Study Center, Boston College), 13.1–13.24.

Wagner, I., Loesche, P., and Bißantz, S. (2022). Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing. *Eur. J. Psychol. Educ.* 37, 531–549. doi: 10.1007/s10212-021-00532-6

Woods, C. M. (2006). Careless responding to reverse-worded items: implications for confirmatory factor analysis. *J. Psychopathol. Behav. Assess.* 28, 186–191. doi: 10.1007/s10862-005-9004-7

Yin, L., and Fishbein, B. (2020). "Creating and interpreting the TIMSS 2019 context questionnaire scales," in *Methods and Procedures: TIMSS 2019 Technical Report*, eds. M. O. Martin, M. von Davier, and I. V. S. Mullis (Boston, MA: TIMSS and PIRLS International Study Center, Boston College), 16.1–16.331. Available online at: https://timssandpirls.bc.edu/timss2019/methods/chapter-16.html (accessed April 22, 2025).