

Who responds inconsistently to mixed-worded scales? Differences by achievement, age group, and gender

Isa Steinmann, Jianan Chen & Johan Braeken

To cite this article: Isa Steinmann, Jianan Chen & Johan Braeken (15 Feb 2024): Who responds inconsistently to mixed-worded scales? Differences by achievement, age group, and gender, Assessment in Education: Principles, Policy & Practice, DOI: [10.1080/0969594X.2024.2318554](https://doi.org/10.1080/0969594X.2024.2318554)

To link to this article: <https://doi.org/10.1080/0969594X.2024.2318554>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 15 Feb 2024.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

Who responds inconsistently to mixed-worded scales? Differences by achievement, age group, and gender

Isa Steinmann ^a, Jianan Chen ^a and Johan Braeken ^b

^aDepartment of Primary and Secondary Teacher Education, Oslo Metropolitan University, Oslo, Norway;

^bCentre for Educational Measurement, University of Oslo, Oslo, Norway

ABSTRACT

We investigated two research questions: which students are more likely to respond inconsistently to mixed-worded questionnaire scales, and which country samples have larger shares of inconsistent respondents? We defined an inconsistent response as strongly agreeing or disagreeing with both positively and negatively worded items of the same scale. Since we assumed that inconsistent responding occurs due to a lack of carefulness, reading, or cognitive skills, we expected to find that inconsistent responding was associated with lower achievement, younger age, being a nonnative speaker, and being a boy. We used data from all 38 countries that participated in the fourth- and eighth-grade assessments of TIMSS (Trends in International Mathematics and Science Study) 2019. Using the mean absolute difference method, we identified shares of 1–21% inconsistent respondents across samples. The results generally supported our hypotheses, especially the hypothesis that inconsistent responding is more common among students and countries with lower mathematics achievement levels.

ARTICLE HISTORY

Received 6 October 2023

Accepted 6 February 2024

KEYWORDS

Mixed-worded scales;
questionnaire design;
inconsistent respondents;
careless respondents; TIMSS

Mixed-worded questionnaire scales combine positively and negatively worded items to measure the same underlying construct. In the Trends in International Mathematics and Science Study (TIMSS), the student questionnaires contain a mathematics self-concept scale that combines positively worded items like ‘I usually do well in mathematics’ and negatively worded items like ‘I am just not good at mathematics’ with the same response categories from 1 (*agree a lot*) to 4 (*disagree a lot*) (Martin et al., 2020). Questionnaire developers use this mixed item wording to improve the quality of responses by preventing respondents from reading and answering the items in a rapid, superficial way (Likert, 1974; Nunnally & Bernstein, 1994; Podsakoff et al., 2003).

Previous research suggests, however, that using mixed-worded scales might have unintended consequences. A common finding is that data from mixed-worded scales do not fit the intended latent structure and that the scales have a lower reliability than expected (e.g. DiStefano & Motl, 2006; Gnamb & Schroeders, 2020; Kam & Meyer, 2015; Marsh, 1996). The mathematics self-concept scale in TIMSS, for example, was designed

CONTACT Isa Steinmann  isa.steinmann@oslomet.no  Department of Primary and Secondary Teacher Education, Oslo Metropolitan University, Pilestredet 42, Oslo 0167, Norway

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

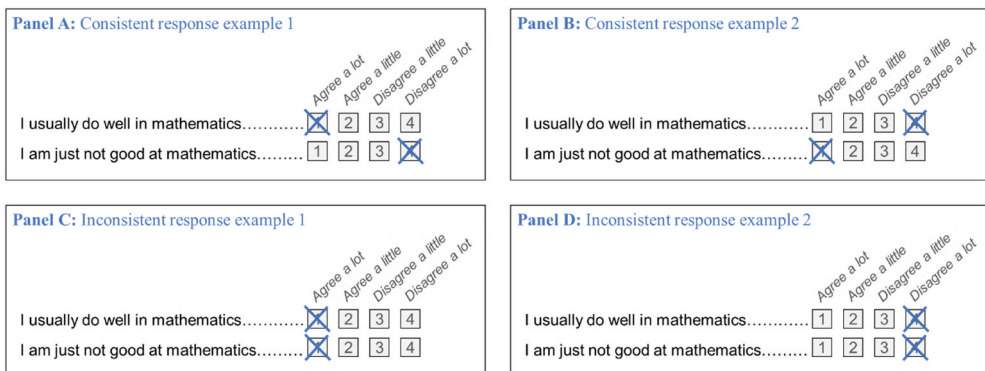
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

to be one-dimensional, but in confirmatory factor analyses, the empirical data appeared to be more complex. The use of a second factor capturing additional covariance among the negatively worded items improved the model fit in previous studies (Marsh et al., 2013; Michaelides, 2019).

Consistent and inconsistent respondents

There are different explanations as to why mixed item wording has unintended consequences for data quality. In this study, we assume that some respondents do not answer the positively and negatively worded items consistently (cf. Steinmann, Strietholt, et al., 2022; Swain et al., 2008). This is illustrated in Figure 1. A consistent response to the mixed item wording requires the respondent to agree with one item type and disagree with the other (see Panels A and B in Figure 1). This implies that some respondents have to use double negations (e.g. disagreeing with not being good at mathematics in Panel A in Figure 1). Respondents need to have the information the item asks about, read and understand the item stem and response options, retrieve the requested information from memory, integrate the required and retrieved information, and then choose the best response option (e.g. Tourangeau et al., 2000). We argue that the mix of positively and negatively worded items makes it harder to respond appropriately because readers may not notice the change in item wording or may have difficulties handling the negative form (Baumgartner et al., 2018).

Therefore, we expect some respondents to fail to respond consistently because of a lack of attention or carefulness, or a lack of the reading or cognitive abilities required to notice and handle the changing item wording (Baumgartner et al., 2018; Steinmann, Strietholt, et al., 2022; Swain et al., 2008). Instead, respondents erroneously answer the negatively worded items as if they were positively worded ones; they either agree with both the positively and negatively worded items or disagree with them. This leads to inconsistent responses across item types that suggest both a positive and negative mathematics self-concept (see Panels C and D in Figure 1). We deliberately use the term ‘inconsistent respondents’ because the term allows for different explanations of the



Panel A: Consistent response example 1

	1	2	3	4
I usually do well in mathematics.....	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am just not good at mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Panel B: Consistent response example 2

	1	2	3	4
I usually do well in mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am just not good at mathematics.....	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Panel C: Inconsistent response example 1

	1	2	3	4
I usually do well in mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am just not good at mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Panel D: Inconsistent response example 2

	1	2	3	4
I usually do well in mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am just not good at mathematics.....	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 1. Display of example consistent and inconsistent responses. *Note.* Items stem from the TIMSS 2019 student questionnaire (Martin et al., 2020). Figure adapted from Steinmann, Strietholt, et al. (2022).

phenomenon. Previous studies have also used other terms such as ‘careless’, ‘disengaged’, or ‘insufficient effort’ respondents.

Numerous previous studies have identified inconsistent respondents empirically. These studies have applied different identification methods. An early study found that 23% of respondents gave the same answer to pairs of antonym items in a US sample of adults (Melnick & Gable, 1990). Using a similar methodology, Swain et al. (2008) found that 20% of respondents were inconsistent in samples of undergraduate students in the US. Some studies applied the mean absolute difference method – which flags respondents as inconsistent if they answer too similarly to positively and negatively worded items of the same scale – and found that 7% (Hong et al., 2020) and 10% (Steedle et al., 2019) of respondents were inconsistent in samples of secondary school students in the US. Bulut and Bulut (2022) applied the mean absolute difference method to primary school student samples from six English-speaking countries and found that an average of 8% of respondents answered inconsistently. Another study applied the mean absolute difference method to an international sample and found that 2–36% of primary school children in 37 countries were inconsistent respondents (Steinmann, Sánchez, et al., 2022). Other studies have used factor mixture analysis models to separate consistent and inconsistent respondents (latent classes) to mixed-worded items of the same, supposedly one-dimensional scale (latent factor). Studies have found that 7–20% of respondents are inconsistent in samples of children and adolescents from the US, Australia, and Germany (Steinmann, Strietholt, et al., 2022); 4–10% of adult samples in the US (Arias et al., 2020) and 8% of young adults in the Dominican Republic are inconsistent respondents (García-Batista et al., 2021).

In summary, numerous previous studies have identified inconsistent respondents in empirical analyses, predominantly using data from the US. The shares of inconsistent respondents vary greatly between studies and between countries in the same studies. Overall, the range is 2–36%, implying very small shares of inconsistent respondents in some instances and considerable (though minority) shares in other instances.

Effects of inconsistent responding

Inconsistent responding to mixed-worded questionnaire scales is problematic because it is not a logical response according to the mixed wording design (e.g. Baumgartner et al., 2018; Steinmann, Strietholt, et al., 2022). The presence of inconsistent respondents in datasets also leads to an overestimation of the data dimensionality, as shown in simulation studies (Schmitt & Stults, 1985; Woods, 2006) and in empirical analyses that have compared data with and without inconsistent respondents (Arias et al., 2020; Bulut & Bulut, 2022; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). Removing inconsistent respondents from datasets has led to improved reliability in some studies (Arias et al., 2020; Steinmann, Sánchez, et al., 2022) but had mixed or neutral effects in others (Hong et al., 2020; Steedle et al., 2019). There are mixed or neutral findings from previous studies on the effects of removing inconsistent respondents on associations between the scores of the mixed-worded scales and external variables (Bulut & Bulut, 2022; Hong et al., 2020; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). Steinmann, Sánchez, et al. (2022) showed that inconsistent responding led to an attenuation-to-the-mean bias for the scale scores

of inconsistent respondents. They also found that correlations between the scale scores and external variables of interest were artificially lower for inconsistent respondents compared to consistent respondents.

These findings indicate that the phenomenon of inconsistent respondents could explain the common observation of overly complex latent structures and low reliabilities in mixed-worded scales (cf. Steinmann, Strietholt, et al., 2022). This study addresses another perspective and focuses on whether inconsistent responding is more common among specific groups of respondents and in certain countries. This could imply that, contrary to their purpose, mixed-worded questionnaires do not work equally well for all respondents or in all countries. If, for example, inconsistent responding occurs in individuals who lack the reading or cognitive skills required to notice and handle the changing item wording, mixed-worded scales could introduce measurement issues (such as scale scores attenuated to the mean, scale score variances restricted in range, and a reduction in correlations between scale scores and external variables) for low-achieving students and countries.

Correlates of inconsistent responding

A few studies have compared the characteristics of consistent and inconsistent respondents. In line with the theoretical assumption that mixed wording is difficult to handle, several studies have found that inconsistent respondents have considerably lower levels of education or poorer cognitive, reading, or other academic abilities (Chen et al., [in print](#); Melnick & Gable, 1990; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022; Steinmann, Strietholt, et al., 2022). Another study applied the same mixed-worded questionnaire and inconsistent respondent identification method in fifth- and ninth-graders, finding a lower share of inconsistent respondents among older students than younger students (Steinmann, Strietholt, et al., 2022). Two studies investigated associations between inconsistent responding and self-reported conscientiousness, a proxy for answering the questionnaire carefully. One study found (conforming to expectations) that inconsistent responding was more common among students with lower conscientiousness (Chen et al., [in print](#)), while the other found no significant association (Steinmann, Strietholt, et al., 2022). If inconsistent responding is associated with person characteristics, intra-individual stability could also be expected. Two studies found a small overlap between inconsistent responding to multiple mixed-worded scales in the same questionnaires (Steinmann, Sánchez, et al., 2022; Steinmann, Strietholt, et al., 2022), but a longitudinal study did not find inconsistent response stability over four years (Steinmann, Strietholt, et al., 2022).

This literature review focuses on empirical studies that have identified consistent and inconsistent respondents empirically. We did not include the numerous variable-centred studies that have pointed to a potential connection between the mixed wording issue and person characteristics (e.g. Bolt et al., 2020; Kam & Chan, 2018; Lindwall et al., 2012; Marsh, 1996; Marsh et al., 2013; Michaelides, 2019; Quilty et al., 2006). Bulut and Bulut (2022) showed, for instance, that one-dimensional factor analysis models fit better with data from mixed-worded scales for high-achieving respondents than for low-achieving ones. In our view, however, these variable-centred methods are not well suited to measuring the person correlates of inconsistent responding.

In summary, previous research provides some support for the theoretical assumption that inconsistent responding might be rooted in a failure to read and handle the mixed wording appropriately (possibly due to a lack of reading or cognitive abilities), and may be especially likely in young children. There is less evidence for a lack of attention or carelessness in filling out questionnaires as an explanation for inconsistent responding. A study investigating inconsistent responding in an experimental setting with eye tracking also lent support to the difficulty explanation over the inattention explanation. Baumgartner et al. (2018) found that inconsistent respondents looked at negatively worded items for longer than positively worded items, meaning that they probably noticed the change in wording but still failed to give a consistent response. However, both carelessness and a lack of skills are plausible explanations for the inconsistent response behaviour.

If we find support for the assumption that inconsistent responding is associated with respondent characteristics, this could also explain the finding of considerable variation in shares of inconsistent respondents between countries (Steinmann, Sánchez, et al., 2022). If low achievers are more likely to respond inconsistently, countries with lower mean achievement should also have larger shares of inconsistent respondents. However, to the best of our knowledge, no previous study has investigated the association between shares of inconsistent respondents and other country characteristics.

This study

This study addresses two research questions:

- (1) Which students are more likely to respond inconsistently to mixed-worded questionnaire scales?
- (2) Which country samples have larger shares of inconsistent respondents to mixed-worded questionnaire scales?

Regarding the first research question, we expect to find negative within-country associations between inconsistent responding and better achievement scores, higher student age, (almost) always speaking the test language at home, and being a girl. The hypothesis about achievement has some support in the literature (Chen et al., *in print*; Melnick & Gable, 1990; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022; Steinmann, Strietholt, et al., 2022), but the others are more exploratory. Older student age should be associated with greater maturity (e.g. Bedard & Dhuey, 2006; Steinmann & Olsen, 2022), which should improve a student's ability to handle mixed-worded items. One previous study supported this assumption and found larger shares of inconsistent respondents among younger students (Steinmann, Strietholt, et al., 2022). We expect to find an association between inconsistent responding and speaking another language at home, as reading in another language is more demanding than reading in a native language (e.g. Brevik et al., 2016; Koda, 2007). We assume we will find fewer inconsistent respondents among girls than boys because girls typically have better reading abilities than boys (e.g. Mullis et al., 2017).

Regarding the second research question, we expect to find lower shares of inconsistent respondents in countries with higher mean achievement, in grade 8 compared to grade 4,

and in countries with larger shares of students who (almost) always speak the test language at home. These hypotheses are based on the hypotheses for students; there is no previous literature. We did not conduct country-level analyses on gender, since the shares of girls should be roughly 50% in all participating countries.

Materials and methods

Sample

We used data from TIMSS 2019 (TIMSS & PIRLS International Study Center, 2019). TIMSS is an international large-scale assessment that is conducted every four years. We included the two assessment populations – populations A and B – pertaining to school Grades 4 and 8, respectively. Some participants deviated from the international target grade under certain conditions (see LaRoche et al. (2020) for more information). TIMSS participants were either whole countries or benchmarking participants (i.e. parts of countries). In the following, we use the term *country* to describe all participants for the sake of simplicity. In tables and figures, we use the ISO 3166 codes as used in the TIMSS datasets for country abbreviations.

In this study, we included all students from all 38 countries that took part in both the fourth- and eighth-grade assessments. A stratified two-stage cluster sampling design was applied in each country (LaRoche et al., 2020). In country-specific strata, schools were sampled in the first step and at least one intact classroom per school in the second step. Different schools were usually sampled in the fourth- and eighth-grade assessments. The target sample sizes encompassed at least 150 schools and 4000 students per country, if the population was large enough. The samples were representative for students at the respective grade levels. We excluded students with no responses to any of the nine mixed-worded items. The average exclusion rate across countries was 3% in Grade 4 and 2% in Grade 8. The exclusion rates were below 10% in all countries except in Grade 4 in Quebec, Canada (exclusion rate = 16%). The effective sample sizes ranged from 2,968 in Hong Kong in Grade 4 to 25,834 in the United Arab Emirates in Grade 4.

Instruments

Mixed-worded scale as a basis for identifying inconsistent respondents

We used the mixed-worded mathematics self-concept scale, administered in the student questionnaires of the assessments for both Grades 4 and 8 (see Table 1). This scale consisted of four positively and five negatively worded items. The wording was almost identical in the Grade 4 and Grade 8 questionnaires. The international versions of the questionnaires were translated into numerous test languages (see Ebbs et al. (2020) for more information). The shares of missing values ranged between 0% and 11% in Grade 4 and between 0% and 7% in Grade 8 at the item level across countries.

Student-level covariates

To investigate the association between inconsistent responding and student characteristics within countries, we included four sets of variables: mathematics achievement scores, student age, the frequency of speaking the test language at home, and gender.

Table 1. Mixed-worded mathematics self-concept scale administered in TIMSS 2019 Grades 4 and 8.

	Wording
1. I usually do well in mathematics	+
2. Mathematics is harder for me than for many of my classmates ¹	-
3. I am just not good at mathematics ²	-
4. I learn things quickly in mathematics	+
5. Mathematics makes me nervous	-
6. I am good at working out difficult mathematics problems	+
7. My teacher tells me I am good at mathematics	+
8. Mathematics is harder for me than any other subject	-
9. Mathematics makes me confused	-

Note. The response categories were 1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, and 4 = *disagree a lot*. The items were administered in both TIMSS 2019 Grades 4 and 8.

¹In Grade 8, the exact wording was 'Mathematics is more difficult for me than for many of my classmates'. ²In Grade 8, the exact wording was 'Mathematics is not one of my strengths'.

We used mathematics test scores as proxies for cognitive and reading abilities, as these are not available in TIMSS. Previous research has shown that scores in such academic achievement and cognitive ability tests correlate highly (e.g. Rohde & Thompson, 2007). In TIMSS, plausible value variables are available to reflect students' overall performance in complex mathematics and science tests, either administered in an electronic or paper-based format (see Foy et al. (2020) for more information). In the fourth-grade assessment, the mean mathematics achievement scores ranged between $M = 374$ ($SD = 96$) in South Africa and $M = 626$ ($SD = 76$) in Singapore. In the eighth-grade assessment, the mean mathematics achievement scores ranged between $M = 389$ ($SD = 66$) in Morocco and $M = 616$ ($SD = 86$) in Singapore. There was no missing data in the achievement plausible values.

Student age was derived from the test date and birth month and year, assessed in the student questionnaires. In the fourth-grade assessment, mean age ranged between $M = 9.65$ ($SD = 0.33$) in Italy and $M = 11.52$ ($SD = 0.89$) in South Africa. In the eighth-grade assessment, mean age ranged between $M = 13.67$ ($SD = 0.58$) in Abu Dhabi and $M = 15.53$ ($SD = 1.10$) in South Africa. The shares of missing values ranged between 0–2% across countries and grades.

In the student questionnaires, the frequency of speaking the test language at home was assessed with the categories 1 (*always*), 2 (*almost always*), 3 (*sometimes*), and 4 (*never*). We collapsed categories and recoded them to form a binary variable with 0 (*sometimes or never*) and 1 (*always or almost always*). Across countries and grades, the shares of students who (almost) always spoke the test language at home ranged between 28% in South Africa in Grade 8 and 100% in Korea in Grade 8. The shares of missing values ranged between 0–14% across countries and grades.

The student questionnaires also asked about gender (binary variable with 0 = *boy* and 1 = *girl*). Across countries, the shares of girls in the samples ranged from 46% in Hong Kong in Grade 4 to 54% in Kuwait in Grade 8. The shares of missing values were lower than 1% in every country.

Country-level covariates

To investigate the association between the shares of inconsistent respondents and country characteristics, we focused on the mean mathematics achievement, grade level (0 = *Grade 4*, 1 = *Grade 8*), and shares of students who (almost) always spoke the test language at home. As the mean mathematics achievement and the shares of students who (almost) always spoke the test language at home were aggregates of the imputed student-level covariates, these variables did not contain any missing values.

Analysis

Identification of inconsistent respondents

We applied the mean absolute difference method to identify inconsistent respondents to mixed-worded questionnaire scales (Hong et al., 2020; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). The mean absolute difference method quantifies the extent to which respondents displayed a response pattern that conformed to the inconsistent pattern of either agreeing or disagreeing with both positively and negatively worded items (Steinmann, Sánchez, et al., 2022). For each student, we computed the absolute difference between the average responses to the positively worded items and the average of the reverse-coded responses to the negatively worded items. An extreme, consistent respondent could, for instance, agree a lot with all positively worded items (i.e. mean response to positively worded items = 1) and disagree a lot with all negatively worded items (i.e. mean reverse-coded response to negatively worded items = 1) (see Panel A in Figure 1). In this case, the mean absolute difference would be 0 ($|1 - 1| = 0$). An extreme, inconsistent respondent could, by contrast, agree a lot with all positively worded items (i.e. mean response to positively worded items = 1) and agree a lot with all negatively worded items (i.e. mean reverse-coded response to negatively worded items = 4) (see Panel C in Figure 1). The mean absolute difference would thus equal 3 ($|1 - 4| = 3$). Thus, the mean absolute difference varies between 0 (*extreme, consistent response*) and 3 (*extreme, inconsistent response*) if the mixed-worded scale has a four-point Likert scale.

We categorised all respondents with a mean absolute difference of 1.75 or higher as inconsistent (0 = *consistent*, 1 = *inconsistent*). This corresponds to an average distance of almost two units on the 4-category Likert response scale, and is the same threshold used by Steinmann, Sánchez, et al. (2022). It means that respondents who agree a lot with three positively worded items and agree a little with one positively worded item (i.e. mean response to positively worded items = 1.25), for example, have to agree at least a little with all negatively worded items (i.e. mean reverse-coded response to negatively worded items ≥ 3) to be flagged as inconsistent respondents.

Analyses to address first research question on student-level correlates of inconsistent responding

To address the first research question on student-level correlates of inconsistent responding, we ran student-level logistic regression models. Specifically, we investigated the association between inconsistent responding (0 = *consistent response*, 1 = *inconsistent response*) and mathematics achievement scores, student age in years, the frequency of speaking the test language at home (0 = *sometimes or never* and 1 =

always or almost always), and gender (0 = boy, 1 = girl). In the first set of models, we ran simple logistic regression models. In a second set of models, we ran multiple logistic regression models, including all four student-level covariates at the same time. The results of the simple logistic regression models are displayed in [Appendix A](#). We replicated the analyses for all 38 countries and for both Grade 4 and 8 assessments. To summarise the abundant student-level regression results across countries, we fitted the meta-analytic random-effects model and computed the across-country average log odds ratio and its corresponding confidence and prediction interval (CI and PI).

Analyses to address second research question on country-level correlates of inconsistent responding

To address the second research question on country-level correlates of inconsistent responding, we ran country-level simple linear regression models and a country-level paired t-test. We regressed the shares of inconsistent respondents on the variables for mean mathematics achievement and shares of students who almost (always) spoke the test language at home. Using a paired t-test, we investigated whether the shares of inconsistent respondents differed between Grade 4 and 8 assessments. Note that these analyses were based on only 38 country observations.

Analysis details

We addressed missing data using multiple imputation by chained equations (van Buuren, 2011). The imputation models were constructed using all four predictor variables (mathematics achievement scores, student age, the frequency of speaking the test language at home, and gender) and all nine mixed-worded items. We used the proportional odds model to impute mixed-worded item responses, and logistic regression to impute the gender and language variables. We generated five imputed datasets to capture the uncertainty around imputed values. Next, we created resamples that contained one imputed dataset and one of the five achievement plausible values (cf., von Davier et al., 2009). We ran the subsequent analyses separately for each of the five resamples and combined the results using Rubin's (1987) rules. This applies to both the student-level regression analyses and the estimation of the country-level variables. We accounted for the stratified two-stage cluster sampling design and the nesting of students in schools by using sampling weights and clustered standard errors (cf., Huang, 2016; Meinck, 2020). In the country-level regressions, all countries had the same weight.

Transparency and openness

We have reported how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we have followed the journal article reporting standards (Kazak, 2018). All data are publicly available at the TIMSS & PIRLS International Study Center and can be accessed at <https://timss2019.org/international-database/>. The analysis code is available as online supplementary material. Data were analysed using R, version 4.3.0 (R Core Team, 2022), and the packages mice, version

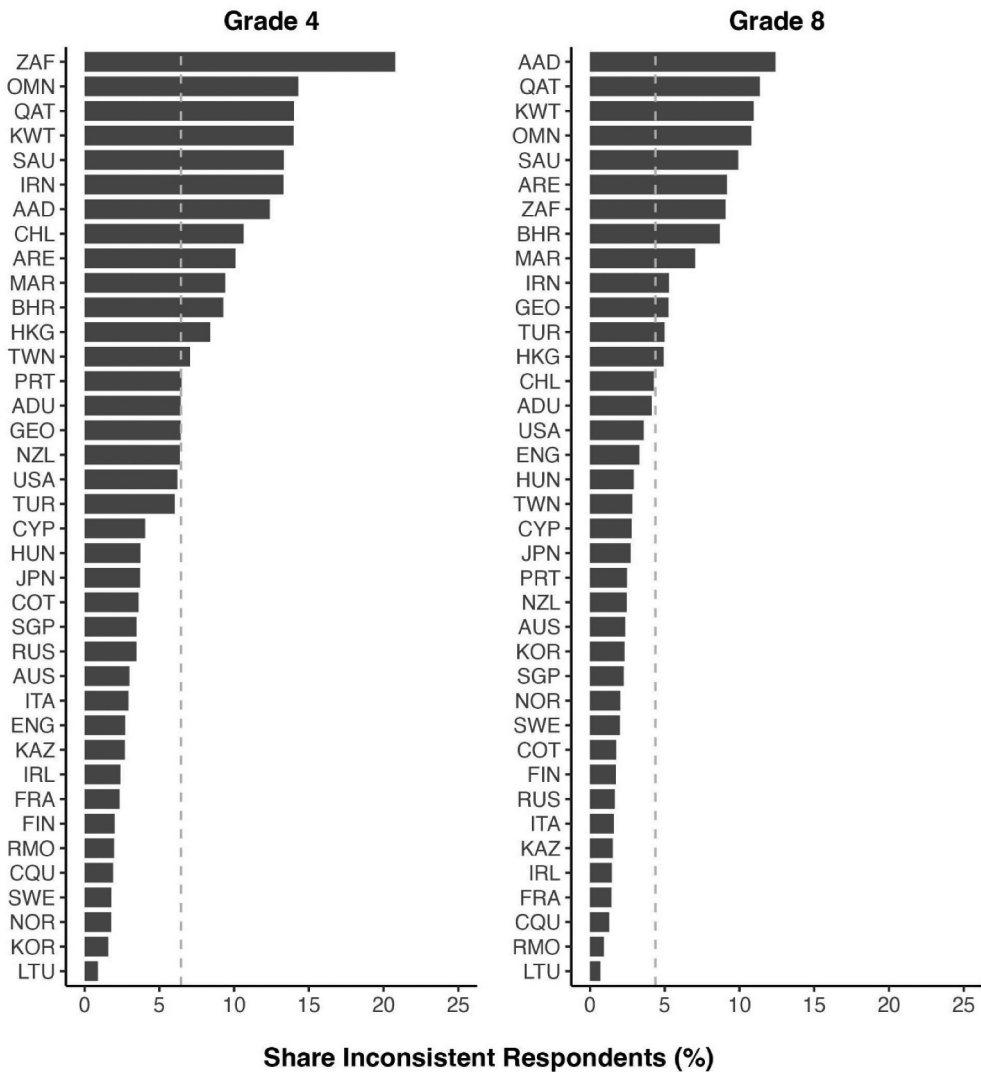


Figure 2. Prevalence of inconsistent respondents in Grades 4 and 8. *Note.* Displayed are the shares of inconsistent respondents (x-axes) across countries (y-axes). Vertical dashed lines indicate mean shares of inconsistent respondents across countries.

3.15.0 (van Buuren, 2011), mitools, version 2.4 (Lumley, 2019), metafor, version 4.2–0 (Viechtbauer, 2010), and survey, version 4.2–1 (Lumley, 2023). This study's design and its analyses were not preregistered.

Results

Using the mean absolute difference method with a threshold of 1.75, 6% of Grade 4 students and 4% of Grade 8 students were flagged as inconsistent respondents on average across countries, ranging from 1% in Lithuania in Grade 8 up to 21% in South Africa in Grade 4 (see Figure 2).¹

Findings regarding first research question on student-level correlates of inconsistent responding

Figure 3 shows the results of regressing inconsistent responding on all four student-level covariates at once (mathematics achievement, student age, language at home, and gender) in the different countries and the two grade levels. After accounting for the other variables, we found that mathematics achievement was the student-level covariate that had the most consistent, significant association with inconsistent responding across countries and grade levels. In all cases, students with higher mathematics achievement scores were less likely to be flagged as inconsistent respondents; this is also reflected in the all-negative prediction intervals for the corresponding log odds ratio (Grade 4: 95%-PI [-1.67, -0.36]; Grade 8: 95%-PI [-1.37, -0.45]).

After adjusting for the other covariates, we found that student age was not significantly associated with inconsistent responding in almost all cases and across countries (Grade 4: 95%-PI [-0.16, 0.09]; Grade 8: 95%-PI [-0.07, 0.16]; see Figure 3). This pattern aligned somewhat with our expectations; we assumed that inconsistent responding would be more common among younger students because of their lower maturity (i.e. their lower ability levels). It is thus not surprising that the association with age was mostly not statistically significant after adjusting for mathematics achievement, among other variables.

The results for the language-at-home predictor were similar to that for age when including all covariates at once (see Figure 3). The language variable was not significantly associated with inconsistent responding in the majority of cases, as well as across countries (Grade 4: 95%-PI [-0.15, 0.26]; Grade 8: 95%-PI [-0.67, 0.36]; see Figure 3). In a few cases, students who (almost) always spoke the test language at home were less likely to respond inconsistently.

Lastly, in the majority of cases, girls were less likely to be flagged as inconsistent respondents in the models that included all four covariates (Grade 4: 95%-PI [-0.81, -0.05]; Grade 8: 95%-PI [-1.23, -0.12]; see Figure 3); in the others, the association was not statistically significant (see confidence intervals in Figure 3).

The results of the simple logistic regression models are displayed in Appendix A.

Findings regarding second research question on country-level correlates of inconsistent responding

Regressing shares of inconsistent respondents on mean mathematics achievement

Figure 4 displays the results of the simple linear regression of the shares of inconsistent respondents on the mean mathematics achievement of the 38 country samples. We ran these analyses separately for Grade 4 and 8 data. At both grade levels, we found that countries with higher mean achievement had significantly lower shares of inconsistent respondents. Specifically, the regression results suggest that a 100-scale-point-higher mathematics achievement mean was associated with a 6-percentage-point-lower share of inconsistent respondents in Grade 4 ($b = -0.06$, $SE = 0.01$, $p < .001$) and a 4-percentage-points-lower share in Grade 8 ($b = -0.04$, $SE = 0.01$, $p < .001$). These findings aligned with our expectations and suggested that there are more inconsistent respondents in lower-achieving countries.

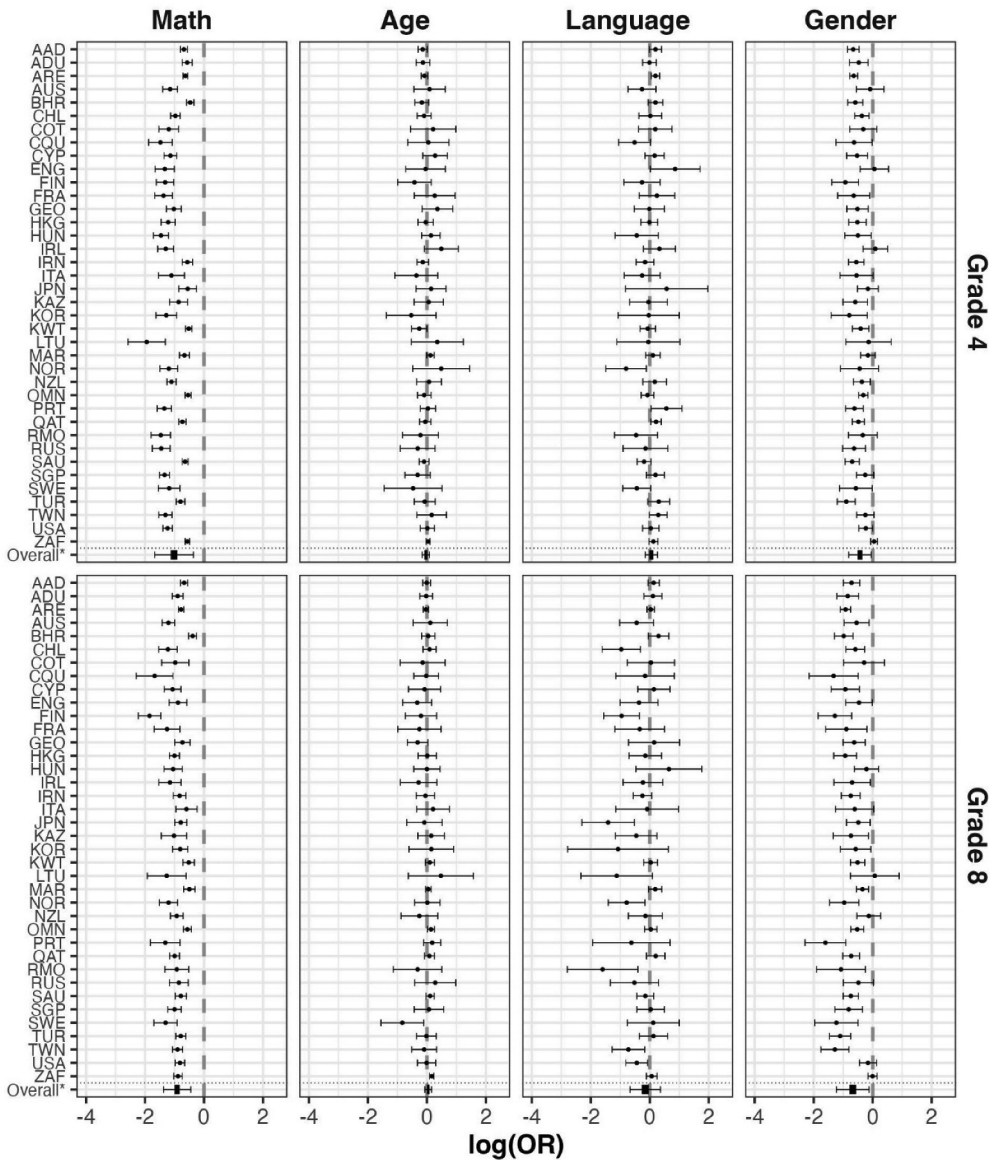


Figure 3. Results of regressing inconsistent responding variable on mathematics achievement, student age, language at home, and gender in grades 4 and 8. Note. Displayed are log odds ratios ($\log(\text{OR})$) (x-axes) across countries (y-axes). A positive/negative $\log(\text{OR})$ indicates that the odds of being classified as an inconsistent respondent is higher/lower for students scoring 100 scale score points higher on the mathematics achievement test, for students who are one year older, who (almost) always speak the language of the test at home, and for girls. The grey dashed vertical line is drawn at $\log(\text{OR}) = 0$, corresponding to independence between the covariate and the random respondent classification. Horizontal whiskers indicate 95% confidence intervals. *Overall: The rectangle represents the 95% confidence interval around the estimated average log odds ratio across countries, and the whiskers extending the rectangle represent the corresponding 95% prediction interval.

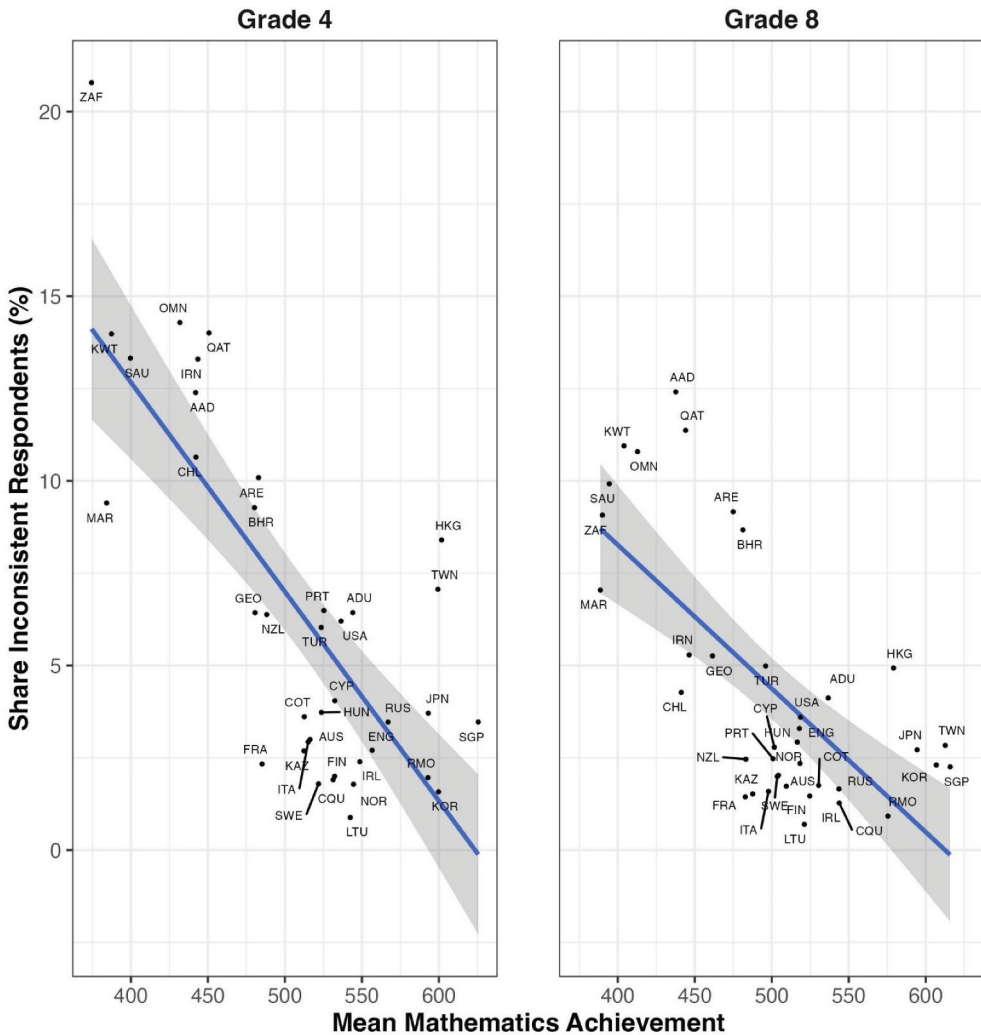


Figure 4. Results of regressing shares of inconsistent respondents on mean mathematics achievement in Grades 4 and 8. *Note.* Displayed are the results of regressing the shares of inconsistent respondents (y-axes) on the mean mathematics achievement scores (x-axes) in simple linear regressions (blue lines) in 38 countries in Grades 4 and 8 in scatterplots.

Regressing shares of inconsistent respondents on grade level

The results of the paired t-test showed a significantly ($\Delta = 2.07\%$, $t(df = 37) = 5.21$, $p < .001$) lower within-country share of inconsistent respondents in Grade 8 (4%) compared to Grade 4 (6%) (see Figure 2). This means that adolescents were less likely to respond inconsistently to the same scales than children from the same countries.

Regressing shares of inconsistent respondents on shares of students who (almost) always spoke the test language at home

Figure 5 shows the results of regressing the shares of inconsistent respondents on the share of students who (almost) always spoke the test language at home

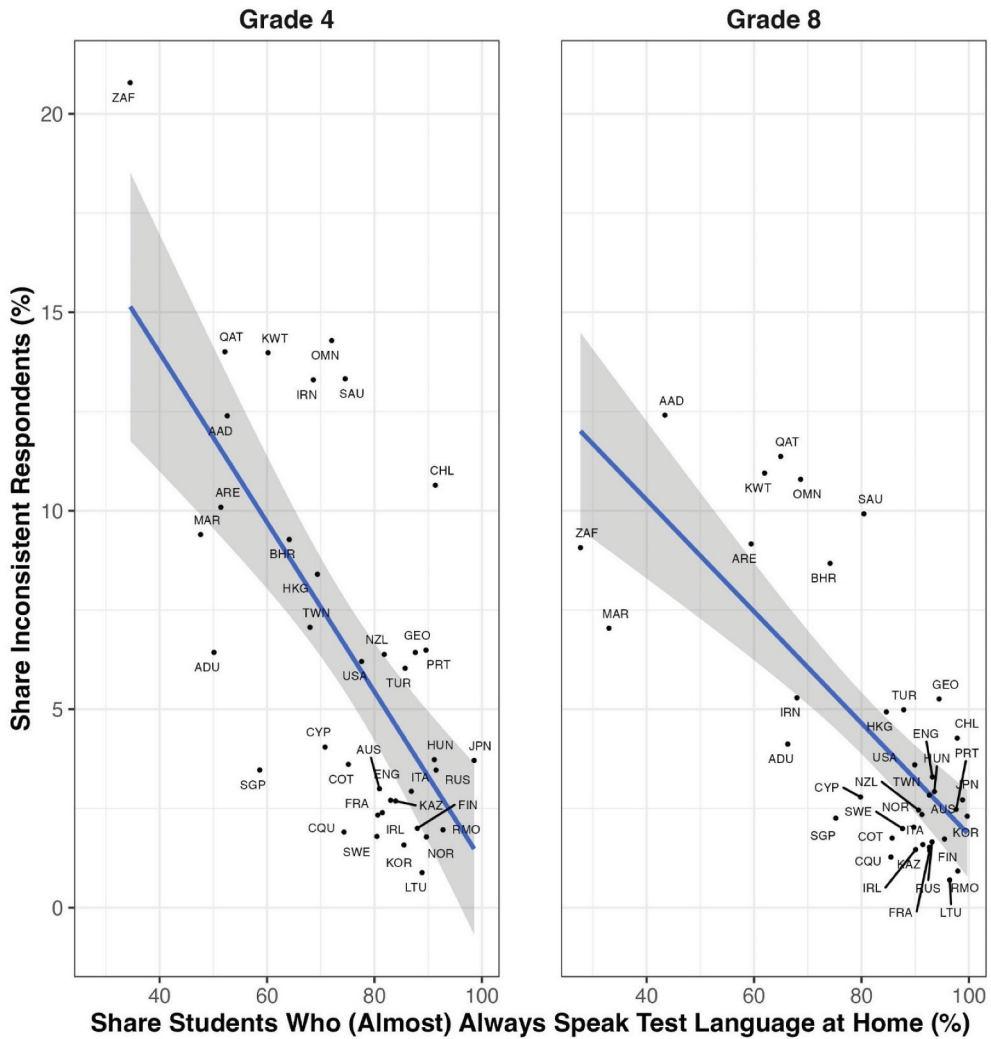


Figure 5. Results of regressing shares of inconsistent respondents on share of students who (almost) always speak test language at home in Grades 4 and 8.

separately in Grades 4 and 8. As expected, we found that countries where more students spoke the test language almost always or always had lower shares of inconsistent respondents. Specifically, a 10-percentage-points-larger share of students who (almost) always spoke the test language at home was associated with a 2-percentage-points-lower share of inconsistent respondents in Grade 4 ($b = -0.21$, $SE = 0.04$, $p < .001$) and a 1-percentage-point lower share in Grade 8 ($b = -0.14$, $SE = 0.02$, $p < .001$). This suggests that countries with more native speakers had lower shares of inconsistent respondents.

Discussion

This study investigated which students are more likely to respond inconsistently to mixed-worded questionnaire scales, and which country samples have larger shares of inconsistent respondents to such scales. At the student level, we investigated four predictor variables of inconsistent responding, separately and in a joint model: mathematics achievement, student age, language at home, and gender. Across the different models, countries, and grade levels, mathematics achievement stood out as the strongest and most consistent predictor of inconsistent responding. In all analyses, students with higher mathematics achievement scores were less likely to respond inconsistently to a mixed-worded questionnaire scale. For the other three predictors, the associations with inconsistent responding were not always significantly different from zero, and for age and language at home, we found different patterns of results in the simple and multiple logistic regressions.

In the models that included all four predictors, the results suggested (i) that in many countries, girls are less likely to respond inconsistently than boys, (ii) that in some countries, students who (almost) always speak the test language at home are less likely to respond inconsistently than those who sometimes or never speak the test language at home, and (iii) that in a few countries, older students are less likely to respond inconsistently than younger classmates. At the country level, we investigated three predictor variables separately: mean mathematics achievement, grade level, and the share of students who (almost) always spoke the test language at home. Conforming to expectations, we found lower shares of inconsistent respondents in (i) countries with higher mean achievement, (ii) in Grade 8 compared to Grade 4, and (iii) in countries with larger shares of students who (almost) always speak the test language at home.

One interpretation of these findings is that answering consistently to mixed-worded questionnaire items is indeed challenging; low-achieving students are more likely to give inconsistent responses. Previous studies have followed the same line of argument and found similar associations with achievement (Melnick & Gable, 1990; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022; Steinmann, Strietholt, et al., 2022). However, causal inference is impossible to draw with the data and study design at hand; other mechanisms could well explain this association. Careless, unmotivated respondents could give inconsistent responses to the mixed-worded items because they do not notice the changing item wording *and* score low on the mathematics achievement test because they do not really try to solve the mathematical problems. Our study cannot directly differentiate between the two common explanations for inconsistent responding: a lack of skills or carelessness.

However, we included predictor variables other than mathematics achievement to further investigate whether inconsistent responding can be traced back to a lack of skills or carelessness. We showed that in many countries, inconsistent responding was more common among boys than girls, younger than older classmates, and nonnative speakers than native speakers (in models that also adjusted for mathematics achievement). These findings seem to weigh against the carelessness explanation, as this would imply that boys, younger students, and nonnative speakers were often more careless than girls, older students, and native speakers. Also at the country level, we found larger shares of inconsistent respondents in countries with many nonnative speakers, and larger shares

in Grade 4 than in Grade 8. According to the carelessness explanation, larger shares of inconsistent respondents would be expected in Grade 8, since adolescents could be expected to show less diligence in responding than children in low-stakes assessments such as TIMSS (Silm et al., 2020). Thus, our findings can be interpreted to tentatively support the lack-of-skills explanation over the carelessness explanation for inconsistent responding. However, our design does not allow us to directly test the two explanations. Studies that have used eye tracking (Baumgartner et al., 2018) or response time measures (Swain et al., 2008) suggest that inconsistent respondents take more time to respond to negatively worded items; this suggests that they notice the negative wording, but still give an inconsistent response. These findings also seem to discourage the carelessness explanation. However, more research is needed to help us understand what leads to inconsistent responses.

Limitations

A major limitation of this study is the lack of indicators for careless responding (e.g. instructed response items or response time). With the data at hand, we were unable to directly differentiate between inconsistent responding rooted in a lack of skills or in a lack of carefulness. It would have been preferable to use a reading covariate instead of the mathematics achievement proxy, and it would have been interesting to include several more student-level covariates (e.g. dyslexia, personality, and cognitive abilities). However, we chose the TIMSS data because it enabled us to compare the shares of inconsistent respondents in Grade 4 and Grade 8, and across many countries. Another issue is that we did not investigate further potential differences in inconsistent responding between countries, such as those related to language or culture. Mixed wording may be less commonly used or more difficult to spot in some languages than others.

On a more general note, our study proceeds from the assumption that strongly agreeing or disagreeing with both positively and negatively worded items of the same scale implies an inconsistent response. This follows from the design principles of mixed-worded scales that state that the mixed-worded items should work in an opposite manner, and that after reverse-coding the negatively worded items, scale scores measure a continuum of, for instance, very low to very high mathematics self-concept. However, for some respondents, strongly agreeing or disagreeing with both item types might be an intended, subjectively meaningful response. Understanding of what an inconsistent statement is might also vary between countries due to different degrees of tolerance for contradiction in different cultures (Peng & Nisbett, 1999).

Another limitation of this study concerns the nature of the mixed-worded questionnaire scale that we used. We were not able to systematically distinguish between different types of negatives (e.g. 'bad' and 'not good'; Baumgartner et al., 2018). It might also be preferable to use mixed-worded questionnaire scales with antonym items and more than four response categories in research on inconsistent responding, rather than those available in TIMSS. With antonym items, it should be clearer that an inconsistent response is indeed perceived as contradiction for all respondents. Having more response categories would enable a more fine-grained understanding of more and less extreme inconsistent response patterns.

Implications

The central implications of our findings lie in three fields: What do our findings imply for the development of future questionnaire scales, for research that uses data from mixed-worded scales, and for future research on inconsistent responding?

Since both the implications for the development of future questionnaire scales and the implications for research that uses data from mixed-worded scales highly depend on which of the two explanations for inconsistent responding holds – a lack of skills or carelessness – future research on inconsistent responding should address the question: Why do some respondents answer inconsistently to mixed-worded scales? Such research should consider that different reasons may apply for different individuals and age groups and that some inconsistent responses might even be due to both a lack of attention *and* skills.

If inconsistent responding is rooted in a lack of attention or carefulness in filling out questionnaires, implementing mixed-worded scales allows researchers to identify such careless respondents and remove them from the data to improve data quality (cf. Patton et al., 2019). If inconsistent responding is, however, rooted in an inability to handle the mixed wording, implementing mixed-worded questionnaire scales would create the phenomenon of inconsistent responses and decrease data quality. Depending on the reason for inconsistent responding, mixed wording might either make a problem visible or create it in the first place. There are cases to be made for either implementing mixed-worded questionnaire scales as data-cleaning tools (e.g. Hong et al., 2020) or for avoiding mixed-worded scales (e.g. Lenzner & Menold, 2016) in future questionnaire scale development. Relatedly, if inconsistent responding is rooted in a lack of skills, research that uses data from mixed-worded scales should consider only using the positively worded items to create scale scores, for example (e.g. Marsh, 1996). If inconsistent responding is rooted in carelessness, however, research that uses data from mixed-worded scales should flag inconsistent respondents and remove them from the data to attain datasets with only diligent respondents (e.g. Patton et al., 2019).

The question of whether to include mixed-worded scales in questionnaires ultimately relies on whether they lead to better or worse data quality. Experimental studies that compare positively worded and mixed-worded questionnaire scales seem to favour positively worded ones (e.g. Cole et al., 2019; Roszkowski & Soven, 2010; Zeng et al., 2020). Numerous studies, like this study, find that for whatever reason, some respondents answer in an inconsistent fashion to mixed-worded scales (e.g. Bulut & Bulut, 2022; Hong et al., 2020; Steinmann, Strietholt, et al., 2022), which is not a logical response according to the mixed-wording design. Removing such inconsistent respondents from datasets improves certain data quality characteristics, such as dimensionality and reliability (e.g. Arias et al., 2020; Steedle et al., 2019; Steinmann, Sánchez, et al., 2022). Even if our study could not determine causes of inconsistent responding, the correlations with student- and country-level covariates suggest that inconsistent responding is not a randomly occurring phenomenon. This is worrisome, especially in international large-scale assessments such as TIMSS, since the questionnaires are supposed to work equally well for different student groups and in different countries (Desa et al., 2018).

In conclusion, this study and previous studies suggest that there are good reasons for avoiding mixed wording in questionnaire scales, especially if the respondents are young beginner readers. The notion that mixed wording introduces an additional reading and cognitive difficulty in responding is plausible and somewhat supported by empirical evidence (e.g. Baumgartner et al., 2018; Steinmann, Strietholt, et al., 2022), and the questionnaires should be as easy to fill out for everyone as possible (cf. Schulz & Carstens, 2020). There are alternative methods for detecting careless, inattentive respondents instead of using mixed-worded scales (e.g. overly fast response time, straight lining across scales, or instructed-response items). These might be preferable.

Note

1. We adopted two other thresholds to conduct sensitivity analyses. With a looser threshold of 1.5, 9% of Grade 4 students and 6% of Grade 8 students were flagged as inconsistent respondents on average across countries. With a more stringent threshold of 2, 5% of Grade 4 students and 3% of Grade 8 students were flagged as inconsistent respondents on average across countries. Although the threshold affected the shares of inconsistent respondents flagged, the general patterns of the remaining core results were stable.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the International Association for the Evaluation of Educational Achievement (IEA) Research and Development Fund. Some of the findings were reported at the 2023 IEA International Research Conference, Dublin, Ireland. We would like to thank David Mottram for copyediting the manuscript.

Notes on contributors

Isa Steinmann is associate professor at the Department of Primary and Secondary Teacher Education at Oslo Metropolitan University, Norway. One strand of her research focuses on determining how education systems and schools affect student achievement and educational inequality outcomes. In this field, she enjoys applying methods that aim for causal inference from large-scale assessment data. Another strand of her research interests concerns how properties of international large-scale assessments are linked to their results and interact with the respondents.

Jianan Chen was a research assistant at Oslo Metropolitan University when collaborating on the article. Currently, she is a PhD candidate at CREATE - Centre for Research on Equality in Education, affiliated with CEMO - Centre for Educational Measurement at the University of Oslo. Her research focuses on the construction and validation of progressive tests for screening language skills based on modern design principles. Besides, she is interested in investigating aberrant response patterns in large-scale assessments.

Johan Braeken is professor in psychometrics at CEMO, the Centre for Educational Measurement at the University of Oslo, Norway. His research interests are in latent variable modelling, modern test design, and meta-research at the item-response level of international large-scale assessments.

This includes among others computerized adaptive testing and aberrant item response patterns on surveys.

ORCID

Isa Steinmann  <http://orcid.org/0000-0002-9940-4413>

Jianan Chen  <http://orcid.org/0000-0002-0421-6058>

Johan Braeken  <http://orcid.org/0000-0002-2119-3222>

References

- Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, 52(6), 2489–2505. <https://doi.org/10.3758/s13428-020-01401-8>
- Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *Journal of Marketing Research*, 55(6), 869–883. <https://doi.org/10.1177/0022243718811848>
- Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics*, 121(4), 1437–1472. <https://doi.org/10.1162/qjec.121.4.1437>
- Bolt, D., Wang, Y. C., Meyer, R. H., & Pier, L. (2020). An IRT mixture model for rating scale confusion associated with negatively worded items in measures of social-emotional learning. *Applied Measurement in Education*, 33(4), 331–348. <https://doi.org/10.1080/08957347.2020.1789140>
- Brevik, L. M., Olsen, R. V., & Hellekjær, G. O. (2016). The complexity of second language reading: Investigating the L1-L2 relationship. *Reading in a Foreign Language*, 28(2), 161–182. <https://doi.org/10.125/66899>
- Bulut, H. C., & Bulut, O. (2022). Item wording effects in self-report measures and reading achievement: Does removing careless respondents help? *Studies in Educational Evaluation*, 72, 101126. <https://doi.org/10.1016/j.stueduc.2022.101126>
- Chen, J., Steinmann, I., & Braeken, J. (in print). Competing explanations for inconsistent responding to a mixed-worded self-esteem scale: Cognitive abilities or personality? *Personality and Individual Differences*.
- Cole, K. L., Turner, R. C., & Gitchel, W. D. (2019). A study of polytomous IRT methods and item wording directionality effects on perceived stress items. *Personality and Individual Differences*, 147, 63–72. <https://doi.org/10.1016/j.paid.2019.03.046>
- Desa, D., Van De Vijver, F. J. R., Carstens, R., & Schulz, W. (2018). Measurement invariance in international large-scale assessments: Integrating theory and method. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods* (pp. 879–910). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118884997.ch40>
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6
- Ebbs, D., Wry, E., Wagner, J.-P., & Netten, A. (2020). Instrument translation and layout verification for TIMSS 2019. In Martin, M. O., Von Davier, M., & Mullis, I. V. S. Eds., *Methods and procedures: TIMSS 2019 technical report* (pp. 5.1–5.23). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). <https://timssandpirls.bc.edu/timss2019/methods/chapter-5.html>

- Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In Martin, M. O., Von Davier, M., & Mullis, I. V. S. eds., *Methods and procedures: TIMSS 2019 technical report* (pp. 12.1–12.146). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). <https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html>
- García-Batista, Z. E., Guerra-Peña, K., Garrido, L. E., Cantisano-Guzmán, L. M., Moretti, L., Cano-Vindel, A., Arias, V. B., & Medrano, L. A. (2021). Using constrained factor mixture analysis to validate mixed-worded psychological scales: The case of the Rosenberg self-esteem scale in the Dominican Republic. *Frontiers in Psychology, 12*, 636693. <https://doi.org/10.3389/fpsyg.2021.636693>
- Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg self-esteem scale. *Assessment, 27*(2), 404–418. <https://doi.org/10.1177/1073191117746503>
- Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80*(2), 312–345. <https://doi.org/10.1177/0013164419865316>
- Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education, 84*(1), 175–196. <https://doi.org/10.1080/00220973.2014.952397>
- Kam, C. C. S., & Chan, G. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences, 129*, 83–87. <https://doi.org/10.1016/j.paid.2018.03.022>
- Kam, C. C. S., & Meyer, J. P. (2015). Implications of item keying and item valence for the investigation of construct dimensionality. *Multivariate Behavioral Research, 50*(4), 457–469. <https://doi.org/10.1080/00273171.2015.1022640>
- Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist, 73*(1), 1–2. <https://doi.org/10.1037/amp0000263>
- Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development: Reading and language learning. *Language Learning, 57*(s1), 1–44. <https://doi.org/10.1111/0023-8333.101997010-il>
- LaRoche, S., Joncas, M., & Foy, P. (2020). Sample design in TIMSS 2019. In M. O. Martin, M. V. Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 31–333) TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). <https://timssandpirls.bc.edu/timss2019/methods/chapter-3.html>
- Lenzner, T., & Menold, N. (2016). GESIS survey guidelines: Question wording (2.0). SDM-Survey Guidelines (GESIS Leibniz Institute for the Social Sciences). https://doi.org/10.15465/gesis-sg_en_017
- Likert, R. (1974). The method of constructing an attitude scale. In G. M. Maranell (Ed.), *Scaling. A sourcebook for behavioral scientists* (pp. 233–243). Aldine Publ.
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment, 94*(2), 196–204. <https://doi.org/10.1080/00223891.2011.645936>
- Lumley, T. (2019). *Mitools: Tools for Multiple Imputation of Missing Data* (2.4) [Computer Software]. <https://cran.r-project.org/web/packages/mitools/index.html>
- Lumley, T. (2023). *Survey: Analysis of Complex Survey Samples* (4.2) [Computer Software]. <https://cran.r-project.org/web/packages/survey/index.html>
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810–819. <https://doi.org/10.1037/0022-3514.70.4.810>
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., Xu, M. K., Nagengast, B., & Parker, P. (2013). Factorial, convergent, and discriminant validity of

- TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, 105(1), 108–128. <https://doi.org/10.1037/a0029907>
- Martin, M. O., Von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
- Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (Vol. 10, pp. 113–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_7
- Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems. *Educational Research Quarterly*, 14(3), 31–36.
- Michaelides, M. P. (2019). Negative keying effects in the factor structure of TIMSS 2011 motivation scales and associations with reading achievement. *Applied Measurement in Education*, 32(4), 365–378. <https://doi.org/10.1080/08957347.2019.1660349>
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement (IEA). <https://pirls2016.org/wp-content/uploads/structure/CompletePDF/P16-PIRLS-International-Results-in-Reading.pdf>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill. <http://www.loc.gov/catdir/description/mh022/93022756.html>
- OECD. (2019). *PISA 2018 results (volume II): Where all students can succeed*. OECD Publishing. <https://doi.org/10.1787/b5fd1b8f-en>
- Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, 44(3), 309–341. <https://doi.org/10.3102/1076998618825116>
- Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist*, 54(9), 741–754. <https://doi.org/10.1037/0003-066X.54.9.741>
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg self-esteem scale method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 99–117. https://doi.org/10.1207/s15328007sem1301_5
- R Core Team. (2022). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35(1), 83–92. <https://doi.org/10.1016/j.intell.2006.05.004>
- Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, 35(1), 113–130. <https://doi.org/10.1080/02602930802618344>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9(4), 367–373. <https://doi.org/10.1177/014662168500900405>
- Schulz, W., & Carstens, R. (2020). Questionnaire development in international large-scale assessment studies. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (Vol. 10, pp. 61–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_5

- Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, 31, 100335. <https://doi.org/10.1016/j.edurev.2020.100335>
- Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social-emotional learning competencies. *Educational Measurement Issues & Practice*, 38(2), 101–111. <https://doi.org/10.1111/emip.12256>
- Steinmann, I., & Olsen, R. V. (2022). Equal opportunities for all? Analyzing within-country variation in school effectiveness. *Large-Scale Assessments in Education*, 10(1), 2. <https://doi.org/10.1186/s40536-022-00120-0>
- Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2022). The impact of inconsistent

- responders to mixed-worded scales on inferences in international large-scale assessments. *Assessment in Education Principles, Policy & Practice*, 29(1), 5–26. <https://doi.org/10.1080/0969594X.2021.2005302>
- Steinmann, I., Strietholt, R., & Braeken, J. (2022). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods*, 27(4), 667–702. <https://doi.org/10.1037/met0000392>
- Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed likert items. *Journal of Marketing Research*, 45(1), 116–131. <https://doi.org/10.1509/jmkr.45.1.116>
- TIMSS & PIRLS International Study Center. (2019). *TIMSS: Trends in International Mathematics and Science Study*. <https://timssandpirls.bc.edu/timss-landing.html>
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Routledge.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 36(3). <https://doi.org/10.18637/jss.v036.i03>
- von Davier, M., Gonzalez, E. J., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 2, pp. 9–36). IERI.
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191. <https://doi.org/10.1007/s10862-005-9004-7>
- Zeng, B., Wen, H., & Zhang, J. (2020). How does the valence of wording affect features of a scale? The method effects in the undergraduate learning burnout scale. *Frontiers in Psychology*, 11, 585179. <https://doi.org/10.3389/fpsyg.2020.585179>

Appendix A: Simple Regression Results Regarding First Research Question

Regressing Inconsistent Responding on Mathematics Achievement

Figure A1 shows the results of the simple logistic regressions of inconsistent responding on mathematics achievement in all countries and in Grades 4 and 8. Conforming to our expectations, students with higher mathematics test scores were significantly less likely to be flagged as

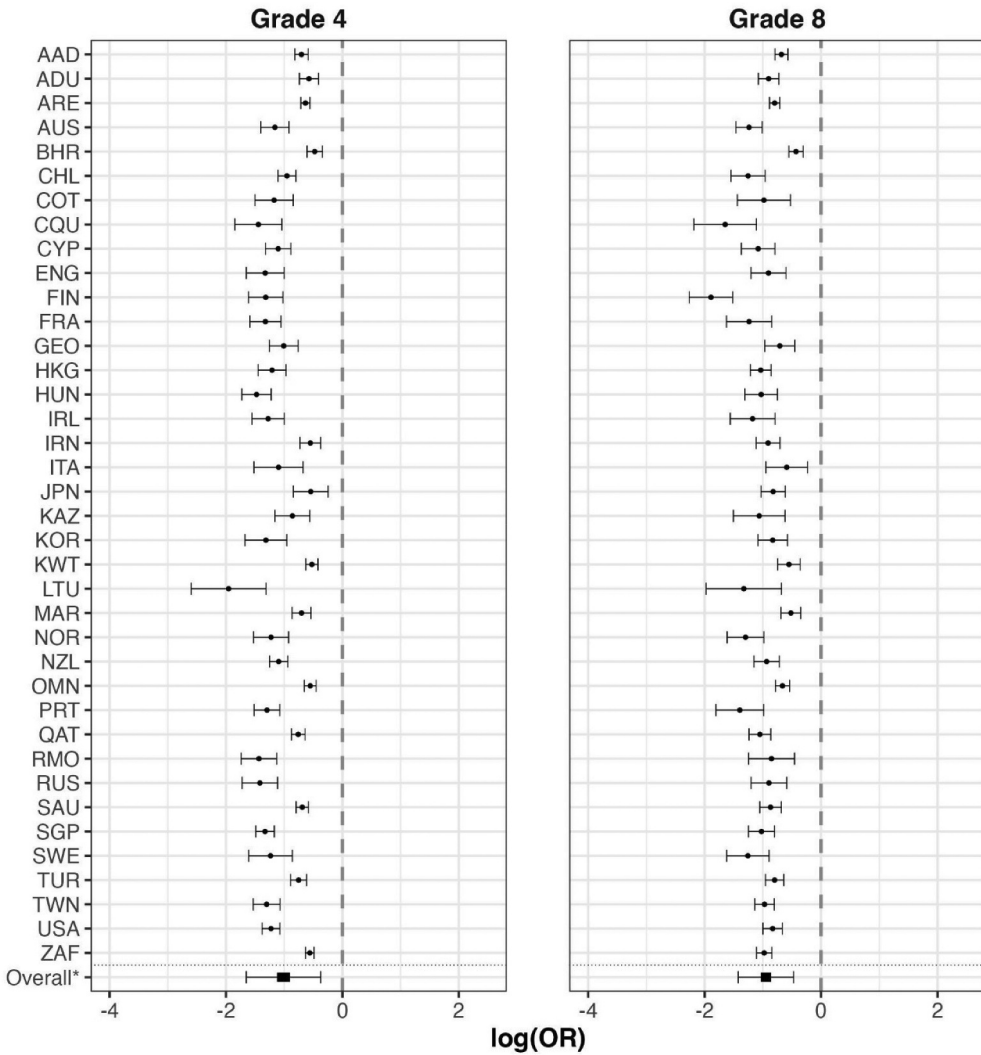


Figure A1. Results of regressing inconsistent responding variable on mathematics achievement in Grades 4 and 8. Note. Displayed are log odds ratios ($\log(\text{OR})$) (x-axes) across countries (y-axes). A positive/negative $\log(\text{OR})$ indicates that the odds of being classified as an inconsistent respondent is higher/lower for students scoring 100 scale score points higher on the mathematics achievement test. The grey dashed vertical line is drawn at $\log(\text{OR}) = 0$, corresponding to independence between the covariate and the random respondent classification. Horizontal whiskers indicate 95% confidence intervals. *Overall: The rectangle represents the 95% confidence interval around the estimated average log odds ratio across countries, and the whiskers extending the rectangle represent the corresponding 95% prediction interval.

inconsistent respondents in all countries and both grade levels. On average across countries, the odds of being flagged as inconsistent respondents for students who scored 100 points higher on the mathematics test were about one third of the odds for lower-performing students in both Grade 4 ($\exp(-1.01) = 0.36$, $\log(\text{OR}) = -1.01$, 95%-CI $[-1.12, -0.90]$) and Grade 8 ($\exp(-0.95) = 0.39$, $\log(\text{OR}) = -0.95$, 95%-CI $[-1.03, -0.86]$).

Regressing Inconsistent Responding on Student Age

Figure A2 shows the results of the simple logistic regressions of inconsistent responding on student age in all countries and in Grades 4 and 8. In more than two-thirds of cases, students' age did not significantly predict inconsistent responding. However, in some cases, older students were more likely to be flagged as inconsistent respondents (e.g. Hungary in Grade 4), whereas in other cases, older students were less likely to be flagged as inconsistent respondents (e.g. Korea in Grade 4). On average across countries, the random effects meta-analysis suggests that the odds of being flagged as inconsistent respondents for one-year-older students were about the same as the odds for younger students in Grade 4 ($\log(\text{OR}) = -0.01$, 95%-CI $[-0.11, 0.09]$), but 1.20 times higher in Grade 8 ($\log(\text{OR}) = 0.18$, 95%-CI $[0.09, 0.27]$).

The cases where older students were more likely to respond inconsistently did not support our a priori assumption that inconsistent responding would be more common among younger students due to differences in maturity. The association between inconsistent responding and age could, however, have been affected by third-variable effects in some countries. If countries applied flexible school entry or grade repetition or acceleration policies, more academically mature students might have been promoted, while less mature students might have been held back (cf., Bedard & Dhuey, 2006; Steinmann & Olsen, 2022). In this case, older students would not necessarily be more academically mature than their younger classmates. The results of the multiple logistic regressions (see Figure 3) that included mathematics achievement as an additional covariate put the results of these simple logistic regressions into perspective.

Regressing Inconsistent Responding on Language Spoken at Home

Figure A3 shows the results of the simple logistic regressions of inconsistent responding on language spoken at home in all countries and in Grades 4 and 8. In the majority of cases, the language at home variable did not significantly predict being flagged as an inconsistent respondent. Yet in some cases (e.g. Qatar in Grades 4 and 8), students who (almost) always spoke the test language at home were more likely to be flagged and in other cases (e.g. Norway in Grades 4 and 8) less likely to be flagged as inconsistent respondents. On average across countries, the odds of being flagged as inconsistent respondents for students who reported (almost) always speaking the test language at home were about the same as the odds for students who reported sometimes or never speaking the test language at home in Grade 4 ($\log(\text{OR}) = -0.06$, 95%-CI $[-0.17, 0.05]$), but 0.65 times lower in Grade 8 ($\log(\text{OR}) = -0.43$, 95%-CI $[-0.62, -0.25]$).

The cases where students who (almost) always spoke the test language at home were more likely to be flagged as inconsistent respondents contradicted our expectations. These unexpected findings could, however, be associated with between-country differences in the performance of nonnative groups; in some countries, students with an immigration background score higher in reading tests than those without (e.g. OECD, 2019). As for the student age above, the results of the multiple logistic regressions (see Figure 3) put these findings into perspective.

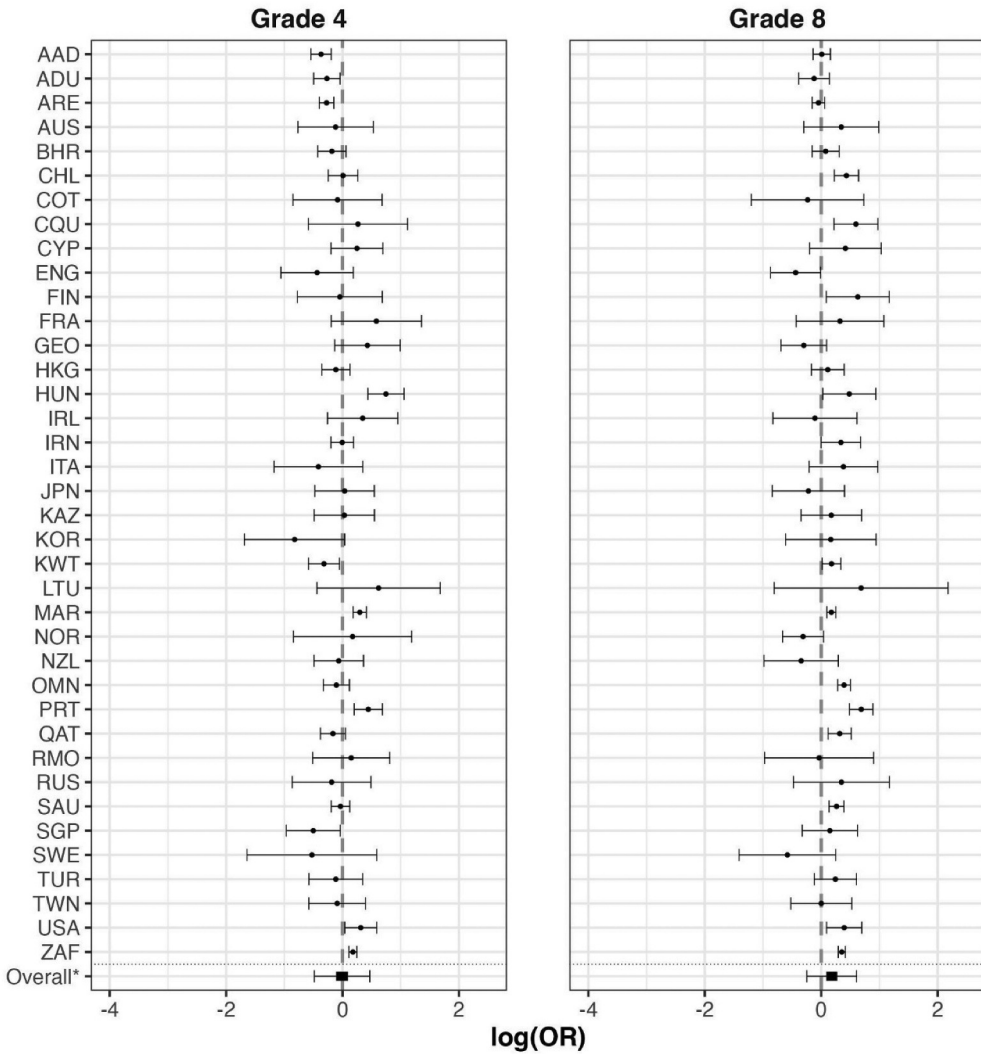


Figure A2. Results of regressing inconsistent responding variable on student age in Grades 4 and 8. *Note.* Displayed are log odds ratios (log (OR)) (x-axes) across countries (y-axes). A positive/negative log(OR) indicates that the odds of being classified as an inconsistent respondent is higher/lower for students that are one year older. The grey dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random respondent classification. Horizontal whiskers indicate 95% confidence intervals. *Overall: The rectangle represents the 95% confidence interval around the estimated average log odds ratio across countries, and the whiskers extending the rectangle represent the corresponding 95% prediction interval.

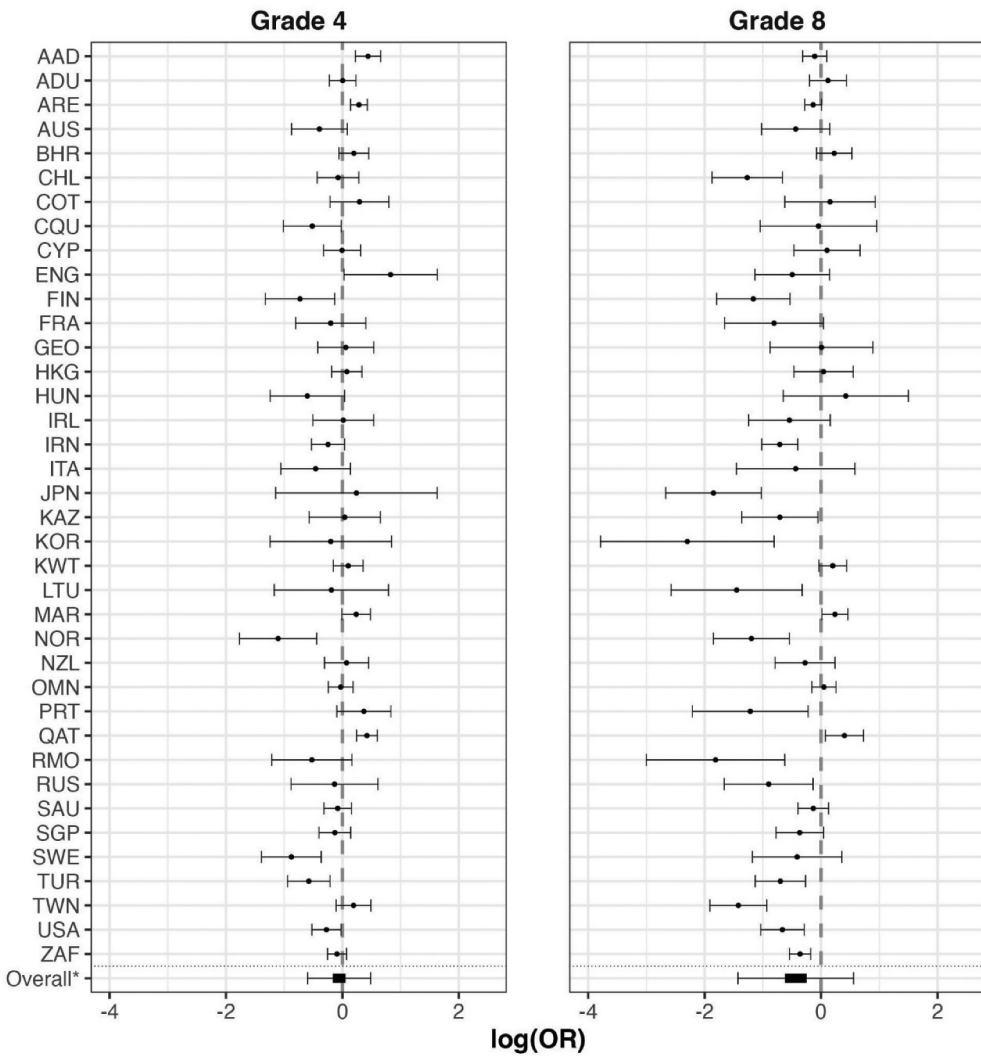


Figure A3. Results of regressing inconsistent responding variable on language at home in Grades 4 and 8. *Note.* Displayed are log odds ratios (log (OR)) (x-axes) across countries (y-axes). A positive/negative log(OR) indicates that the odds of being classified as an inconsistent respondent is higher/lower for students who (almost) always speak the test language at home than for those who do not. The grey dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random respondent classification. Horizontal whiskers indicate 95% confidence intervals. *Overall: The rectangle represents the 95% confidence interval around the estimated average log odds ratio across countries, and the whiskers extending the rectangle represent the corresponding 95% prediction interval.

Regressing Inconsistent Responding on Gender

The results of regressing inconsistent responding on gender (0 = boy and 1 = girl) are displayed in Figure A4 across the different countries and for Grades 4 and 8. Aligning with our hypotheses, we observed that in about two thirds of cases, girls were significantly less likely to be flagged as inconsistent respondents than boys. In the remaining cases, the log odds ratio was not significantly different from zero. On average across countries, the odds of being flagged as inconsistent respondents for girls was about half of the odds for boys (Grade 4: $\log(\text{OR}) = -0.39$, 95%-CI $[-.47, -.43]$; Grade 8: $\log(\text{OR}) = -0.71$, 95%-CI $[-0.83, -0.59]$).

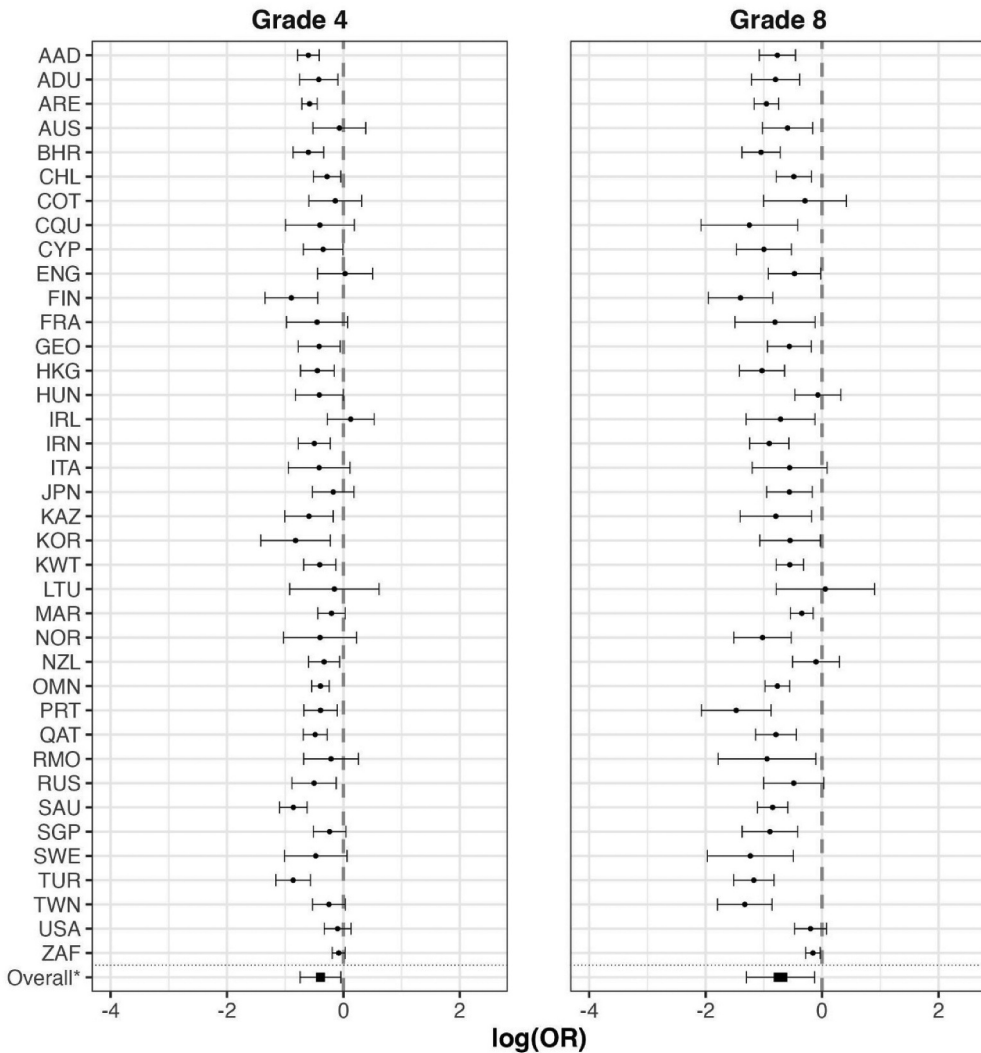


Figure A4. Results of regressing inconsistent responding variable on gender in Grades 4 and 8. *Note.* Displayed are log odds ratios ($\log(\text{OR})$) (x-axes) across countries (y-axes). A positive/negative $\log(\text{OR})$ indicates that the odds of being classified as an inconsistent respondent is higher/lower for girls than for boys. The grey dashed vertical line is drawn at $\log(\text{OR}) = 0$, corresponding to independence between the covariate and the random respondent classification. Horizontal whiskers indicate 95% confidence intervals. *Overall: The rectangle represents the 95% confidence interval around the estimated average log odds ratio across countries, and the whiskers extending the rectangle represent the corresponding 95% prediction interval.