



The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments

Isa Steinmann ^{a,b}, Daniel Sánchez^a, Saskia van Laar^a and Johan Braeken ^a

^aCentre for Educational Measurement, University of Oslo, Oslo, Norway; ^bCenter for Research on Education and School Development, Tu Dortmund University, Dortmund, Germany

ABSTRACT

Questionnaire scales that are mixed-worded, i.e. include both positively and negatively worded items, often suffer from issues like low reliability and more complex latent structures than intended. Part of the problem might be that some responders fail to respond consistently to the mixed-worded items. We investigated the prevalence and impact of inconsistent responders in 37 primary education systems participating in the joint PIRLS/TIMSS 2011 assessment. Using the mean absolute difference method and three mixed-worded self-concept scales, we identified between 2%–36% of students as inconsistent responders across education systems. Consistent with expectations, these students showed lower average achievement scores and had a higher risk of being identified as inconsistent on more than one scale. We also found that the inconsistent responders biased the estimated dimensionality and reliability of the scales. The impact on external validity measures was limited and unsystematic. We discuss implications for the use and development of questionnaire scales.

ARTICLE HISTORY

Received 26 June 2021
Accepted 27 October 2021

KEYWORDS


Inconsistent responders; mixed-worded; PIRLS/TIMSS; dimensionality; reliability; validity

In education survey research, questionnaire scales, especially mixed-worded ones, often suffer from issues such as low reliability and more complex latent structures than intended. This study assumes that the failure of some responders to respond consistently to the mixed item wording is part of the problem. Mixed-worded questionnaire scales include both positively and negatively worded items such as ‘I usually do well in mathematics’ and ‘I am just not good at mathematics’. The mixed wording is meant to encourage more thorough answering behaviour because the responders have to read each item carefully in order to give a consistent response. Questionnaire developers employ mixed wording to improve the measurement properties of scales while assessing the same underlying constructs with both item types (e.g. Idaszak & Drasgow, 1987; Podsakoff et al., 2003).

To respond consistently to such mixed-worded questionnaire scales, the responder must tick opposite sides of the response scale. This is illustrated in Panel A in Figure 1, where a responder strongly agrees with a positively worded item and strongly disagrees

CONTACT Isa Steinmann  isa.steinmann@cemo.uio.no  Centre for Educational Measurement, University of Oslo, Postboks 1161 Blindern, 0318 Oslo, Norway

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

 Supplemental data for this article can be accessed [here](#)

© 2021 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.

Panel A
Consistent Response

	agree a lot	agree a little	disagree a little	disagree a lot
a) I usually do well in mathematics.....	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
b) I am just not good at mathematics.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

Panel B
Inconsistent Response

	agree a lot	agree a little	disagree a little	disagree a lot
a) I usually do well in mathematics.....	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
b) I am just not good at mathematics.....	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 1. Examples of Consistent and Inconsistent Responders to Positively and Negatively Worded Items. *Note.* Panel A displays the responses of a consistent responder to a positively (a) and a negatively worded item (b). Panel B displays an equivalent example of an inconsistent responder. Items stem from the TIMSS 2011 student questionnaire (Martin & Mullis, 2012). Figure adapted from Authors (2021).

with a negatively worded item. If a responder does not switch the side of the response scale, this can be considered inconsistent because the resulting statements are, by implication, implausible. This is illustrated in Panel B of Figure 1, where another responder strongly agrees with both item types. One possible reason for such inconsistent responses is that people with low reading or cognitive skills might fail to notice the change in item wording or fail to adjust their responses accordingly (Bolt et al., 2020; Melnick & Gable, 1990; Steedle et al., 2019; Steinmann et al., 2021; Weems et al., 2003). Similarly, responders who are distracted, not highly committed, or in a hurry to fill out the questionnaire might not be careful enough to detect the changing item wording and to adjust their responses properly (Kam & Meyer, 2015; Quilty et al., 2006; Weems et al., 2003). Therefore, the mixed wording might make it more difficult to respond meaningfully to questionnaire scales (e.g. Marsh, 1986; Schmitt & Stults, 1985; Swain et al.,

2008). This is worrisome because questionnaires should be sufficiently easy to fill out so that the responses are interpretable, substantively meaningful, and not confounded with ability. However, mixed-worded scales are frequently used in education surveys like international large-scale assessments; the items in [Figure 1](#) come from a questionnaire for fourth-grade students (Trends in International Mathematics and Science Study (TIMSS) 2011; Martin & Mullis, 2012).

There are different ways to detect such inconsistent responders in questionnaire data. An early study identified inconsistent responders by analysing the answer patterns to pairs of parallel antonym items (one positively and one negatively worded item per pair) (Melnick & Gable, 1990). In a sample of US parents of school children, this study found that 23% of responders to a school-effectiveness scale gave inconsistent responses to the positively and negatively worded items (i.e. they either agreed or disagreed with both). Another study reanalysed primary data from 9 studies in which US college students and other adults responded to scales that contained at least one negatively worded item (Swain et al., 2008). They found that, on average, 18% of responders selected the same side of the Likert response scale when rating negatively and positively worded items. Two other studies utilised the mean absolute difference between responses to positively and negatively worded items to detect inconsistent responders. They found that 7% (Hong et al., 2020) or 10% (Steedle et al., 2019) of the observed US secondary school students responded too similarly to positively and negatively worded items of inventory scales on motivation, social engagement, and self-regulation. Bolt et al. (2020) used an item response theory mixture model and found that between 2% and 13% of a sample of primary and secondary school students in the US responded inconsistently to one negatively worded growth mindset and three positively worded self-management, self-efficacy, and social awareness scales. Since the positively and negatively worded items belonged to separate scales, however, inconsistencies in responding to them could also reflect substantive, plausible differences between the observed constructs. By contrast, another study used a constrained factor mixture model to identify inconsistent responders to actual mixed-worded scales on global self-esteem, reading self-concept, and mathematics self-concept (Steinmann et al., 2021). In representative samples of primary and secondary school students from the United States, Australia, and Germany, this study found that between 7% and 20% of responders were inconsistent.

Previous Findings on the Impact of Inconsistent Responders

In summary, previous studies that used different mixed-worded attitude scales and detection methods found small shares of inconsistent responders – between 2% and 23% – in samples of students, adolescents, and adults. With one exception, these studies used US samples. Yet the question of how these inconsistent responders affected analyses that use these mixed-worded scales remains unaddressed.

Since inconsistent responders do not switch the side of the response scale they select when faced with mixed item wording, additional, construct-unrelated covariance between positively and negatively worded items should consequently emerge. There are a number of possible implications of this wording-related covariance. To begin with, the scale should have a more complex latent structure. Two simulation studies investigated whether this was indeed the case and found that if as few as 10% of

responders gave inconsistent answers to unidimensional mixed-worded scales, more than one latent factor would be needed to attain an adequate model fit in factor analyses (Schmitt & Stults, 1985; Woods, 2006). Similarly, in the above-mentioned empirical study by Steedle et al. (2019), the removal of inconsistent responders led to small model fit improvements regarding the intended latent structures.

A second implication of the wording-related covariance due to inconsistent responding should be an underestimation of internal consistency reliability measures. Nevertheless, recent simulation (Hong et al., 2020) and empirical studies (Hong et al., 2020; Steedle et al., 2019) found mixed or neutral effects of excluding inconsistent responders on reliability measures.

A third implication is that inconsistent responders might systematically bias associations between the mixed-worded scales and external variables, although the direction of this bias might be difficult to anticipate (cf., Steedle et al., 2019). Steedle et al. (2019) found slightly increased correlations with external variables and among subscales after removing inconsistent responders. In contrast, Hong et al. (2020) found slightly decreased associations with external variables when removing inconsistent responders.

The Present Study

In conclusion, very few studies, predominantly from the US, have identified inconsistent responders as defined above and investigated whether dimensionality, reliability, and external validity measures change after excluding them. The few available studies have found inconclusive results. This study addressed this research gap and investigated whether inconsistent responders to mixed-worded scales biased the estimated dimensionality, reliability, and external validity measures. We therefore first identified inconsistent responders using the mean absolute difference method (cf. Hong et al., 2020; Steedle et al., 2019) and then compared results of analyses that either included or excluded them. To investigate the generalisability of our analyses, we used data from international primary school children who responded to three mixed-worded scales (on reading, mathematics and science self-concept) that were designed to be one-dimensional. Specifically, we focused on four research questions:

(1) *How many inconsistent responders are there internationally and what characteristics do they have?*

Following the literature review above, we expected inconsistent responders to be a minority in all education systems and to have lower achievement scores than consistent responders. Due to the assumed association between response behaviour and personal characteristics such as a lack of skills or a lack of carefulness, we furthermore expected inconsistent responding to be rather persistent. In other words, someone who responds inconsistently to one mixed-worded scale should also be more likely to respond inconsistently to other mixed-worded scales in the same questionnaire.

(2) *How does excluding inconsistent responders affect the mean scores of the mixed-worded questionnaire scales?*

When computing mean scores of mixed-worded scales, the responses to either the positively or negatively worded items have to be reverse-coded, first. Since the inconsistent responders give the same responses to both item types, their mean scale scores should be biased towards the middle of the response scale. When removing the inconsistent responders, the mean scale scores could therefore shift away from the middle of the response scales. In the case of academic self-concepts of primary school students, we would expect that consistent responders express rather positive self-concepts, on average.

(3) How does excluding inconsistent responders affect the dimensionality and reliability of mixed-worded questionnaire scales?

As discussed above, inconsistent responders should introduce wording-related covariance between the items of mixed-worded scales that is unrelated to the substantive constructs of interest. Therefore, factor analyses that include inconsistent responders should indicate a more complex latent structure than the intended, substantive one. Factor analyses that only include consistent responders should, by contrast, support the intended one-dimensionality of the constructs. In the same vein, the fact that inconsistent responders introduce construct-unrelated covariance should lead to higher scale reliability estimates when only including consistent responders.

(4) How does excluding inconsistent responders affect associations between the mixed-worded scales and external variables?

Generally, the three mixed-worded scales (reading, mathematics, and science self-concept) can be expected to correlate with each other and with achievement scores in the same domains. Although we did not have clear expectations about the effects of excluding inconsistent responders on these associations, we explored them because we regarded the impact on external validity measures as a relevant additional perspective.

Materials and Methods

Sample

We used data from the joint assessment of PIRLS (Progress in International Reading Literacy Study) and TIMSS in 2011, since it included test and questionnaire data from primary school students for the three domains of reading, mathematics, and science. We included all 37 education systems that participated in the joint PIRLS/TIMSS assessment. In the figures, we abbreviated the education systems with the ISO 3166 codes used in the PIRLS/TIMSS datasets.

The education systems drew representative samples of their fourth-grade student population – the exceptions were Honduras and Botswana, which sampled sixth-grade students. The studies used a multi-stage sampling procedure – the education systems first drew stratified school samples and then sampled at least one fourth-grade (or sixth-grade) class from each of these schools. For each education system, the minimum target sample size was $n = 4000$ students from a minimum of 150 schools, provided that the population was large enough. Further details on the sampling and assessment procedures

Table 1. Item Wording and Item Names of the Reading, Mathematics, and Science Self-Concept Scales.

Scale name Item wording	Wording direction	Item name
<i>Reading Self-Concept</i>		
I usually do well in reading.	+	ASBR08A
Reading is easy for me.	+	ASBR08B
Reading is harder for me than for many of my classmates.	–	ASBR08C
If a book is interesting, I don't care how hard it is to read.	+	ASBR08D
I have trouble reading stories with difficult words.	–	ASBR08E
My teacher tells me I am a good reader.	+	ASBR08F
Reading is harder for me than any other subject.	–	ASBR08G
<i>Mathematics Self-Concept</i>		
I usually do well in mathematics.	+	ASBM03A
Mathematics is harder for me than for many of my classmates.	–	ASBM03B
I am just not good at mathematics.	–	ASBM03C
I learn things quickly in mathematics.	+	ASBM03D
I am good at working out difficult mathematics problems.	+	ASBM03E
My teacher tells me I am good at mathematics.	+	ASBM03F
Mathematics is harder for me than any other subject.	–	ASBM03G
<i>Science Self-Concept</i>		
I usually do well in science.	+	ASBS06A
Science is harder for me than for many of my classmates.	–	ASBS06B
I am just not good at science	–	ASBS06C
I learn things quickly in science.	+	ASBS06D
My teacher tells me I am good at science.	+	ASBS06E
Science is harder for me than any other subject.	–	ASBS06F

Note. For each item, the response categories included 1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, and 4 = *disagree a lot*. Positively worded items are indicated by '+' and negatively worded items by '–'.

are contained in the technical documentation (Martin & Mullis, 2012), which is freely available online together with the datasets (<https://timssandpirls.bc.edu/timsspirls2011/international-database.html>).

Measures

We included three mixed-worded attitude scales (reading, mathematics, and science self-concept) as well as three achievement scores (reading, mathematics, and science achievement) in our analyses. In the joint PIRLS/TIMSS assessment, the students responded to all three self-concept scales in the student questionnaires as well as to all three achievement tests.

Each of the three self-concept scales contained three negatively worded items and three (science self-concept scale) or four (reading and mathematics self-concept scales) positively worded items. The wording in the international questionnaires is depicted in Table 1 in the original item order – these were subsequently translated. For all items, the Likert response scales ranged from 1 = *agree a lot* to 4 = *disagree a lot*. Each scale was designed to measure one domain-specific academic self-concept construct.

The paper-pencil achievement tests contained multiple-choice and constructed response items. For each achievement domain, five plausible values, computed based on test and background information using conditioning techniques, are given for each student. The achievement measures are scaled to have an international mean of 500 and standard deviation of 100.

Statistical Analysis

Inconsistent Responder Detection

For each domain and education system, we used the mean absolute difference method (cf. Hong et al., 2020; Steedle et al., 2019) to identify students whose item response pattern aligned with an inconsistent one. The mean absolute difference between the average response to the negatively worded items ('M-') and the reverse-coded average responses to the positively worded items ($5 - 'M+'$) quantifies the degree to which responders do not switch sides when marking the 4-category Likert response scale in accordance with the item wording ($|'M-' - (5 - 'M+')|$). The higher this absolute difference, the more inconsistent the response pattern.

This approach is illustrated in Figure 2. If a perfectly consistent responder on average selects 1 (*agree a lot*) for the positively worded items, the average response to the negatively worded items should be 4 (*disagree a lot*). In this case, the mean absolute difference would be equal to zero (i.e. $|4 - (5 - 1)|$). The solid main diagonal line in Figure 2 connects all pairs of average scores corresponding to perfectly consistent responders across the response scale. The further a responder orthogonally deviates from the diagonal in either direction, the less consistent the observed pair of response averages. A perfectly inconsistent responder would, for instance, tick the box for 1 (*agree a lot*) on average for both item types. In this case, the mean absolute difference is 3 (i.e. $|1 - (5 - 1)|$). In the present study, we set the threshold of the mean absolute difference for an inconsistent responder at 1.75 scale points (indicated by the dashed diagonal lines in Figure 2). Notice that the threshold is independent of the education systems and is set as a function of the common Likert scale ranging from 1 to 4. In Figure 2, the responders who were identified as inconsistent are depicted in grey. By implication, students who select 2 (*agree a little*) on average for both item types were not identified as inconsistent responders (i.e. $|2 - (5 - 2)| = 1$) and are therefore depicted in black. In further analyses, the sensitivity of the results to the specific value of the threshold was checked by applying a stricter and a more liberal threshold.

Analyses of the International Prevalence and Characteristics of Inconsistent Responders

To address the first research question on the prevalence of inconsistent responders, we computed the ratio of students who were identified as inconsistent responders for all 37 education systems and for all three mixed-worded scales. We expected inconsistent responders to be a minority in all cases. To investigate whether consistent responders outperformed inconsistent ones, we estimated Glass's delta (using the country-specific standard deviations of student achievement) as an effect size measure of the differences in mean achievement scores of the consistent versus inconsistent responders. For each education system, we conducted this comparison domain-wise (i.e. reading achievement differences between consistent and inconsistent responders to reading self-concept scale, etc.). We furthermore estimated an individual's relative risk of being identified as an inconsistent responder in either of the two other domains after they had been identified

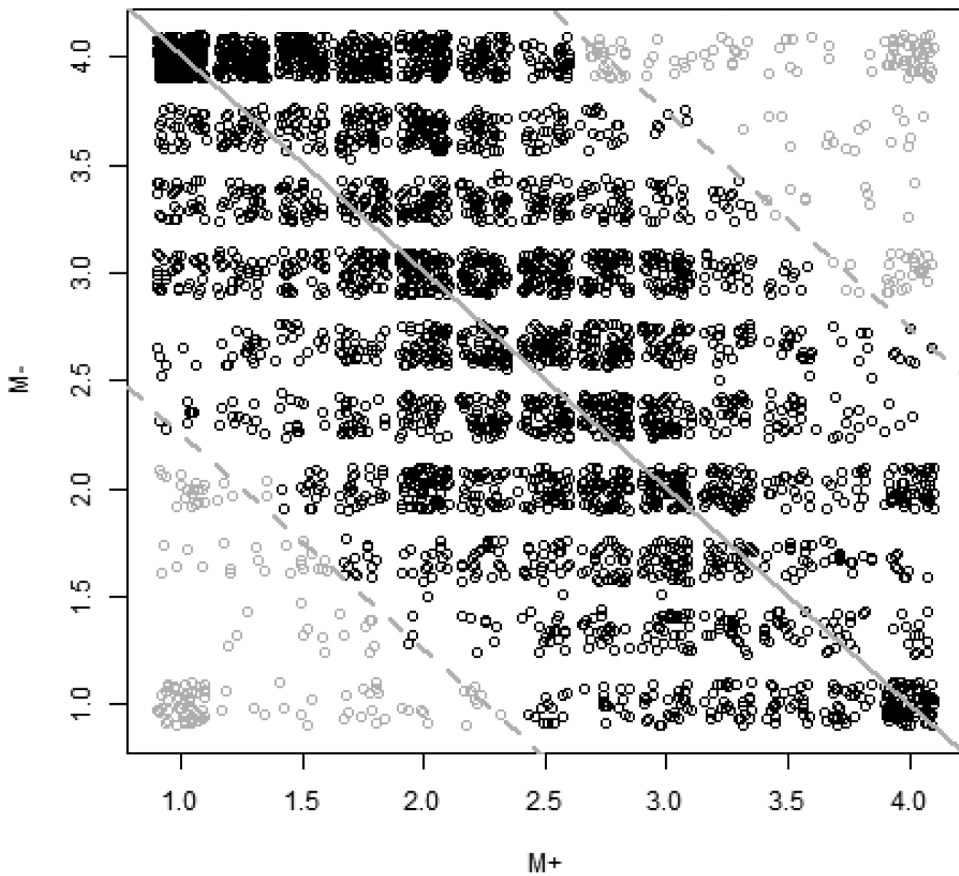


Figure 2. Illustration of the Mean Absolute Difference Approach to Identify Inconsistent Responders. *Note.* The horizontal axis represents the average response to positively worded items (M+) on the 4-category response scale. The vertical axis represents the corresponding average response on negatively worded items (M-). Every responder is represented by a black (consistent responder) or grey (inconsistent responder) dot. To avoid overplotting due to the categorical nature of the underlying item responses, the data points are jittered. The grey solid line represents the expected combinations of response averages for perfectly consistent responders. A larger orthogonal distance from this diagonal corresponds to less consistent response behaviour between positively and negatively worded items. The grey dashed lines represent the threshold value for the absolute mean threshold of 1.75 scale points. The plot reflects the example of the mathematics self-concept scale in Taiwan (Chinese Taipei).

as inconsistent in the given domain compared to when they had not been identified as such. We estimated this for each self-concept scale and education system to investigate the stability of the inconsistent response behaviour.

Analyses on the Impact of Excluding Inconsistent Responders on Mean Scale Scores

Per education system and self-concept scale, we compared the mean scores of the mixed-worded questionnaire scales when both consistent and inconsistent responders were included with the mean scores when only the consistent responders were included. The

mean scores of the questionnaire scales were computed as the average of the item responses after reverse-coding the positively worded items. This way, high values imply positive self-concepts for all items.

Analyses on the Impact of Excluding Inconsistent Responders on Scale Dimensionality and Reliability

To assess whether excluding inconsistent responders indeed resolved the observed multidimensionality issue and improved the estimated reliabilities of the scales, we compared findings when both consistent and inconsistent responders were included with findings from when the inconsistent responders were excluded. We assessed the dimensionality using the empirical Kaiser criterion (Braeken & van Assen, 2017), where the number of latent factors underlying the scale is based on a comparison of eigenvalues of the observed item correlation matrix to reference eigenvalues. Furthermore, the relative size of the first sample eigenvalue (i.e. $\lambda_1\%$) is used as a proxy for the percentage of reduction in total residual variance of the scale items due to a single common latent variable, or stated more colloquially, for ‘how unidimensional the scale is’.

We assessed the scale reliability in two ways. First, we used the correlation $r(M+,M-)$ between the average responses to positively worded items and to negatively worded items as a specific type of split-half reliability. Second, we took Cronbach’s alpha as an average across all possible split-half reliabilities. We expected the first reliability measure to be the most sensitive to the presence of inconsistent responders as it directly mapped onto the positive vs. negative wording contrast and ignored the within-wording consistency. The within-wording consistency made up a large part of the Cronbach’s alpha, which we thus expected to be more robust to the presence of inconsistent responders.

Analyses on the Impact of Excluding Inconsistent Responders on External Validity Measures

We explored changes in external validity measures when including versus excluding inconsistent responders in two ways. On the one hand, we expected the three mixed-worded scales to be moderately positively correlated because they all assessed academic self-concepts. On the other hand, we expected the three scales to correlate moderately positively with achievement scores in the same domains (i.e. we expected reading self-concept to be associated with reading achievement, etc.). In these analyses, we used mean scale scores for each student and domain. When computing these mean scores, we first inversely recoded the responses to the positively worded items so that high values imply positive self-concepts for all items. We computed both the between-domain intercorrelations among the self-concept scales and the within-domain correlations between self-concept and achievement scores.

Implementation Details

We ran all statistical analyses in R version 4.03 and took care to properly handle sampling weights and plausible values as advised for large-scale assessment data (Rutkowski et al., 2010). We provide a dataset with all relevant estimates by education system, a code book,

and the R script to replicate the analyses as supplemental material online. We briefly report on the robustness of results when we set a more conservative or liberal threshold for identifying the inconsistent responders.

Results

International Prevalence and Characteristics of Inconsistent Responders

The prevalence of inconsistent responders per domain across education systems is summarised in Figure 3. The average prevalence across all education systems and domains was about 9%. In general, there appeared to be more variation across education systems than between the three self-concept scales. Across education systems, the average percentages of inconsistent responders were quite similar – they were 10%, 8%, and 8% for the reading, mathematics, and science self-concept scales, respectively. For most

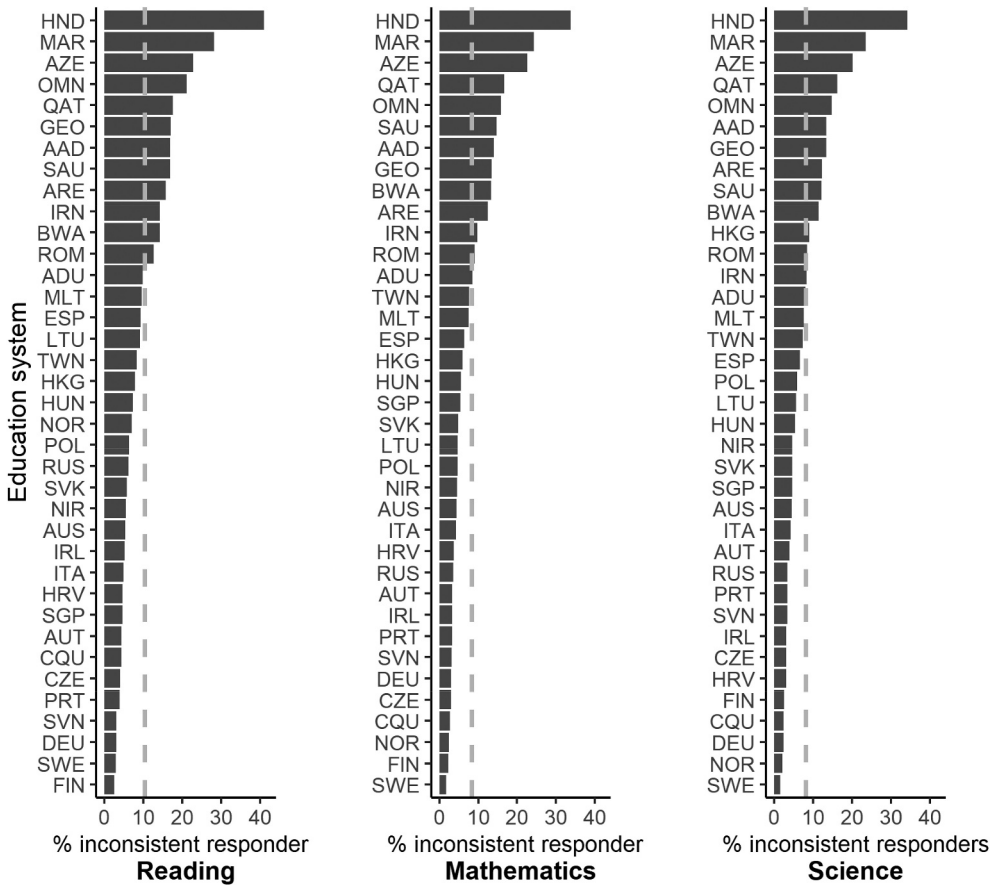


Figure 3. Prevalence of Inconsistent Responders to the Reading, Mathematics, and Science Self-Concept Scales across Education Systems. Note. The figure displays the shares of inconsistent responders to the reading (left), mathematics (centre), and science (right) self-concept scales on the horizontal axes. The vertical dashed lines depict the average shares of inconsistent responders across education systems per domain.

education systems, the proportions of inconsistent responders were quite similar across the domains, although the share of inconsistent responders to the reading self-concept scale was considerably higher than in the other two domains in some education systems. Sweden, Finland, Norway, Germany, Canada (Quebec), Czech Republic, and Slovenia had generally lower percentages of inconsistent responders, while Honduras, Morocco, Azerbaijan, Qatar, Oman, and Georgia had higher percentages. Across domains, Sweden had the lowest average share of inconsistent responders (2%) and Honduras had the highest (36%).

Figure 4 depicts the distributions of consistent and inconsistent responders' mean scale scores across domains and education systems. The means of consistent responders were closer to the high self-concept end of the scale, while the inconsistent responders scored closer to the midpoints of the 4-category Likert scales, which reflected their inconsistent response behaviour. As expected, the inconsistent responders' scores also showed a lower standard deviation than those of the consistent responders.

We expected the consistent responders to outperform the inconsistent responders in the achievement tests and found support for this hypothesis, since the inconsistent responders scored almost one standard deviation below the other students' average achievement in reading (Glass's delta: $M = .91$, range = $\{.19, 1.41\}$), mathematics (Glass's delta: $M = .87$, range = $\{.27, 1.21\}$), and science (Glass's delta: $M = .90$, range = $\{.30, 1.29\}$). The education systems with noticeably smaller achievement differences between consistent and inconsistent responders tended to be those systems with higher prevalences of inconsistent responders.

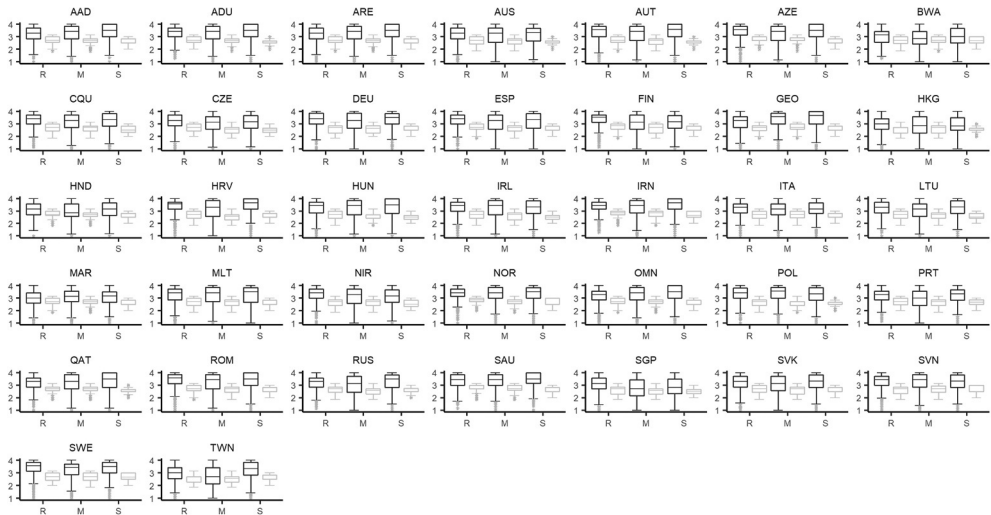


Figure 4. Distributions of Consistent and Inconsistent Responders' Scores on Mean Scales in Three Domains and Across Education Systems. *Note.* The three single letters on the horizontal axes stand for the self-concept mean scale scores in the three domains of reading, mathematics, and science, respectively. Black boxplots represent the consistent responders and grey boxplots show the inconsistent responders.

In the same vein, we expected an individual who responded inconsistently to one mixed-worded scale to also be more likely to respond inconsistently to other mixed-worded scales in the same questionnaire. This hypothesis was supported, as the risk of being identified as an inconsistent responder in either of the other domains was, on average across education systems, 6 to 7 times higher for those who were identified as inconsistent compared to those who were not in that given domain. This was similar for reading (relative risk: $M = 6.18$, range = {2.58, 12.27}), mathematics (relative risk: $M = 7.06$, range = {2.38, 16.52}), and science (relative risk: $M = 7.07$, range = {2.39, 16.26}).

Impact of Excluding Inconsistent Responders on Mean Scale Scores

We expected that removing the inconsistent responders from the samples should increase the mean scores on the mixed-worded scales, because the inconsistent responders score more at the middle of the response scales (see Figure 4) and because the consistent responders should report rather positive academic self-concepts, on average. Figure 5 depicts the mean scale scores when including and excluding the inconsistent responders. In the samples that included inconsistent responders, the mean scale scores expressed medium positive academic self-concepts in reading (mean scale score: $M = 3.20$, range = {2.94, 3.39}), mathematics (mean scale score: $M = 3.10$, range = {2.75, 3.30}), and science (mean scale score: $M = 3.20$, range = {2.85, 3.44}). When removing the inconsistent responders, they were higher in the reading (mean scale score: $M = 3.25$, range = {3.00, 3.41}), mathematics (mean scale score: $M = 3.14$, range = {2.77, 3.34}), and science (mean scale score: $M = 3.26$, range = {2.86, 3.50}) self-concept scales. Removing the inconsistent responders increased the mean scores by about 0.04 points on average, expressing more positive self-concepts in reading (change in mean scale score: $M = .05$,

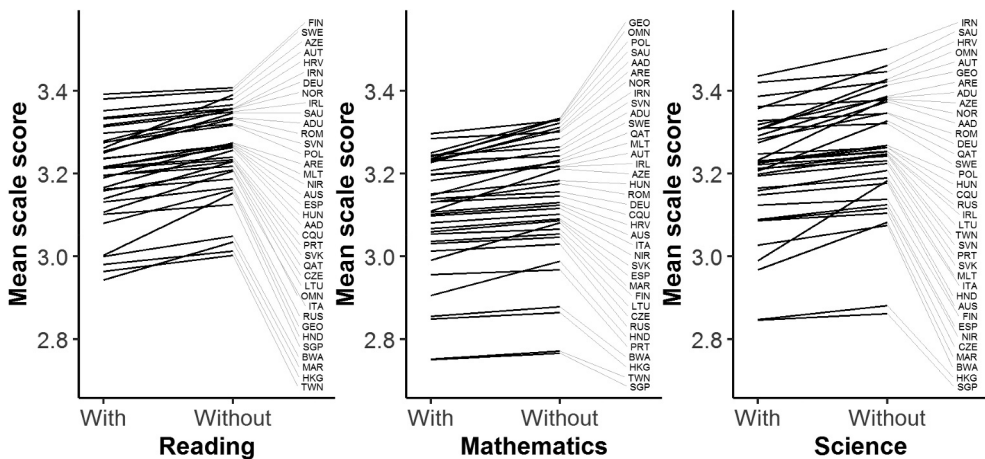


Figure 5. Mean Scores of the Three Self-Concept Scales with and without Inclusion of Inconsistent Responders. *Note.* The figure displays the mean scores on the reading (left), mathematics (centre), and science (right) self-concept scales when including (with) or excluding (without) inconsistent responders for all education systems. The mean scores are computed as the mean of item responses after reverse-coding the responses to the positively worded items. High values imply positive academic self-concepts.

range = {.02, .15}), mathematics (change in *mean scale score*: $M = .04$, range = {.01, .10}), and science (change in *mean scale score*: $M = .05$, range = {.01, .19}). The changes in the average scores seem modest, but it has to be taken into account that the lowest and highest mean scale scores across education systems were only about 0.50 points apart. We observed larger increases in the education systems with higher prevalences of inconsistent responders (e.g. Azerbaijan, Georgia, Morocco, Oman, Qatar, Saudi Arabia, and Honduras).

Impact of Excluding Inconsistent Responders on Scale Dimensionality and Reliability

Dimensionality

Since we expected inconsistent responders to introduce wording-related covariance that is independent of the substantive self-concept constructs, we believed that the factor analyses would suggest more complex latent structures than the intended one-dimensional structures. We expected the one-dimensional models to represent the data well after removing these inconsistent responders. Consistent with this, the empirical Kaiser criterion suggested two latent underlying factors in 96 out of 111 samples that included inconsistent responders. In the remaining samples, it suggested the intended one-dimensional structure (12 education systems in mathematics {Australia, Austria, Croatia, Czech Republic, Finland, Germany, Ireland, Norway, Portugal, Russia, Sweden, and Canada (Quebec)} and 3 in the science self-concept scale {Czech Republic, Germany, Russia}). These 15 education systems had all less than 5% inconsistent responders. Conforming to expectations, the empirical Kaiser criterion supported the intended one-dimensionality after removing the inconsistent responders in 104 of 111 cases. In the remaining seven cases, it suggested two factors. These all stemmed from the reading self-concept scale and contained mostly education systems with a relatively high prevalence of inconsistent responders (Honduras, Morocco, Qatar, Botswana), but also more moderate cases (Hong Kong SAR, Northern Ireland, Russia).

To quantify the size of this bias to dimensionality in more continuous terms, we also computed the relative size of the first sample eigenvalue (i.e. $\lambda_1\%$). In the samples that included inconsistent responders, this index was about 40% on average in the reading ($\lambda_1\%$: $M = 37\%$, range = {25%, 46%}), mathematics ($\lambda_1\%$: $M = 49\%$, range = {28%, 62%}), and science ($\lambda_1\%$: $M = 48\%$, range = {29%, 58%}) self-concept scales. After removing the inconsistent responders, the index increased in the cases of reading ($\lambda_1\%$: $M = 40\%$, range = {29%, 49%}), mathematics ($\lambda_1\%$: $M = 54\%$, range = {33%, 65%}), and science ($\lambda_1\%$: $M = 52\%$, range = {34%, 63%}) self-concept scales. On average, this increase was about 4 percentage points in reading (change in $\lambda_1\%$: $M = 3\%$, range = {1%, 5%}), mathematics (change in $\lambda_1\%$: $M = 4\%$, range = {2%, 9%}), and science (change in $\lambda_1\%$: $M = 4\%$, range = {2%, 9%}), with extremes up to 9 percentage points difference (e.g. Honduras in mathematics self-concept scale; Taiwan and Hong Kong in science self-concept scale).

Reliability

We expected to find higher scale reliability estimates after removing inconsistent responders because of the additional construct-unrelated covariance they introduced. The first reliability measure, the correlation $r(M+,M-)$ between average responses to positively and negatively worded items specifically targeted the internal consistency between responses to the two item types. In the samples that included inconsistent responders, this correlation was close to zero in some education systems in the reading ($r(M+,M-)$: $M = -.30$, range = $\{-.50, .07\}$), mathematics ($r(M+,M-)$: $M = -.46$, range = $\{-.66, -.03\}$), and science ($r(M+,M-)$: $M = -.38$, range = $\{-.58, .02\}$) self-concept scales. Such zero correlations are incompatible with the expectation that positively and negatively worded items measure the same latent constructs in an opposite way, which would instead imply large negative correlations. As expected, the average responses to positively and negatively worded items were more negatively correlated in education systems with fewer inconsistent responders. As depicted in the upper half of [Figure 6](#), the exclusion of inconsistent responders made the correlation $r(M+,M-)$ more negative in the reading (change in $r(M+,M-)$: $M = -.22$, range = $\{-.49, -.08\}$), mathematics (change in $r(M+,M-)$: $M = -.21$, range = $\{-.51, -.07\}$), and science (change in $r(M+,M-)$: $M = -.22$, range = $\{-.48, -.09\}$) self-concept scales.

When we included inconsistent responders, we found quite low Cronbach's alpha results in some education systems in the reading ($M = .66$, range = $\{.41, .77\}$), mathematics ($M = .81$, range = $\{.55, .89\}$), and science ($M = .76$, range = $\{.49, .85\}$) self-concept scales. As depicted in the lower half of [Figure 6](#), after removing the inconsistent responders, we noted slight increases in Cronbach's alpha in the reading (change in Cronbach's alpha: $M = .04$, range = $\{.01, .12\}$), mathematics (change in Cronbach's alpha: $M = .03$, range = $\{.01, .13\}$), and science (change in Cronbach's alpha: $M = .04$, range = $\{.01, .10\}$) self-concept scales. Note how generally, the reliability of the reading self-concept scale remained comparably low even after excluding inconsistent responders. Regarding the mathematics and science self-concept scales, our finding of Cronbach's alpha measures below .8 after removing inconsistent responders coincided with higher prevalences of inconsistent responders (i.e. about 9% and higher).

Impact of Excluding Inconsistent Responders on External Validity Measures

To investigate whether removing inconsistent responders affected external validity measures, we compared the correlations between the three self-concept scales when including and excluding inconsistent responders. There were great differences between education systems in the observed intercorrelations among the three self-concept scales when we included inconsistent responders ($r(\text{reading,mathematics})$: range = $\{.16, .55\}$; $r(\text{reading,science})$: range = $\{.29, .61\}$; $r(\text{mathematics,science})$: range = $\{.13, .61\}$). However, these correlations remained rather stable when we excluded inconsistent responders (change in $r(\text{reading,mathematics})$: $M = .00$; change in $r(\text{reading,science})$: $M = .00$; change in $r(\text{mathematics,science})$: $M = .01$). As [Figure 7](#) shows, even the most pronounced changes were still below a change in correlation of .1 (absolute change in $r(\text{reading,mathematics})$: $\text{Max} = .04$; absolute change in $r(\text{reading,science})$: $\text{Max} = .04$; absolute change in $r(\text{mathematics,science})$: $\text{Max} = .05$). This might be because the inconsistent responders scored around the scale midpoints (see [Figure 4](#)) and therefore did not have the same pull

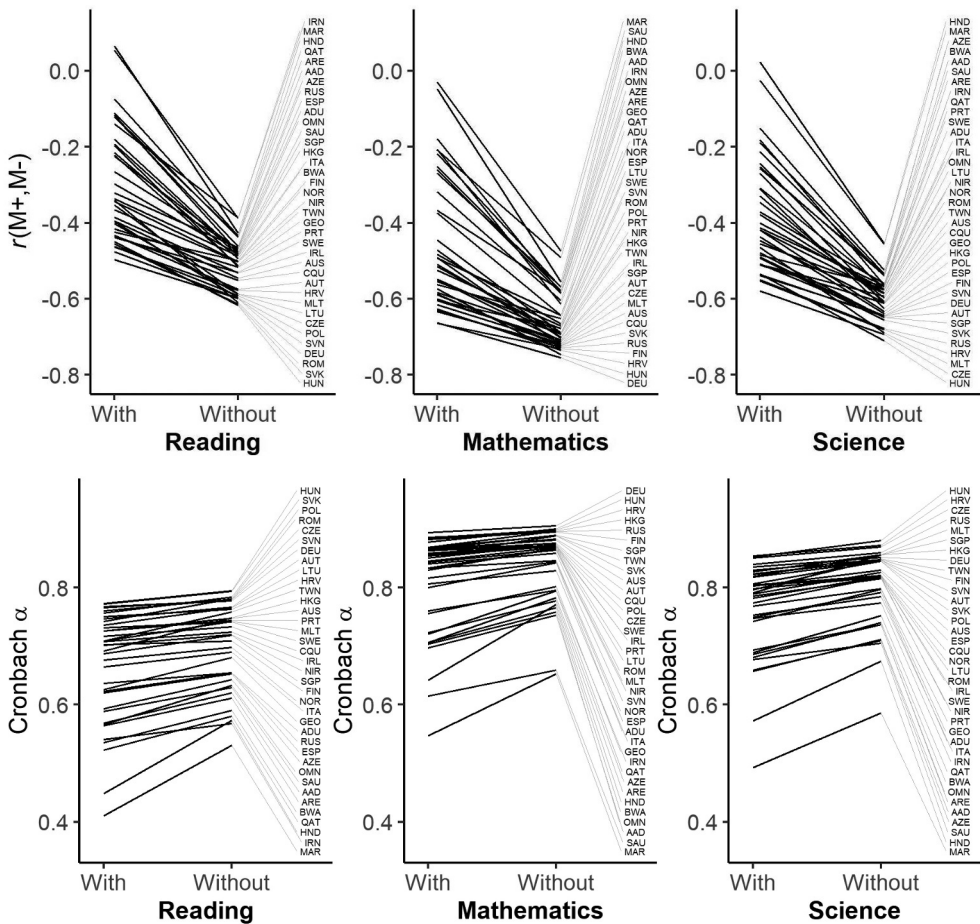


Figure 6. Reliability of the Three Self-Concept Scales with and without Inclusion of Inconsistent Responders. *Note.* The figure displays the correlation between average responses to positively and negatively worded items (upper half) and Cronbach's alpha (lower half) reliability measures for all education systems and the reading (left), mathematics (centre), and science (right) self-concept scales when including (with) or excluding (without) inconsistent responders.

on the observed correlation measures. Some impact could be expected in education systems in which almost no students report a low self-concept, in which inconsistent responding in one domain would be strongly related to self-concept in another domain, or in which the prevalence of inconsistent responders was high. The latter was the case for the education systems that showed the most pronounced changes in correlation (i.e. Honduras, Honduras, and United Arab Emirates (Abu Dhabi) in the reading, mathematics, and science scales, respectively).

A second analysis of changes in external validity measures concerned the correlations between the self-concept scales and achievement scores in the same domain. Across education systems and when including inconsistent responders, these correlations were small to moderate in reading (range = {.30, .53}), mathematics (range = {.17, .45}), and science (range = {.17, .40}). As shown in [Figure 8](#), these correlations tended to attenuate

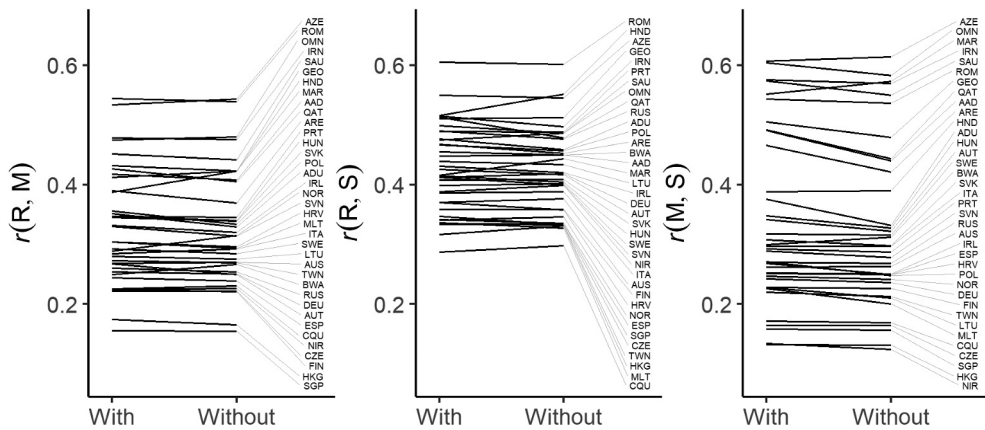


Figure 7. Intercorrelations among the Three Self-Concept Scales with and without Inclusion of the Inconsistent Responders. *Note.* The figure displays the inter-scale correlations between the domains of reading and mathematics (left), reading and science (centre), and mathematics and science (right) for all education systems when including (with) or excluding (without) inconsistent responders.

somewhat when excluding the inconsistent responders in the domains of reading (change in $r(\text{self-concept, achievement})$: $M = -.02$), mathematics (change in $r(\text{self-concept, achievement})$: $M = -.02$), and science (change in $r(\text{self-concept, achievement})$: $M = -.03$). The maximal changes for specific education systems and domains were still below a change in correlation of .1 in the domains reading (change in $r(\text{self-concept, achievement})$: $\text{Max} = -.06$), mathematics (change in $r(\text{self-concept, achievement})$: $\text{Max} = -.05$), and science (change in $r(\text{self-concept, achievement})$: $\text{Max} = -.08$). The education systems with slightly more pronounced changes were mostly, though not

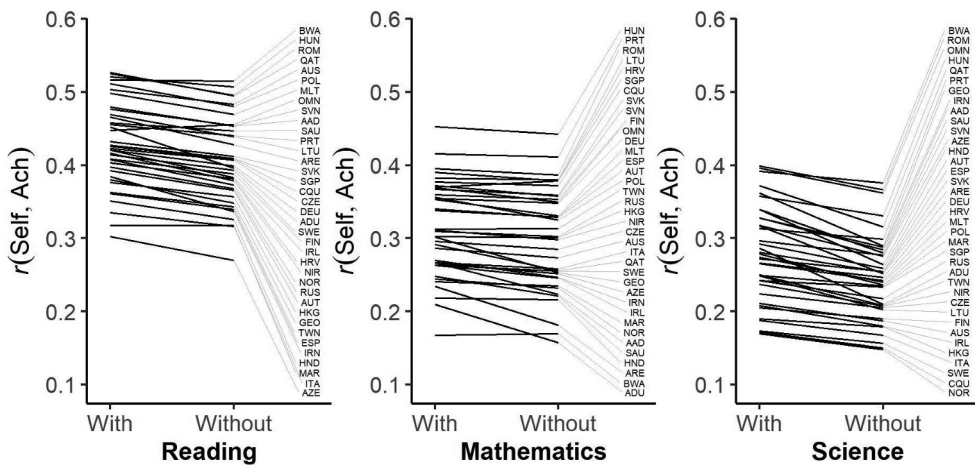


Figure 8. Correlation between Self-Concept and Achievement in Three Self-Concept Scales with and without Inclusion of the Inconsistent Responders. *Note.* The figure displays the correlations between self-concept and achievement in the domains of reading (left), mathematics (centre), and science (right) for all education systems when including (with) or excluding (without) inconsistent responders.

always, the ones with higher prevalences of inconsistent responders. All in all, we conclude that the scale correlations were rather robust to the presence of inconsistent responders because they scored at the midpoint of the scales (see [Figure 4](#)).

Sensitivity of Results to the Exact Threshold Value for Identifying Inconsistent Responders

In the main analyses, we applied a threshold value of 1.75 on the mean absolute difference to identify inconsistent responders. This was set as a function of the Likert scale that ranged from 1 to 4. As a sensitivity check, we replicated the analyses with a more liberal (1.50) and a more strict (2.00) threshold value. The education systems that already had bigger shares of inconsistent responders were more responsive to shifts in the threshold. The maximum difference in the percent of inconsistent responders was a 6% reduction when using the liberal value and an 8% increase when applying the stricter threshold (both times in the case of Honduras). Education systems with lower shares remained fairly stable, both in terms of the shares of inconsistent responders (i.e. in the range of a 1–3% absolute difference) and in the considered outcome measures. Dimensionality and reliability outcomes were fairly robust, although the specific correlation between the average scores on negatively and positively worded items was somewhat more sensitive. External validity outcomes remained robust throughout (i.e. absolute differences in correlation did not exceed .05). This expected pattern of results supports the relative robustness of the findings.

Discussion

The present study addressed the prevalence and impact of inconsistent responders to mixed-worded questionnaire scales. First, we investigated the prevalence of inconsistent responders internationally and what characteristics they had. Looking at all three mixed-worded scales in all 37 education systems, we found that a minority of primary school students did not change the side of the response scale according to positive and negative item wording and were therefore identified as inconsistent responders. This, as well as the finding that these inconsistent responders had lower average achievement scores than other students conformed with our expectations. Furthermore, students who responded inconsistently to one mixed-worded scale had a much higher risk of responding inconsistently to either of the other two scales. We interpret these findings as internal validation for our assumption that inconsistent responding is associated with personal characteristics. Specifically, we assumed that responders who lack the necessary reading or cognitive skills or the necessary care when filling out questionnaires might not notice the mixed item wording or fail to adjust their responses accordingly (see also Marsh, 1986; Schmitt & Stults, 1985; Steinmann et al., 2021; Swain et al., 2008).

Second, we investigated whether excluding inconsistent responders affected the average mean scores on the mixed-worded questionnaire scales. Conform to expectations, we found that the inconsistent responders scored more in the middle of the response scale, while consistent responders scored higher, expressing more positive academic self-concepts. Therefore, removing the inconsistent responders shifted the mean scale scores, especially in education systems with large shares of inconsistent responders, to the extent

that the relative position of the mean scale scores of the education systems changed (see [Figure 5](#)). This suggests that since the education systems vary in the shares of inconsistent responders, and since the inconsistent responders score more at the middle of the response scales, country comparisons can be biased. This type of country comparison problem even adds to other potential measurement invariance issues in the international comparative studies.

Third, we studied whether excluding inconsistent responders affects the estimated dimensionality and reliability of mixed-worded questionnaire scales. In most cases, the empirical Kaiser criterion suggested that the latent structure of the mixed-worded scales would be more than one-dimensional in the full samples (cf. bias of about 4%–9% in the first eigenvalue). In the samples without inconsistent responders, the empirical Kaiser criterion usually suggested a simple one-dimensional structure as adequate to represent the data. This finding conformed to previous research (Schmitt & Stults, 1985; Steedle et al., 2019; Woods, 2006) and to our expectations. It also explains why previous research usually found data structures that were more complex than intended in mixed-worded questionnaire scales (e.g. DiStefano & Motl, 2009; Gnams & Schroeders, 2020; Lindwall et al., 2012; Marsh et al., 2013; Quilty et al., 2006). In the same vein, we found that removing the inconsistent responders systematically improved Cronbach's alpha and the estimated correlation between average responses to the positively and negatively worded items. This specific type of split-half reliability measure directly reflects consistent responding to the two wording types and is therefore more sensitive to excluding inconsistent responders, while Cronbach's alpha also reflects the consistency in responding to items with the same wording. However, these findings might explain why the reliabilities of mixed-worded questionnaire scales are often rather low and variable between countries (e.g. Barnette, 2000; Marsh et al., 2013; Martin & Mullis, 2012).

Fourth, we investigated how excluding inconsistent responders affected external validity measures. Between analyses with and without inconsistent responders, we found no systematic group-level difference in the association between the mixed-worded scales with each other, nor between the mixed-worded scales and achievement scores in the same domain. This might be because, by implication, inconsistent responders score in the middle of simple mean scale scores of mixed-worded scales that contain about as many positively worded items as negatively worded ones. This implies that they do not affect correlations between the mean scale scores and external variables much to begin with. This might also explain why previous research likewise found inconclusive effects (Hong et al., 2020; Steedle et al., 2019).

Further interesting findings concern the large differences between education systems and scales. Between education systems, the rate of inconsistent responders ranged from below 5% to more than one third of students. It is possible that inconsistent responding is more common in countries where students have lower achievement levels overall, where fewer students are native speakers of the language of the questionnaire, where the mixed-wording is more difficult to handle due to language characteristics, or where using negatively worded statements and double negations is less common. Although beyond the scope of the present study, this appears to be an interesting area for future research. Furthermore, we observed relatively similar prevalences of inconsistent responders across the three scales within education systems. In some education systems, however, there were more inconsistent responders to the reading self-concept scale than to the

other two scales. It was again beyond the scope of this study to investigate these differences further. However, it is important to recall that the three self-concept scales were not equivalent across the subjects (see [Table 1](#)). The reading self-concept scale stemmed from the PIRLS part of the assessment, which was developed and adapted to the languages independently from the mathematics and science self-concept scales from the TIMSS assessment. Further, the reliability of the reading self-concept scale was generally lower than that of the other two self-concept scales.

Limitations

Although our analyses drew on representative student samples from 37 education systems and three independent mixed-worded questionnaire scales, we would like to stress some limitations to the generalisability of our findings. The three mixed-worded scales all addressed self-attitudes, contained an almost balanced share of three or four positively worded items and three negatively worded ones, and started with a positively worded item. All scales and items contained the same four response categories that started with *agree a lot* on the left side of the response scale. All responders were primary school students and therefore beginning readers. The PIRLS/TIMSS assessment was a low-stakes survey in all participating education systems. We therefore cannot generalise our findings to other scales, responder groups, or high-stakes contexts. There are some indications in the literature that inconsistent responders can also be found in scales on other constructs (e.g. [Hong et al., 2020](#); [Melnick & Gable, 1990](#); [Steedle et al., 2019](#)). However, the phenomenon might be less common among older responders (e.g. [Bolt et al., 2020](#); [Steinmann et al., 2021](#)) and under high-stakes conditions (cf. [Huang et al., 2012](#)), for instance. Furthermore, we used the mean absolute difference method to identify inconsistent responders. Different identification methods are available (e.g. [Bolt et al., 2020](#); [Hong et al., 2020](#); [Steinmann et al., 2021](#)) and might lead to different identification results.

Conclusion

While the use of mixed wording in questionnaire scales is intended to retain responders' attention and elicit more thorough responses, our study showed that some primary school students failed to shift from one side of the response scale to the other in accordance with the wording. This inconsistent response behaviour appears to be related to student characteristics such as achievement measures and to vary between education systems. These inconsistent responders hence have invalid scores on the mixed-worded scales. Furthermore, we found that they biased dimensionality and reliability analyses. Due to the invalid individual scores, inconsistent responding furthermore risks confounding the associations between the mixed-worded scales and external variables. In our study, however, we found no systematic empirical support for such effects at the group level.

We therefore concluded that the mixed wording does not accomplish the goal of improving the psychometric properties of questionnaire scales (e.g. [Idaszak & Drasgow, 1987](#); [Podsakoff et al., 2003](#)), at least for young responder groups. We would therefore recommend that questionnaire developers who do not plan to conduct research on the

inconsistent responder phenomenon use scales with only one type of item wording, especially for the primary school sector (see also Dunbar et al., 2000; Lindwall et al., 2012; Marsh, 1996; Melnick & Gable, 1990; Weems et al., 2003). This conforms to the guiding principle that questionnaires should be as easy to fill out as possible to avoid a confounding of responses with reading or cognitive abilities. If researchers nevertheless decide to use data from mixed-worded questionnaire scales, we recommend removing inconsistent responders before evaluating the scales' dimensionality and reliability and before conducting inferential analyses (cf., Patton et al., 2019; Schmitt & Stults, 1985; Woods, 2006).

Author Notes

Acknowledgements of support: We thank Roisin Cronin for copy-editing the manuscript.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Norges Forskningsrådet [FRIPRO-HUMSAM261769].

Notes on contributors

Isa Steinmann is a postdoctoral researcher at the Centre for Educational Measurement at the University of Oslo (CEMO) in Norway and at the Center for Research on Education and School Development (IFS) at TU Dortmund University in Germany. In her research, she mostly uses national and international large-scale assessment data for secondary analysis to determine how education systems and schools affect student achievement and educational inequalities. Another strand of her research investigates how properties of the international large-scale assessments affect their results and interact with the respondents.

Daniel Sánchez is a Research Assistant at the Centre for Educational Measurement at the University of Oslo (CEMO) in Norway. His research focuses on substantive-methodological synergisms in the broad area of international large-scale assessments. He holds a Master of Science in Measurement, Assessment, and Evaluation from the University of Oslo.

Saskia van Laar received her master's degree in Methodology and Statistics in Psychology from Leiden University, the Netherlands. She is currently a doctoral research fellow at the Centre for Educational Measurement at the University of Oslo, Norway. Her current work focuses on validity and response behaviour in international large-scale educational 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 measurement-related methodological issues in large-scale assessments. This includes among others computerised adaptive testing and response validity of survey responses.

ORCID

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

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

References

- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361–370. <https://doi.org/10.1177/00131640021970592>
- 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>
- Braeken, J., & van Assen, M. A. L. M. (2017). An empirical Kaiser criterion. *Psychological Methods*, 22(3), 450–466. <https://doi.org/10.1037/met0000074>
- DiStefano, C., & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(1), 134–146. <https://doi.org/10.1080/10705510802565403>
- Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). Question wording effects in the assessment of global self-esteem. *European Journal of Psychological Assessment*, 16(1), 13–19. <https://doi.org/10.1027//1015-5759.16.1.13>
- 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, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Idaszak, J. R., & Drasgow, F. (1987). A revision of the job diagnostic survey: Elimination of a measurement artifact. *Journal of Applied Psychology*, 72(1), 69–74. <https://doi.org/10.1037/0021-9010.72.1.69>
- Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, 18(3), 512–541. <https://doi.org/10.1177/1094428115571894>
- 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>
- 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>
- Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22(1), 37–49. <https://doi.org/10.1037/0012-1649.22.1.37>
- 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>
- Martin, M. O., & Mullis, I. V. S. (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College.
- Melnick, S. A., & Gable, R. K. (1990). The use of negative item stems: A cautionary note. *Educational Research Quarterly*, 14(3), 31–36. <https://psycnet.apa.org/record/1991-34765-001>

- 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>
- 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. *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
- Rutkowski, L., Gonzalez, E., Joncas, M., & Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>
- 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>
- 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 and Practice*, 38(2), 101–111. <https://doi.org/10.1111/emip.12256>
- Steinmann, I., Strietholt, R., & Braeken, J. (2021). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods* Advance Online Publication. <http://dx.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>
- Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education*, 28(6), 587–606. <https://doi.org/10.1080/0260293032000130234>
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 189–194. <https://doi.org/10.1007/s10862-005-9004-7>