Chen *et al.*
Large-scale Assessments in Education (2023) 11:37
https://doi.org/10.1186/s40536-023-00184-6

Large-scale Assessments in Education

RESEARCH Open Access

Who are those random responders on your survey? The case of the TIMSS 2015 student questionnaire

Jianan Chen[1], Saskia van Laar[1,2*] and Johan Braeken[1]

*Correspondence:
saskia.
vanlaar@maastrichtuniversity.nl

[1] CEMO: Centre for Educational Measurement at the university of Oslo, Faculty of Educational Sciences, University of Oslo, Oslo, Norway
[2] Department of Educational Development and Research, School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract

A general validity and survey quality concern with student questionnaires under low-stakes assessment conditions is that some responders will not genuinely engage with the questionnaire, often with more random response patterns as a result. Using a mixture IRT approach and a meta-analytic lens across 22 educational systems participating in TIMSS 2015, we investigated whether the prevalence of random responders on six scales regarding students' engagement and attitudes toward mathematics and sciences was a function of grade, gender, socio-economic status, spoken language at home, or migration background. Among these common policy-relevant covariates in educational research, we found support for small group differences in prevalence of random responders (OR $\geq$ 1.22) (average prevalence of 7%). In general, being a student in grade 8 (vs. grade 4), being male, reporting to have fewer books, or speaking a language different from the test language at home were all risk factors characterizing random responders. The expected generalization and implications of these findings are discussed based on the observed heterogeneity across educational systems and consistency across questionnaire scales.

Keywords: Rrandom responders, International large-scale educational assessment, Mixture IRT, TIMSS

Introduction

International large-scale assessments in education (ILSAE), such as IEA's Trends in International Mathematics and Science Study (TIMSS) or OECD's Programme for International Student Assessment (PISA), can provide input on current policy-relevant research questions with respect to inequality and inequity (e.g., Hopfenbeck et al., 2018). ILSAE tend to consist of both an achievement test component and a questionnaire component. The collected data allows for educational research that assesses potential differences in achievement and/or attitudes between, for instance, students of differing gender, socio-economic status, or migration background (e.g., Hopfenbeck et al., 2018), often in combination with a search for protective or risk factors with respect to such differences by comparing classroom practices and other contextual factors. In this way, ILSAE can help shape educational policy by clarifying standards and providing a wide basis of reference comparisons for

education systems, informing curriculum reforms, identifying investment targets based on poor performance in certain subject domains or by specific groups, and guiding resource allocation for optimization of classroom practices and teacher training (for a review, see e.g., Hernández-Torrano and Courtney, 2021).

Although a potential treasure trove, ILSAE have some inherent limitations such as providing less fine-grained learning achievement outcomes than the regular system of school exams (Clarke & Luna-Bazaldua, 2021) and relying on self-report measures for many relevant contextual factors or background variables (e.g., Hopfenbeck & Maul 2011; Rutkowski & Rutkowski 2010), and all this in a low-stakes assessment context (e.g., Eklöf, 2010). There is no immediate feedback nor negative or positive consequences for the students participating in the ILSAE. Hence, data quality and validity issues are of concern for everyone involved in these huge projects. A general concern is that not all students are providing genuine responses and that this might distort results to the extent that it could lead to misguided conclusions and educational policy recommendations. Random responding by students on questionnaire scales of the survey is one type of invalid response behavior that comes across as especially threatening or harmful. Random responding is described as providing "responses without meaningful reference to the test questions" (Berry et al., 1992, p. 340) often ascribed to among others insufficient effort, carelessness, thoughtlessness, disengagement, or lack of seriousness and motivation on the part of the person responding to the survey (e.g., Huang, Curran, Keeney, Poposki, & DeShon, 2012). Hence, it is rather intuitive to understand the validity concerns (e.g., Cronbach, 1950; Messick, 1984) that having *random responders* on your survey would raise.

Although observable responses are still provided by the person, a random responder can be seen as causing a form of *nonresponse error*, because we end up lacking accurate data on the genuine attitude or information the person is surveyed about. Hence, as with nonresponse rates (e.g., Cochran, 1951, Bethlehem, 2009), low prevalence of random responders in the sample can be regarded as a quality indicator of both survey and corresponding survey data, whereas a high prevalence makes the quality of survey results open for critical debate. Similar to more traditional nonresponse (e.g., Groves & Peytcheva 2008; Hedlin, 2020), the biasing impact will not only depend on the prevalence but also on the underlying mechanism as commonly framed in terms of Rubin's (1976) framework of missing completely at random (MCAR), at random (MAR), or not at random (MNAR). Hence, it might be useful to think in similar terms about random responders when considering their potential impact. If minority groups or groups with other specific characteristics have a higher prevalence of random responders, such systematic disproportionate differences can lead to selective fallout in the sample, and if the propensity of engaging in random responding relates to the survey outcomes of interest, this can potentially skew, bias, and invalidate any inferences/conclusions based on the questionnaire scales (for a similar point on nonresponse, see e.g., Richiardi, Pizzi, & Pearce, 2013).

This study

In this study, we performed an initial exploration of this validity issue for survey scales inquiring about students' engagement and attitudes toward mathematics and science in the TIMSS 2015 assessment (Martin et al., 2016). We conducted a study across 22 participating educational systems, comparing whether student groups—defined in terms of

research- and policy-relevant covariate information on grade, gender, socio-economic status, spoken language at home, and migration background—differed in their odds of having been classified as a random responder on six TIMSS student questionnaire scales about students' engagement and attitudes toward Mathematics and Science. Findings will inform about the potential differential prevalence of random responders among the student groups.

Identifying random responders

Detection methods for random responding are either based on auxiliary information at the item level such as item response times or are based on the actual item response pattern across a questionnaire scale. The response-time approach leads to an operationalization in terms of so-called 'rapid guessing', where an item response is given in too little time for the person to have actively processed the actual survey question (Wise, 2017). Although very fine-grained, this approach requires the availability and precise measurement of response time at the item level, as well as the setting of a reasonable threshold for when a response is considered 'too fast'. For surveys where items on questionnaire scales are not presented one at a time, such auxiliary item-level information is not obvious to obtain (in contrast to achievement tests where it is more typical to show one problem at a time). The item response pattern approach requires methods to quantify unexpected variability across responses compared to a typical consistent pattern of responses across the questionnaire scale (e.g., Curran, 2016). This makes the approach less suitable for questionnaire scales that are not targeting a reflective construct (as compared to a more formative construct such as socio-economic status) and not feasible for single items (due to a lack of related items as a comparison base).

In absence of useful auxiliary information at the item level, we conducted scale level detection of random responding following a mixture item response theory (IRT) approach. More specifically, we used an extension of the HYBRID model by Yamamoto (1989) to the polytomous case for survey responses as proposed by van Laar and Braeken (2022). Hence, every student was classified as a random responder or a typical responder on the questionnaire scales under investigation.

Survey scales

Among the survey scales present in TIMSS 2015, we focused on those related to students' engagement and attitudes toward mathematics and science. This is an active and relevant area of research in education where there is a general worry about the decline in positive attitudes and beliefs with increasing age and grade or educational level (Potvin & Hasni, 2014). How these attitudes and beliefs relate to educational achievement varies on what exactly is surveyed. Students' confidence in mathematics or science tends to be positively related to achievement in the corresponding subject (Wigfield & Eccles, 2002), whereas achievement's relation with valuing the subject is typically weaker (Lee & Stankov, 2018). Educational stakeholders and governments are invested in these topics as a common educational policy objective aims to encourage students to choose more STEM-related subjects (Science–Technology–Engineering–Mathematics) in higher education to fill job market shortages in those areas and support technological innovation.

TIMSS 2015 surveyed both grade 4 and grade 8 students on their views on engaging teaching, their confidence, and how they like learning in each of the two subject domains (Mathematics and Science) separately, and this in a multitude of educational systems across the world. The three type of scales were almost exactly the same across the subject domains and grades in both format and wording, and a thorough translation process was applied to support the international administration of the survey. Thus, this set of survey scales (3 types $\times$ 2 domains $\times$ 2 grades $=$ 12 scales) in TIMSS 2015 offered a good variety that helps to set the context for the potential generalization of the study's findings.

Covariates for the differential prevalence study

When considering potential group differences in the prevalence of random responders, we followed the implicit hypothesis that if a participant needs to mentally push him/herself to read and respond to the items on a survey scale, the participant will be more inclined to answer randomly as a low-effort efficient reaction or due to misunderstanding of the survey question and/or response options. This implicit hypothesis and the relevance to educational policy were the two criteria that informed our choice of covariates to study. A third, more methodological criterion that came into play is that one wants to avoid having to rely on unreliable self-report group covariate information to define the groups of relevance. The group indicators that are based on self-report were restricted in this study to simple questions, early in the survey, that directly relate to a participant's identity and are expected to be more reliable and elicit higher veracity.

The TIMSS survey was administered to children in grade 4 as well as young adolescents in grade 8. Both *grade* populations responded to quite similar surveys, but they are not guaranteed to respond in a similar fashion. One can argue that questions about engagement and attitudes toward mathematics and sciences might require more effort from those in the lower grades as it might be less obvious for them to relate to or understand the questions (e.g., Mellor & Moore 2013). On the other hand, students in the higher grades are said to be more sceptical and critical toward time and effort investment affecting their response motivation (e.g., Silm, Pedaste, & Täht 2020; Rosenzweig, Wigfield, & Eccles 2019). Hence, although a grade-differential prevalence of random responders sounds not too unreasonable to expect, it is less clear what direction this would take.

Although also available as a self-report measure, information on the gender of a student was directly available as registered by the TIMSS test administrators. With respect to potential *gender* differences in the prevalence of random responders, a literature review by DeMars et al. (2013) concluded that overall, when considering attendance, response times, and self-reported effort, females would be expected to put more effort into low-stakes tests than males. The review mostly covered achievement tests, but it sounds reasonable to extend a similar expectation to a survey context. Tentative explanations for such differential prevalence bring up gender-stereotyped personality trait differences in terms of conscientiousness and agreeableness (see also Löckenhoff et al., 2014; Bowling et al., 2016).

In education, the link between *socio-economic status* (SES) and educational outcomes (for an achievement-focused review, see e.g., Sirin, 2005) is a robust finding and reason

for concern and research on educational inequalities and inequity. As a proxy for a student's SES, we used the self-reported estimate of the number of books at home. Based on a comparison with official register data in Sweden, Wiberg and Rolfsman 2023 recommended the use of this self-report measure, with the added benefit that it is simple and has low omission rates. In the survey non-response literature (e.g., Goyder, Warriner, & Miller 2002), it is common to find lower non-response rates with higher SES, and this at all stages of the survey data collection. Reasons for this non-response trend are less clear, but speculated to be linked to socio-psychological factors. Following these findings, we expected to observe a similar difference in the prevalence of random responders between low and high SES groups.

Spoken language at home might be another potential factor related to the differential prevalence of random responders. When the language of the survey is different from the language the student speaks at home, this might require more effort, both cognitively in terms of ease of understanding as well as mentally in terms of engagement/relating to the survey. In the context of achievement tests for young adults, Goldhammer et al. (2017) observed that a difference between test and home language was related to more disengagement as measured by more rapid-guessing. Hence, also for the prevalence of random responders, we expected a similar difference to apply.

We also considered *migration background*, an issue that is often of prime interest for policymakers. Based on the self-reports on whether their respective parents were born in the country where the survey was administrated, a crude student migration background index was constructed. General expectations on the relation of this covariate to the prevalence of random responding are hard to make as the contextual factors surrounding immigration will heavily differ depending on the educational system.

Furthermore, we will map and report resulting patterns of student group differences in the prevalence of random responders across the different *educational systems* participating in TIMSS 2015, but, lacking a well-justified theory on such cross-system differences, no further hypotheses were made.

In sum, the key research question addressed by this study is 'who are the random responders on the students' engagement and attitudes toward mathematics and science survey scales of TIMSS 2015?'. More specifically, we investigated whether being classified as random responder instead of typical responder is associated with student characteristics such as grade, gender, SES, spoken language at home, or migration background.

Methods

TIMSS is an international large-scale assessment of mathematics and science, which has been conducted normally every four years since 1995. TIMSS 2015 provides the sixth assessment of trends in the fourth grade and/or eighth grade of fifty-seven educational systems and seven benchmarking participants, including assessments of mathematics and science achievement as well as context questionnaires collecting background information (Mullis & Martin, 2013).

The student questionnaire is given to each student who takes part in the assessment, with some questions identical for the fourth-graders and eighth-graders. The student questionnaire for eighth grade has an integrated version and a separated version, depending on the implemented science program in the educational system. The

integrated version is for those with science as a single or general subject, while the separated version is for those where science is separated into different subjects, including biology, earth science, chemistry, and physics.

Sample

We considered the educational systems that participated in both the mathematics and the science assessment of TIMSS 2015, with both grade four and grade eight students, and that were not one of the added benchmarking participants. Furthermore, to retain close comparability of student questionnaires between grades four and eight, we only included educational systems with an integrated science program. This ensured that student questionnaires are consistent in terms of questionnaire length, scale items, and scale position. In total, 22 educational systems[1] meet these inclusion criteria: Australia (AUS), Bahrain (BHR), Canada (CAN), Chile (CHL), Chinese Taipei (TWN), England (ENG), Hong Kong SAR (HKG), Iran, Islamic Rep. of (IRN), Ireland (IRL), Italy (ITA), Japan (JPN), Korea, Rep. of (KOR), Kuwait (KWT), New Zealand (NZL), Norway (NOR), Oman (OMN), Qatar (QAT), Saudi Arabia (SAU), Singapore (SGP), Turkey (TUR), United Arab Emirates (ARE), and United States (USA).

TIMSS' target sample size for the number of students to be reached within an educational system is $n = 4000$ across a minimum school sample of 150 schools (if student population size and other practicalities permit; see e.g., *Chapter 3: Sample Design* in Martin et al. (2016)). For the set of educational systems in this study, sample sizes ranged from 3593 grade 4 students in Kuwait to 21177 in the United Arab Emirates, and from 3759 grade 8 students in Saudi Arabia to 18012 in the United Arab Emirates. Tables 1 and 2 in Appendix 1 summarize these and other descriptive statistics.

Measures

The TIMSS 2015 student questionnaire covers basic background questions about the students and their home situation, and it includes questions about the students' school experiences, engagement and attitudes with respect to school subjects and homework.

Survey scales: students' engagement and attitudes toward mathematics and science

The six survey scales measured three types of student engagement and attitudes on two subject domains (mathematics and science): Like Learning Mathematics (variables: 'ASB01A'-'ASB01I' in grade 4, 'BSBS17A'–'BSBS17I' in grade 8), View on Engaging Teaching in Mathematics Lessons (variables: 'ASB02A'-'ASB02J' in grade 4, 'BSBS18A'–'BSBS18J' in grade 8), Confidence in Mathematics (variables: 'ASB03A'-'ASB03I' in grade 4, 'BSBS19A'–'BSBS19I' in grade 8), Like Learning Science (variables: 'ASB04A'-'ASB04I' in grade 4, 'BSBS21A'–'BSBS21I' in grade 8), View on Engaging Teaching in Science Lessons (variables: 'ASB05A'-'ASB04J' in grade 4, 'BSBS22A'–'BSBS22J' in grade 8), and Confidence in Science (variables: 'ASB06A'-'ASB06G' in grade 4, 'BSBS23A'–'BSBS23H' in grade 8). These were Likert scales of between 7 and 10 items using four categories ranging from 'agree a lot', over 'agree a little'/'disagree a little', to 'disagree a lot'.

[1] Their corresponding (ISO) code will be used as the label in figures and tables.

More information on the specific items in the survey scales can be found in the different versions of the TIMSS 2015 student questionnaire which have been made publicly available.[2]

The Like Learning scales (abbreviated as Like-M and Like-S) contain items related to how the student perceives and enjoys the subject and are also referred to as measuring intrinsic motivation (e.g., Michaelides, Brown, Eklöf, & Papanastasiou 2019). The Confidence scales (abbreviated as Conf-M and Conf-S) contain items related to students' self-concept with respect to the subject domain (e.g., Michaelides et al., 2019). The View scales (abbreviated as View-M and View-S) contain items related to how the student perceives their teacher's interaction with both the subject and the students. In further analyses, the students' responses on items of the scales form the main indicators to flag potential random responders. Note that for scales with mixed-wording, negatively-worded items are reverse coded.

Covariates: five student characteristics

Five student characteristics were considered as covariates potentially related to the prevalence of random responders on the questionnaire scales in an educational system in TIMSS 2015 (for descriptive statistics by grade and per educational system, see Appendix 1: Tables 1 and 2).

Grade

The student's grade is a non-student-reported variable based on whether the student was part of the grade four or grade eight administration of the student questionnaire (TIMSS provides separate datasets per grade by country). This grade variable was dummy coded, with grade four coded as zero, and grade eight coded as one. Note that some educational systems, for reasons related to curriculum or the current state of education, decided to participate with different grades than four and eight (i.e., Norway, England, and New Zealand participated with grade five and grade nine). Regardless, these grades will still be labeled four and eight during the analyses. There were no missing data for this background variable.

Gender

For gender, we used the non-student-reported variable 'ITSEX' from the Student Tracking Form, which is filled out by the test administrators (e.g., Martin et al. 2016). This gender variable was dummy coded, with female coded as zero, and male coded as one. The male-to-female student ratio was about 50/50 in both grades of all participating educational systems (the biggest imbalance was 54% to 46% in Hong Kong). There were no missing data for this background variable.

Self-reported socio-economic status (SES)

The students reported an estimated number of books at home on an ordered scale of five categories: "None or very few (0–10 books)", "Enough to fill one shelf (11–25 books)",

[2] TIMSS 2015 Context Questionnaires: https://timssandpirls.bc.edu/timss2015/questionnaires/index.html.

"Enough to fill one bookcase (26–100 books)", "Enough to fill two bookcases (101–200 books) and "Enough to fill three or more bookcases (more than 200 books)". The five categories of this number of books variable ('ASBG04' and 'BSBG04' in grade 4 and grade 8 student questionnaire, respectively) were recoded to a scale ranging from 0 to 4. The distribution of the number of books variable varied widely across educational systems and grades. For example, in Korea 29% of fourth-graders and 25% of eighth-graders reported having 101–200 books, and 44% and 39% reported having more than 200 books at home, respectively. In contrast, only 11% of fourth-graders reported having more than 100 books in Chile and only 10% of eighth-graders in Kuwait. In most educational systems no more than 5% of the students did not provide a response to this survey question, with the exception of fourth-grade students in Saudi Arabia (9%), and students of both grades in Kuwait (10%).

Self-reported language at home

The students reported their frequency of speaking the language of the achievement test and student questionnaire at home on an ordered scale of four categories. This language variable (i.e., 'ASBG03' in the grade 4 student questionnaire and 'BSBG03' in the grade 8 student questionnaire) was dummy coded, collapsing the categories "never" and "sometimes" to be coded as zero, and collapsing the categories "almost always" and "always" to be coded as one. The proportion of students considering themselves to speak (almost) always the language of the test at home varied largely across the educational systems, from only 19% in Kuwait to 100% in Korea. In most educational systems, eighth graders reported more often than fourth graders to (almost) always speak the language of the test at home, with an average between-grade difference of 9 percentage points. On average, about 5% of the students in an educational system did not respond to this survey question, with the largest proportion of missingness (up to 10%) in Kuwait.

Self-reported migration background

Students were asked whether their mother was native-born and whether their father was native-born. Both the father variable (i.e., 'ASBG06A' in grade 4 and 'BSBG09A' in grade 8)' and the mother variable (i.e., 'ASBG06B' in grade 4 and 'BSBG09B' in grade 8) had three response categories: "Yes", "No", and "I don't know". A dummy variable was created based on whether the student reported, on at least one of the two variables, their parent to be foreign-born. A combination of one native-born and either an omitted or "I don't know" response resulted in a missing score on this dummy variable; the same holds for a combination of responses only consisting of an omitted or "I don't know" response. The proportion of students reporting to have at least one foreign-born parent varies widely across educational systems, from as low as 1% in Korea to as high as 66% in the United Arab Emirates. On average about 6% of the students in an educational system missed a score on the migration dummy, with the largest proportion of missingness (up to 22%) for grade 4 students in Taiwan and the United States.

Outcome: classification as random responder

Following a mixture item response theory (IRT) approach (Sen & Cohen, 2019), we classify a student as a random or as a typical responder, for each of the six survey scales

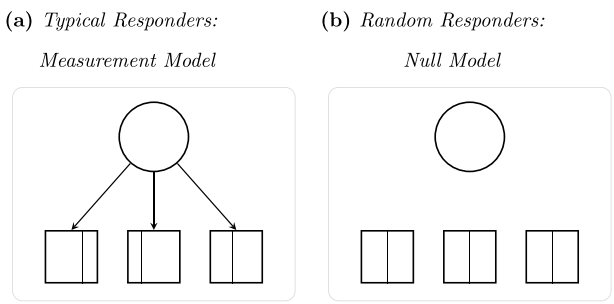

Fig. 1 Mixture IRT model Framework to Define and Operationalize Random Responders in terms of Independence and Uniformity of Item Responses. *Note* Symbols follow standard path diagram conventions, with squares representing observed variables (i.e., item responses); circles, latent variables (i.e., trait to be measured by the scale of items); arrows indicating dependence relations; vertical lines, response category thresholds. Typical responders **a** conditional independence given the latent trait; Random responders **b** mutual independence with uniformly distributed response categories (cf. squares divided into equal parts and no relation with circle or other squares). Reprinted under the terms of CC-BY-NC from van Laar and Braeken (2022), *Journal of Educational Measurement*

considered in this study, using an extension of the HYBRID model by Yamamoto's Yamamoto (1989) to the polytomous case for survey responses as proposed by van Laar and Braeken (2022). Classification is based on the maximum posterior class membership probability of a mixture model consisting of two classes. The approach assumes that there are two distinct, yet unobserved latent groups of responders in the population expressing different response behavior on the survey scale: the class of 'random responders' and the class of 'typical responders' (see Fig. 1). In the class reflecting the typical responders, a student is assumed to provide responses across items in a consistent fashion according to their value on the underlying common latent trait (see Fig. 1a). In the class reflecting the random responders, a student is assumed to provide unrelated responses across items in a more haphazard fashion (see Fig. 1b). More specifically, this comes down to a mixture of (i) a graded response model (Samejima, 1969) for ordered item responses and (ii) a null model with independent item responses that have an equal chance of falling in either of the possible response categories. Note that because the class model for random responders has only fixed known parameters, the mixture model only has one extra parameter to estimate compared to a conventional graded response model, being the mixture class weight which can be seen as the prevalence estimate of random responders in the population.

Estimation

The mixture IRT model was estimated separately for each scale per educational system in each grade. Models were estimated in Mplus Version 8.2 (Muthén & Muthén, 2017) through the MplusAutomation package for R version 0.7-3 (Hallquist & Wiley, 2018) (for an example of Mplus syntax see Appendix 2). We accounted for the total student weights in the TIMSS sampling design and used full-information maximum likelihood estimation with robust standard errors and the expectation-maximization acceleration algorithm with a standard of 400 random starts, 100 final stage optimizations, and 10 initial stage iterations. For each model, the resulting classification variable was a dummy variable with a typical responder being coded zero and a random responder coded as

one. These dummy variables were the main outcome variable for further analyses in the current study.

Quality check

If the mixture model for a specific country-scale combination failed either of two quality checks, the corresponding outcome variable was set to missing. First, the measurement model for the typical responders in the mixture was inspected to ensure that it reflected a clean unidimensional model (i.e., compatible with the assumed common trait for the survey scale). This criterion was not met when two or more standardized item discrimination parameters (i.e., factor loadings) were below 0.40. Secondly, a classification entropy of at least 0.70 was required to ensure that the mixture model was able to provide a good enough distinction between the two latent groups of responders.

Statistical analysis

Odds ratios (OR) were computed as an effect size measure comparing whether the odds of having been classified as a random responder on a specific survey scale are different between the student groups identified by the respective covariate. Odds ratios of 1.22, 1.88, and 3.00 were interpreted as small, medium, and large effect sizes, respectively (Olivier & Bell, 2013) (for negative dependence, the corresponding inversed values are 0.82, 0.53, and 0.33). Computations were student-weighted in accordance with the TIMSS sampling design and run via the R-package 'survey' (Lumley, 2010). For the grade covariate, the data was combined across grades, per educational system by scale combination, to allow for a comparison between grade 4 and grade 8 students. For the grade covariate the odds ratio was computed for each scale based on the across-grades pooled dataset per educational system; For the four other covariates, the odds ratio was computed within each grade, per educational system by scale combination. Hence, a total of 1188 (i.e., $(1 + 4 \times 2) \times 22 \times 6$) odds ratio estimates were obtained.

To summarize the abundance of results, we made use of meta-analytic tools (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2021) via the R-package 'metafor' (Viechtbauer, 2010). The confidence interval of the average log odds ratio (i.e., log(OR)) across educational systems was computed under the random effects meta-analytic model. The latter confidence interval was supplemented by its corresponding prediction interval; the width of the prediction interval relative to the confidence interval reflects the amount of heterogeneity in effect size among the educational systems. The further away the prediction interval stretches from the confidence interval, the more different the effect sizes across systems are. We briefly summarized noticeable system-specific patterns in the text and included forest plots in Appendix 1 that display the individual estimates per covariate for each educational system, per grade by survey scale combination. All analysis scripts were run under R version 4.0.0 (R Core Team, 2020).

Results

Prevalence of random responders

As mentioned before, we had two quality checks to determine whether the resulting classification following the mixture IRT approach to detect random responders could be relied on for further analyses. For the Like and the View scales in both grades and

both Mathematics and Science, the random responder classification passed the quality checks for all educational systems without exception. This was not uniformly the case for the Confidence scales. In grade 4, the classification for seven and eight educational systems (out of 22) did not pass the quality checks for Mathematics (i.e., ARE, BHR, IRN, KWT, OMN, QAT, SAU) and Science (i.e., ARE, BHR, CHL, IRN, KWT, OMN, QAT, SAU), respectively. In grade 8, this was the case for four and three educational systems (out of 22) in Mathematics (i.e., BHR, KWT, OMN, SAU) and Science (i.e., BHR, OMN, SAU), respectively. Notice that it was mostly the same subset of educational systems that did not pass the quality checks for the Confidence scale; mainly due to the questionnaire scale not adhering in those systems to the anticipated unidimensionality of the construct. For the corresponding educational systems not passing the quality checks, no further analyses linking the random responder classification to covariates will be performed, such that they will further appear as missing in the summary graphics and statistics reported.

For countries that passed the quality checks, the average prevalence of having been classified as a random responder on the Like and View scales was around 4%, ranging from 1 to 11% across educational systems and grades (see Fig. 2), while the average prevalence on the Confidence scales was somewhat higher at about 11%, ranging from 2 to 27%). The overall average prevalence (across scales, grades, and educational systems) was around 7%.

Random responder = f(grade)

The relation between having been classified as a random responder and grade differed across the six scales. On average across the 22 educational systems, grade eight students had significantly higher odds of having been classified as a random responder than grade four students on both View scales (OR = 1.79, small to medium effect size) and the Like Science scale (OR = 1.33, small effect size), whereas no such support was found on both Confidence scales and the Like Mathematics scale (see Fig. 3, the confidence intervals (black diamonds) of View-M, View-S, and Like-S exceed zero; the confidence intervals of Conf-M, Conf-S and Like-M include zero). The width of the prediction intervals in Fig. 3 did imply heterogeneity among the educational systems. For instance, Iran showed the most obvious grade difference in the prevalence of random responders (OR = 2.84, medium effect size), especially on the View Mathematics scale (OR = 8.67, large effect size), while Singapore showed an opposite grade difference (i.e., grade 8 < grade 4) in five of the six scales (OR = 0.68, small effect size).

Random responder = f(gender)

On average across the 22 educational systems, male students had significantly higher odds of having been classified as a random responder, and this on all six scales and in both grades (see Fig. 4, all confidence intervals (black diamonds) exceed zero). The average odds ratio for the six scales ranged from 1.10 to 1.58, with a median of 1.46 (i.e., log(OR) =.38), corresponding to a significant but small effect size, and hence gender difference in the prevalence of random responders. Although the gender difference applied quite generally, the width of the prediction intervals in Fig. 4 implied heterogeneity among the educational systems. For instance, Chile and the USA were the educational

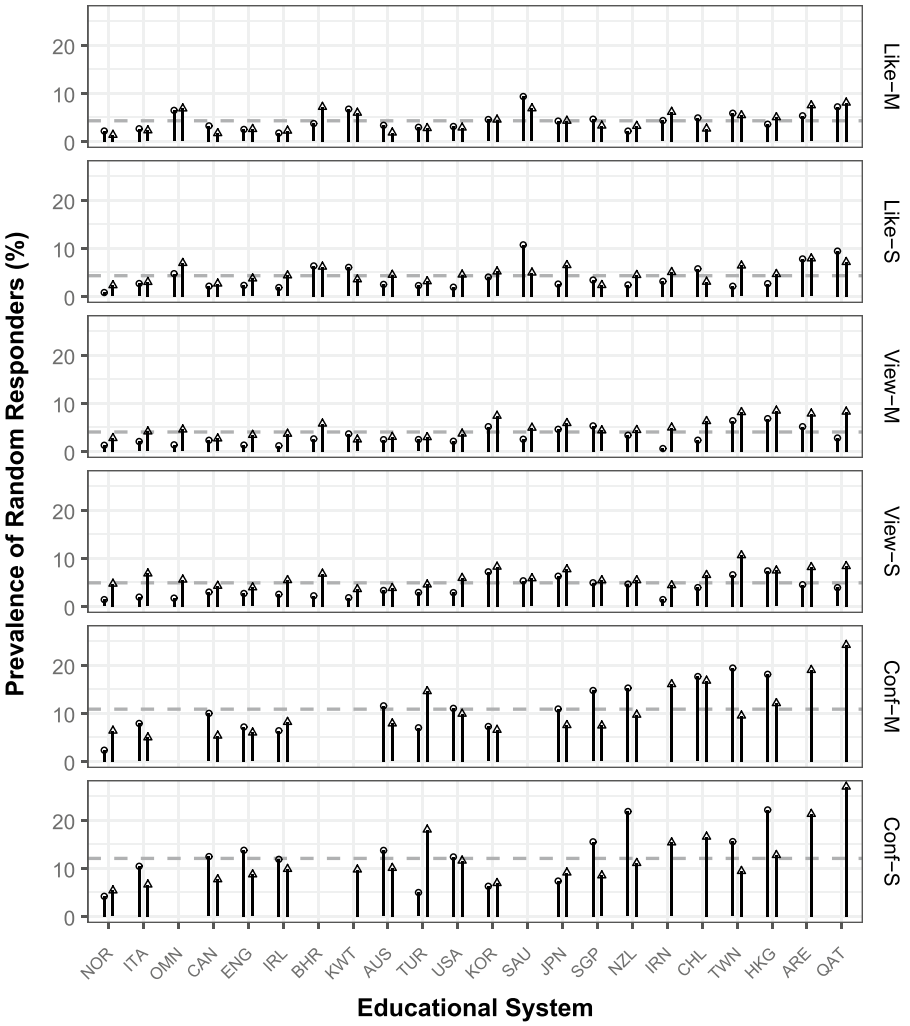

Fig. 2 Estimated Prevalence of Having Been Classified as a Random Responder on the six Questionnaire Scales across Educational Systems. *Note* Circles and triangles represent grade four and grade eight, respectively. Educational systems were ordered by across-grade-and-scale average prevalence, with the gray dashed line being the across systems and across grades average for the scale

systems where the gender difference was almost absent (i.e., log(OR) ≈ 0), whereas Saudi-Arabia and Oman were two educational systems with a more pronounced gender difference in the prevalence of random responders (i.e., medium OR effect sizes). For Norway, there was no support for a gender difference for either View scale, but it had the highest observed gender difference among systems on the Like Mathematics scale (average OR = 2.51 across grades).

Random responder = f[number of books at home (SES)]

On average across the 22 educational systems, students with a higher self-reported number of books at home had significantly lower odds of having been classified as a random responder on the Like scales (average odds ratio across grades: OR = 0.93 for Mathematics; OR = 0.87 for Science) and the Confidence scales (average odds ratio across grades: OR = 0.85 for Mathematics; OR = 0.81 for Science), but no support for such relation

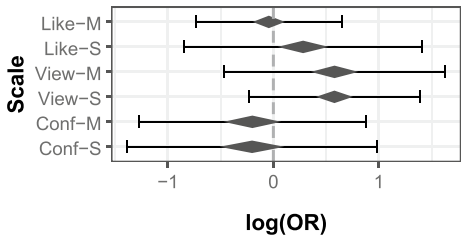

Fig. 3 Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Grade. *Note* The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix 1: Fig. 8 and Tables 1 and 2. A positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for grade 8 than for grade 4 students. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' engagement and attitudes toward Mathematics and Science

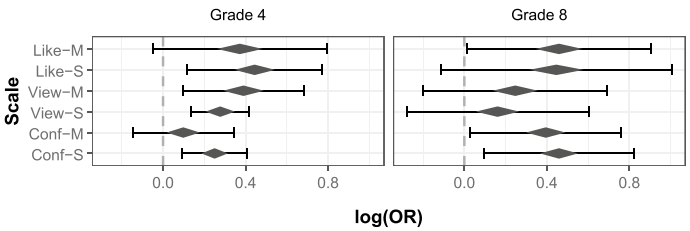

Fig. 4 Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Gender. *Note* The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix 1: Fig. 9 and Tables 1 and 2. A positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for male than for female students. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' engagement and attitudes toward Mathematics and Science

was found on the View scales (see Fig. 5). Note that the number of books covariate had 5 ordered categories, and the interpretation here was for only one category difference, hence the difference between students with the most (more than 200 books) and the fewest (0–10 books) self-reported number of books at home was expected to be four units. For instance, the median of the across-systems average odds ratios for the six scales was 0.91 (i.e., $\log(\text{OR}) = -0.09$) in one unit difference, leading to a small effect size of OR $= 0.70$ when comparing the two scale-extremes (i.e., $\exp(-0.09 \times 4) = \exp(-0.09)^4$). The prediction intervals indicated that most educational systems showed that students reporting to have more books at home had significantly lower odds of having been classified as a random responder on the Confidence scales. Yet, the width of the prediction intervals in Fig. 5, for these and the other four scales, did imply heterogeneity among the educational systems. For instance, Chile and Saudi Arabia were the educational systems where the number of books difference was almost absent (i.e., average $\log(\text{OR}) \approx 0$), while England and New Zealand had the largest OR effect sizes among systems (average OR = 0.82 and 0.83, respectively). For Ireland, there was no support

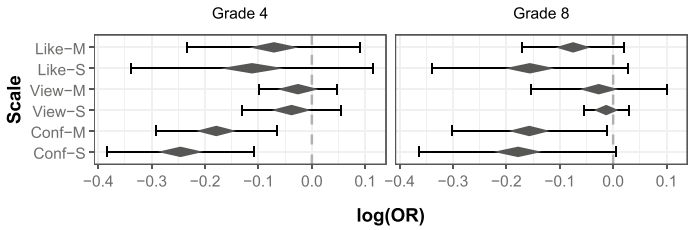

Fig. 5 Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Number of Books at Home. *Note* The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix 1: Fig. 10 and Tables 1 and 2. Number of Books at Home is coded 0 = None or very few (0–10 books)/1 = Enough to fill one shelf (11–25 books)/2 = Enough to fill one bookcase (26–100 books)/3 = Enough to fill two bookcases (101–200 books)/4 = Enough to fill three or more bookcases (more than 200 books), hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher for students who reported having more/fewer books at home. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' engagement and attitudes toward Mathematics and Science

for a number-of-books difference for the Like Mathematics and View Science scales, but it had the highest observed number of books difference among systems on the Confidence in Science scale (average OR = 0.68). At the individual educational system level, the confidence intervals for fourth grade are generally wider than for eighth grade (see Appendix 1: Fig. 10).

Random responder = f(language at home)

On average across the 22 educational systems, students who more often speak the test language at home had significantly lower odds of having been classified as random responders than those who don't speak the same language at home, and this on all six scales and in both grade four and grade eight (see Fig. 6, all confidence intervals are below zero). The average odds ratio for the six scales ranged from 0.68 to 0.89, with a median of 0.84 (i.e., log(OR) = − 0.17), corresponding to an ignorable to small effect size. The width of the prediction intervals in Fig. 6 did imply heterogeneity among the educational systems, with prediction intervals even wider and more negative effect sizes for individual systems in grade eight than in grade four. For instance, Japan showed the most obvious language-related prevalence difference (average OR = 0.37, medium effect size). Note that the language covariate had extreme distributions in some educational systems, such as in Japan and Korea where very few students reported speaking any other language at home (1–2% of fourth and eighth graders in Japan and close to 0% of eighth-graders in Korea), contributing to wider confidence intervals in these systems. Qatar's grade eight was the only educational system where speaking the same language had a positive relation to having been classified as a random responder (average OR = 1.46, small effect size).

Random responder = f(migration background)

On average across the 22 educational systems, no significant relation was found between having at least one foreign-born parent (versus both native-born parents)

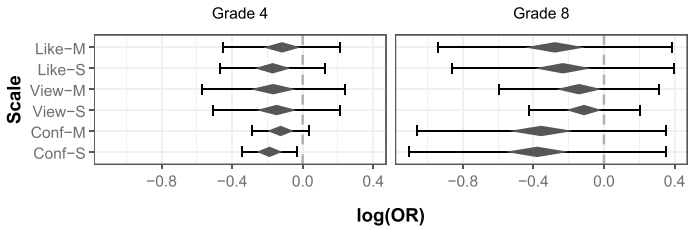

Fig. 6 Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Spoken Language at Home. *Note* The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix 1: Fig. 11 and Tables 1 and 2. Language at Home is coded 1 = Always or almost always speak < language of test < at home/0 = Sometimes or never speak < language of test < at home, hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher for students more/less frequently speaking < language of test < at home. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' engagement and attitudes toward Mathematics and Science

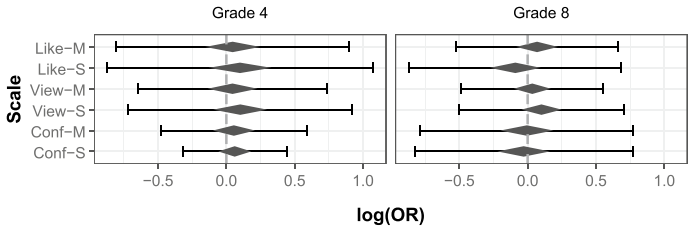

Fig. 7 Meta-analytic confidence and prediction intervals for the odds of having been classified as a Random Responder as a function of the student's Migration Background. *Note* The black diamond represents the confidence interval around the estimated average log odds ratio across educational systems, and the whiskers extending the diamond define the corresponding prediction interval. The gray dashed vertical line is drawn at log(OR) = 0, corresponding to independence between the covariate and the random responder classification. For the estimates per system, see Appendix 1: Fig. 12 and Tables 1 and 2. Migration background is coded 1 = At least one foreign-born parent/0 = Both native-born parents, hence a positive/negative log(OR) indicates that the odds of having been classified as a random responder is higher/lower for students with than without migration background. Results are reported for six scales in the TIMSS 2015 student questionnaire measuring three types of students' engagement and attitudes toward Mathematics and Science

and having been classified as random responders on all six scales for both grades. However, the width of the prediction intervals in Fig. 7 did imply heterogeneity among the educational systems, with different directions of effect sizes for individual systems (see Appendix 1: Fig. 12). For instance, the United Arab Emirates and Qatar showed the strongest negative migration background prevalence differences (i.e., students with at least one foreign-born parent had significantly lower odds of having been classified as random responders than those with both native-born parents, average OR = 0.55 and 0.56, respectively, medium effect sizes), whereas Turkey and Iran showed the strongest positive migration background prevalence differences (i.e., students with at least one foreign-born parent had significantly higher odds of having been classified as random responders than those with both native-born parents, average OR = 2.49 and 1.90, respectively, medium effect sizes). Note that some

educational systems such as Japan and Korea had few students with migration backgrounds (i.e., under 5%), contributing to wider confidence intervals in these systems.

Discussion

Although observable responses are still provided, a random responder can be seen as causing a form of *nonresponse error*, in that we end up lacking accurate data on the genuine attitude or information the student is surveyed about. Similar to more traditional nonresponse, a low prevalence of random responders can be seen as a quality indicator for both the survey and response data resulting from the survey. We found an overall prevalence of random responders ranging from 1 to 27%, with an average of 7% across educational systems for the six TIMSS 2015 scales measuring students' engagement and attitudes toward mathematics and science. Hence, supporting the quality of international large-scale assessments in comparative educational research, this prevalence is relatively low. Yet this 7% average does represent some of those students that typically make up for the stereotypical anecdotes that are underlying general concerns about whether students provide genuine valid responses to the questionnaire in these typical low-stakes assessments. The range of prevalence estimates is comparable to numbers found in the literature for self-report inventories in other fields (e.g., Credé, 2010; Steedle, Hong, & Cheng, 2019).

Differential prevalence of random responders

Similar to nonresponse (e.g., Richiardi et al., 2013), the impact of the prevalence of random responders crucially depends on who they are, these random responders. If minority groups or groups with other specific characteristics have a higher prevalence of random responders, such systematic disproportionate differences can lead to selective fallout in the sample and if the propensity of engaging in random response behavior relates to the survey outcomes of interest this can potentially skew, or at worst invalidate, inferences/conclusions based on the questionnaire scales. The key research objective in this study was to investigate whether random responders were disproportionately present in groups defined by research- and policy-relevant covariates. We used a mixture IRT approach to classify students as random responders and meta-analysis summaries to present our results for each of six questionnaire scales across 22 educational systems and two grades.

We found a small to medium grade difference in prevalence for the View scales (and the Like-S scale), with grade eight students having higher odds of having been classified as a random responder than grade four students on average across educational systems. This was counter to our implicit hypothesis that assumed the questionnaire to be less taxing for the students in the later grade, but it could very well be consistent with a higher intrinsic motivation of younger kids versus young adolescents, similar to the observed decline for achievement tests in students' expectancies and task values (e.g., Rosenzweig et al., 2019). We found a small gender difference in prevalence, with male students having higher odds of having been classified as a random responder than female

students. This seems in line with the stereotype expectation that girls are more diligent and that boys would put in less effort in low-stakes situations (e.g., DeMars et al., 2013). Context-wise, a small SES difference in prevalence was found for all scales except the View scales, with students reporting having fewer books at home also having higher odds of having been classified as a random responder than students reporting having more books at home. This SES difference is in line with findings in the more general nonresponse literature (Goyder et al., 2002). A small to ignorable language difference in prevalence was also found, with students speaking a language at home different from the test language having higher odds of having been classified as a random responder than students with matching language, with the trend being more pronounced in grade four than in grade eight. This is in line with a priori expectations following the ease of understanding and mental engagement, and consistent with findings in the rapid guessing literature (e.g., Goldhammer et al., 2017). For immigration background, no empirical support for a difference in prevalence was found using the crude self-reported parents' birthplace indicator.

Generalizability

The findings of this initial study indicate that who are random responders is not entirely random. The obvious caveat remains that there still might be other crucial covariates than those considered here on which the two groups might systematically differ. As noted in the introduction, some of that covariate information might not always be as easy to measure reliably and validly. One should especially be aware of the catch-22 risk of using self-report measures to characterize responders that might not genuinely report back on those indicators. Furthermore, some of the available covariate indicators might be suboptimal: the number of books for SES or parents' birthplace for migration background might not necessarily be the optimal indicators in all cultures or not all younger kids might in fact be able to reliably provide such information. Thus it would be good not to generalize the null findings for the latter covariates beyond the specific operationalization used in this study.

With respect to the scales, only the View scales showed a grade difference in prevalence and almost no SES difference in prevalence, whereas the Confidence scales had a higher overall prevalence of random responders (i.e., on average 11%). Note that the Confidence scales also tended to fail quality checks for mostly the Middle East countries, indicating larger measurement issues there for the majority of students. Altogether these findings do indicate that whatever the mechanisms are underlying random responding, these won't be all generic or uniformly applicable across scales. This implies a crucial role for scale contents and for how students (i.e., the target population) engage with or understand the questionnaire scale contents.

The observed heterogeneity across educational systems implies that context does matter. Whereas on average a difference in prevalence between two covariate groups might be absent, it might still apply to an individual educational system. For example, a high SES

difference in prevalence was observed in England and New Zealand, and grade eight students in Qatar that spoke the same language at home as the test surprisingly had higher odds of having been classified as a random responder than those who did not. The latter finding is likely due to the somewhat atypical immigrant population in Qatar compared to other systems in our study. Similarly, when considering language and migration background, the native culture was so dominant in Japan and Korea that the minority groups were very small, leading to somewhat larger but also more uncertain prevalence differences than elsewhere.

Handling random responders

Having been classified as a random responder does not necessarily mean that one consciously and purposefully provides random responses. The classification has only a direct binding to the observed response pattern and not to the underlying intentions or response process. Random response patterns can equally arise due to incidental inattention or lack of understanding of the question or uncertainty about the applicability of response options, and so on. In this sense, it is perhaps more natural to qualify the responses given as nonresponse instead of as definite invalid. Hence, we recommend similar approaches as used in the handling of missing data, to deal with data from random responders (e.g., Meng, 2012). This would imply sensitivity analyses comparing inferences with and without the inclusion of the detected random responders and techniques such as multiple imputation on a rich feature set of relevant covariates and survey design variables to comply with a missing-at-random working assumption. Note that the latter does not mean completely at random (for which we have indications it is not), but conditional on the relevant covariate group differences as suggested in explorative studies like the current study.

Conclusion

Similar to missingness rates, prevalence rates of random responders don't tell the whole story, as their influence will depend on the underlying mechanism: the other variables involved and who in effect provides the nonresponses. This study has shown the prevalence of random responders on questionnaire scales in international comparative educational research to be a function of common policy-relevant covariates. Therefore, we call for two actions: (i) For individual researchers using data from the questionnaires of the international large-scale assessments in education, a default practice of sensitivity analyses and robustness checks; (ii) For the larger testing organizations (e.g., OECD or IEA), the default inclusion of a wide arsenal of survey quality indicators including not only prevalence but also relations to covariates, and this for a larger set of non-response behavior including random responders.

Appendix 1

See Figs. 8, 9, 10, 11, 12 and Tables 1, 2.

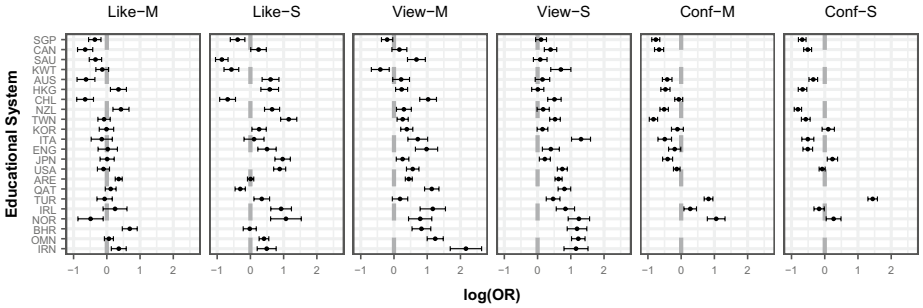

Fig. 8 System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Grade

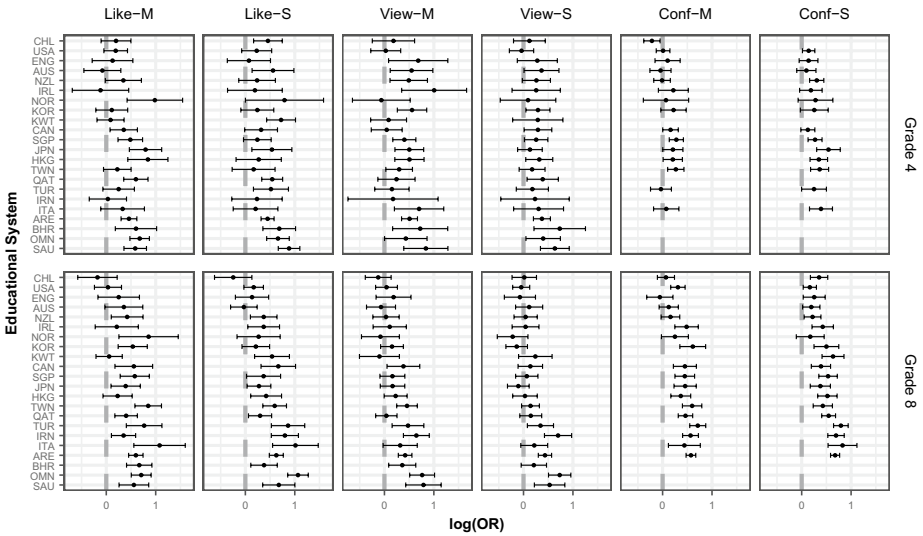

Fig. 9 System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Gender

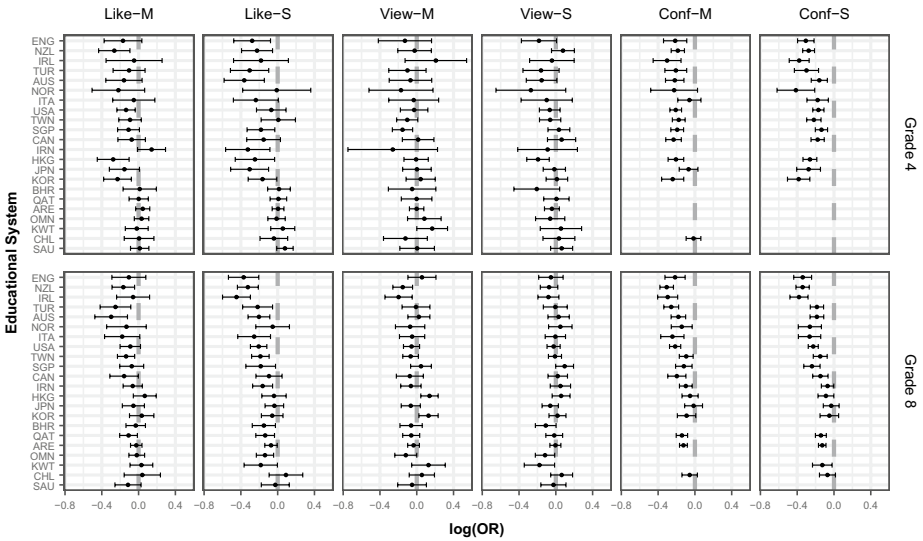

Fig. 10 System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Number of Books at Home

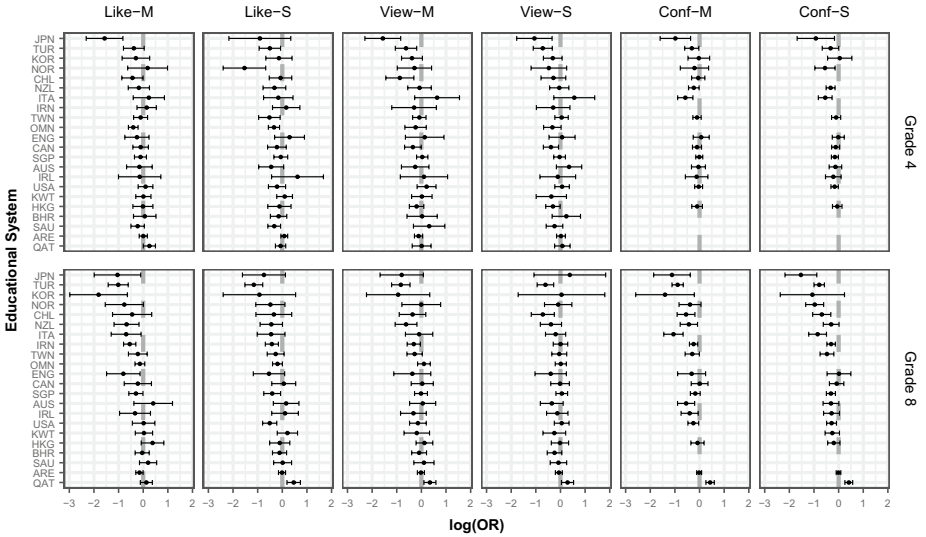

Fig. 11 System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Spoken Language at Home

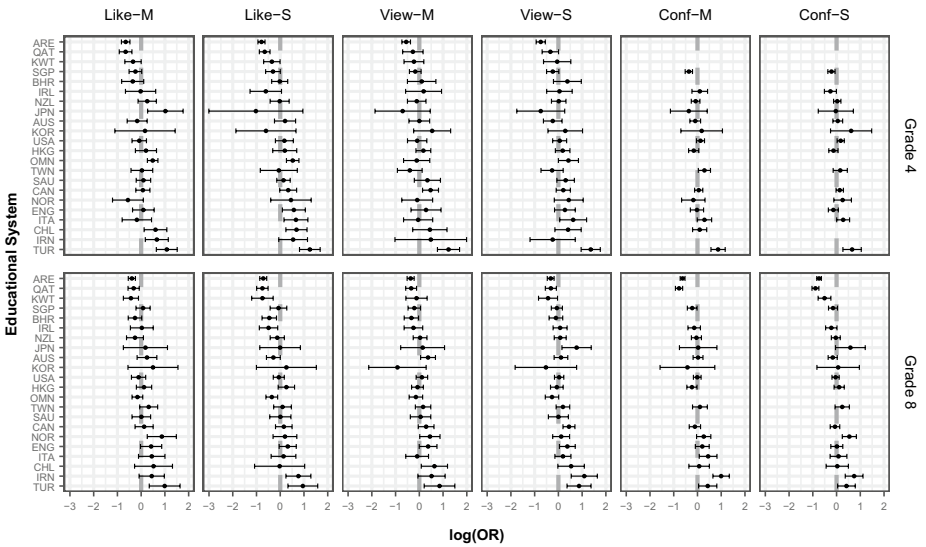

Fig. 12 System-specific confidence intervals for the odds of having been classified as Random Responder as a function of the student's Migration Background

Table 1 Percentage of fourth-grade students in each of covariates' categories

Educational system	Gender: female	Gender: male	SES: 0–10 books	SES: 11–25 books	SES: 26–100 books	SES: 101–200 books	SES: > 200 books	SES: missing	Language: never/ sometimes	Language: (almost) always	Language: missing	Migration: native-born parents	Migration: foreign-born parent	Migration: missing	Sample Size: *n*
ARE	48	52	19	30	27	10	9	4	45	52	4	25	61	15	21177
AUS	49	51	8	18	35	21	16	1	15	84	2	52	37	11	6057
BHR	50	50	20	29	25	11	11	3	31	66	3	52	36	12	4146
CAN	49	51	10	21	38	17	13	2	24	74	2	45	42	12	12283
CHL	49	51	31	33	22	6	5	3	10	87	3	84	8	8	4756
ENG	51	49	10	22	34	18	14	3	17	81	2	52	34	13	4006
HKG	46	54	14	20	32	18	16	1	29	70	1	36	47	17	3600
IRN	49	51	40	27	18	6	6	2	33	65	2	78	12	10	3823
IRL	47	53	9	20	33	20	16	1	12	83	5	69	26	5	4344
ITA	49	51	17	35	28	10	8	1	16	83	1	75	21	4	4373
JPN	50	50	12	29	37	13	8	0	2	98	0	94	3	3	4383
KOR	48	52	4	4	18	29	44	0	8	92	0	96	2	2	4669
KWT	51	49	28	31	17	7	6	10	65	25	10	50	30	20	3593
NZL	49	51	11	19	33	19	17	2	16	83	1	47	40	13	6322
NOR	49	51	7	22	38	19	12	2	15	84	1	70	27	4	4329
OMN	50	50	28	28	23	9	10	4	36	60	4	71	19	11	9105
QAT	51	49	20	28	27	11	11	3	46	52	1	30	56	14	5194
SAU	49	51	32	27	17	7	8	9	19	74	7	66	16	18	4337
SGP	48	52	10	21	37	18	13	0	51	49	0	47	44	9	6517
TUR	49	51	22	33	28	8	5	4	15	80	5	84	7	8	6456
TWN	49	51	19	25	29	13	13	0	40	59	1	65	13	22	4291
USA	51	49	13	23	33	15	13	3	20	76	3	51	27	22	10029

Table 2 Percentage of eighth-grade students in each of covariates' categories

Educational system	Gender: female	Gender: male	SES: 0–10 books	SES: 11–25 books	SES: 26–100 books	SES: 101–200 books	SES: > 200 books	SES: missing	Language: never/ sometimes	Language: (almost) always	Language: missing	Migration: native-born parents	Migration: foreign-born parent	Migration: missing	Sample size: *n*
ARE	50	50	19	29	28	11	10	2	35	63	1	26	66	8	18012
AUS	51	49	11	18	26	20	20	6	7	92	2	51	38	11	10338
BHR	48	52	25	29	26	10	9	2	26	73	1	58	36	6	4918
CAN	51	49	11	21	30	18	17	3	13	85	3	53	42	5	8757
CHL	48	52	25	38	25	7	5	1	5	94	2	93	3	4	4849
ENG	51	49	17	22	28	16	15	2	5	93	2	65	29	6	4814
HKG	47	53	18	25	31	13	13	0	16	84	0	36	52	11	4155
IRN	48	52	27	33	22	8	10	0	33	67	0	95	3	2	6130
IRL	50	50	15	22	28	19	15	1	10	86	4	68	30	2	4704
ITA	49	51	16	25	25	16	18	1	11	88	1	81	17	2	4481
JPN	51	49	12	21	32	17	18	0	1	99	0	96	2	2	4745
KOR	47	53	7	7	22	25	39	0	0	100	0	98	1	0	5309
KWT	50	50	32	30	19	6	4	9	74	19	8	58	30	13	4503
NZL	51	49	14	18	29	19	17	2	7	91	2	57	36	7	8142
NOR	50	50	10	20	29	20	20	1	6	93	1	77	21	2	4697
OMN	48	52	24	33	25	9	8	2	33	66	1	76	22	2	8883
QAT	50	50	22	29	26	12	10	1	30	69	1	32	63	5	5403
SAU	51	49	37	30	19	6	7	2	27	73	1	79	15	6	3759
SGP	49	51	18	27	31	14	11	0	35	65	0	56	40	3	6116
TUR	48	52	16	35	30	11	8	1	10	90	0	95	3	2	6079
TWN	49	51	20	23	27	13	16	0	9	91	0	85	10	5	5711
USA	50	50	17	21	29	17	15	1	9	89	1	68	25	7	10221

Appendix 2: Mplus syntax of mixture IRT model for the 'Like Learning Science' scale in the eighth-grade student questionnaire in Norway

```
TITLE:
Norway_SQIS21;

DATA:
file = "NOR_SQIS21.dat";

VARIABLE:
  names = IDSCHOOL IDSTUD TOTWGT
    BSBS21A BSBS21B BSBS21C BSBS21D
    BSBS21E BSBS21F BSBS21G BSBS21H BSBS21I;
  missing = .;
  usevariables = BSBS21A BSBS21B BSBS21C BSBS21D
    BSBS21E BSBS21F BSBS21G BSBS21H BSBS21I;
  categorical = BSBS21A BSBS21B BSBS21C BSBS21D
    BSBS21E BSBS21F BSBS21G BSBS21H BSBS21I;
  idvariable = IDSTUD;
  weight = TOTWGT;
  cluster = IDSCHOOL;
  classes = c(2);

ANALYSIS:
  type = mixture complex;
  algorithm = INTEGRATION EMA;
  estimator = MLR;
  process = 3;
  starts = 400 100;

MODEL:
%overall%
  F BY BSBS21A-BSBS21I*;
  F@1;
  [F@0];
%c#1%
  F BY BSBS21A-BSBS21I*;
  F@1;
  [F@0];
  [BSBS21A$1-BSBS21I$1];
  [BSBS21A$2-BSBS21I$2];
  [BSBS21A$3-BSBS21I$3];
%c#2%
  F BY BSBS21A-BSBS21I@0;
  F@0;
  [F@0];
  [BSBS21A$1-BSBS21I$1@-1.09861228866811];
  [BSBS21A$2-BSBS21I$2@0];
  [BSBS21A$3-BSBS21I$3@1.09861228866811];

OUTPUT:
  stdyx;

SAVEDATA:
  file = cpr_NOR_SQIS21.dat;
  format = free;
  save = cprobabilities;
```

Note. The item category threshold parameters in Class 2 (i.e., random responder class) are set on a logistic scale and correspond to cumulative response category probabilities

of 25%, 50%, and 75% (i.e., 1/(1+exp(threshold))). A more detailed description of the model can be found in van Laar and Braeken (2022).

Acknowledgements
Not applicable.

Author contributions
All authors contributed to all parts of the manuscript. All authors read and approved the final manuscript.

Funding
This study was supported by a research Grant (FRIPRO-HUMSAM261769) for young research talents of the Norwegian Research Council.

Availability of data and materials
The data sets analysed during this study are available in the TIMSS 2015 International Database, (https://timssandpirls.bc.edu/timss2015/international-database/).

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 September 2022 Accepted: 30 October 2023
Published online: 18 November 2023

References

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*(3), 340.
Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. Wiley.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.
Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218–229.
Clarke, M., & Luna-Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank.
Cochran, W. G. (1951). General principles in the selection of a sample. *American Journal of Public Health, 6*(41), 647–653.
Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*(4), 596–612.
Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3–31.
Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.
DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment, 8*, 69–82.
Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy and Practice, 17*(4), 345–356.
Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education, 5*(1), Article 18.
Goyder, J. C., Warriner, K., & Miller, S. (2002). Evaluating socio-economic status (SES) bias in survey nonresponse. *Journal of Official Statistics, 18*, 1–12.
Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *The Public Opinion Quarterly, 72*(2), 167–189.
Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638.
Hedlin, D. (2020). Is there a 'safe area' where the nonresponse rate has only a modest effect on bias despite non-ignorable nonresponse? *International Statistical Review, 88*(3), 642–657.
Hernández-Torrano, D., & Courtney, M. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *PLoS ONE, 9*, Article 17.
Hopfenbeck, T. N., Lenkeit, J., Masri, Y. E., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353.
Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*(2), 95–121.
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.
Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences, 65*, 50–64.

Löckenhoff, C. E., Chan, W., McCrae, R. R., Fruyt, F. D., Jussim, L., Bolle, M. D., & Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology, 45*(5), 675–694.
Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Wiley.
Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College.
Mellor, D., & Moore, K. A. (2013). The use of Likert scales with children. *Journal of Pediatric Psychology, 39*(3), 369–379.
Meng, X.-L. (2012). You want me to analyze data I don't have? Are you insane? *Shanghai Archives of Psychiatry, 24*(5), 297–301.
Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215–237.
Michaelides, M. P., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. C. (2019). *Motivational profiles in TIMSS mathematics: Exploring student clusters across countries and time*. Springer.
Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 assessment frameworks*. TIMSS & PIRLS International Study Center, Boston College.
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th Ed.). Muthén & Muthén.
Olivier, J., & Bell, M. (2013). Effect sizes for 2 × 2 contingency tables. *PLoS ONE, 8*(3), e58777.
Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at k-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education, 50*(1), 85–129.
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Richiardi, L., Pizzi, C., & Pearce, N. (2013). Commentary: Representativeness is usually not necessary and often should be avoided. *International Journal of Epidemiology, 42*(4), 1018–1022.
Rosenzweig, E. Q., Wigfield, A., & Eccles, J. S. (2019). Expectancy-value theory and its relevance for student motivation and learning. *The Cambridge handbook of motivation and learning* (pp. 617–644). Cambridge University Press.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.
Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies, 42*(3), 411–430.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(1), 1–97.
Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives, 17*(4), 177–191.
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.
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.
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.
van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement, 59*(4), 470–501.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
Wiberg, M., & Rolfsman, E. (2023). Students' self-reported background SES measures in TIMSS in relation to register SES measures when analysing students' achievements in Sweden. *Scandinavian Journal of Educational Research, 67*(1), 69–82.
Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). Academic Press.
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.
Yamamoto, K. (1989). *Hybrid model of IRT and latent class models*. ETS Research Report Series, RR-89-41.

Publisher's Note