Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-023-09412-7

Gender differences in item nonresponse in the PISA 2018 student questionnaire

Kseniia Marcq[1] · Johan Braeken[1]

Received: 13 September 2022 / Accepted: 30 May 2023

Abstract
Gender differences in item nonresponse are well-documented in high-stakes achievement tests, where female students are shown to omit more items than male students. These gender differences in item nonresponse are often linked to differential risk-taking strategies, with females being risk-averse and unwilling to guess on an item, even if it could gain them credits. In low-stakes settings, similar trends should not apply, as the students carry no consequence for their performance. Instead, test-taking motivation is argued to be the pivoting factor, with female students seen as more motivated and omitting fewer items than male students. In contrast to the high- and low-stakes achievement tests, less is known about gender differences in item nonresponse in student background questionnaires. Using cross-classified mixed effects models, we examined gender differences in item nonresponse on the Programme for International Student Assessment (PISA) 2018 student questionnaire across 80 countries and 71 scales. On average, the odds of male students omitting a questionnaire item were double the odds of female students, consistent with the expected trend in the low-stakes setting. However, we show that gender differences in item nonresponse are not merely a function of the stakes involved for individual students but a more complex phenomenon that is context-dependent and not necessarily stable across countries, scales' formats, and contents. We argue that examining differences in item nonresponse patterns could serve as a source of additional information about the students' test-taking behaviour and the quality of the questionnaire.

Keywords Item nonresponse · PISA · Student questionnaire · Gender differences

1 Introduction

Gender differences in item nonresponse — when the respondents of one gender are more prone to not respond to the administered items than the respondents of the opposite gender — are particularly well-documented in high-stakes achievement

✉ Kseniia Marcq
kseniia.marcq@cemo.uio.no

[1] CEMO: Centre for Educational Measurement, University of Oslo, Postboks 1161, 0318 Forskningsparken, Oslo, Norway

Published online: 24 August 2023

testing, where female students have been shown to omit more items than male students (e.g., Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994; Grandy, 1987). These differences in item nonresponse propensities between female and male students are often ascribed to performance differences in the face of competition (e.g., women tend to leave more items unanswered in high-pressure situations; Niederle & Vesterlund, 2010) and attitudes towards risk-taking. In the high-stakes assessments, dominated by multiple-choice items, the latter is commonly associated with varied tendencies to guess, with male students being more willing to guess on multiple-choice items. In contrast, female students are more likely to omit items to which they do not have a definitively correct response (Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994).

The international large-scale assessments in education such as the Programme for International Student Assessment (PISA; OECD, 2019), and the Trends in International Mathematics and Science Study (TIMSS; Mullis & Martin, 2017) are typically not high-stakes for individual students. Participation in such assessments is voluntary, and the students bear little, if any, consequence for their performance. Without risks and competitiveness associated with the consequences of individual performance, students' intrinsic motivation and test-taking effort decline (Wise & DeMars, 2005), and the likelihood of item nonresponse increases (Jakwerth et al., 1999). With male students generally believed to be less conscientious and more work-avoidant (and by extension, exerting less test-taking effort) compared to female students, it is reasonable to assert that in low-stakes settings, male students omit more items than their female counterparts (for an overview of studies on gender differences in test-taking motivation in low-stakes settings, see, e.g., DeMars et al., 2013). For instance, in line with this theory, male students were found more likely than female students not to respond to the low-stakes TIMSS 2015 science and mathematics assessments in a selected subset of countries (Papanastasiou, 2020).

However, several studies have shown gender differences in item nonresponse propensity in both high- and low-stakes achievement settings to be a more complex issue, subject to variation by country, item content, and item format. For instance, while relatively stable gender differences in item nonresponse (with male students omitting fewer items) were documented in the USA (on the Graduate Record Examinations General Test; Grandy, 1987) and Israel (on the Psychometric Entrance Test and the Hadassah battery of aptitude tests, respectively; Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994), Matters & Burnett (1999) found no support for differential item nonresponse patterns by gender on multiple-choice items on the high-stakes Queensland Core Skills Test in Australia. Other studies suggest that gender differences in item nonresponse vary across content areas. On the Iowa Tests of Basic Skills and the Iowa Tests of Educational Development, Von Schrader & Ansley (2006) demonstrated that female students had higher nonresponse rates on the high-stakes mathematics test, whereas males omitted more items on reading and vocabulary tests. On the other hand, contrary to the expected pattern in the low-stakes situations, female students were shown to have higher nonresponse rates than male students in the reading literacy domain of the low-stakes PISA 2009 in Japan (Okumura, 2014), and the mathematics domain of the low-stakes German National Educational Panel Study (while no support was found for gender differences in item nonresponse on the reading domain; Köhler, Pohl, & Carstensen, 2017). Lastly, gender differences in item nonresponse have been shown to be sensitive to different item formats, with male students omitting more open-response items than female students in both high- and low-stakes settings (see, e.g., Matters & Burnett, 1999; Okumura, 2014). This contradicting evidence on gender differences in item nonresponse both within and between high- and low-stakes achievement tests suggests that item nonresponse profiles are not necessarily distinct merely due to the stakes involved for individual students. Rather, the observed gender differences in item nonresponse profiles may additionally be a function of various item response formats and contents, and the extent and directionality of the difference may not be stable across countries.

In contrast to the relatively extensive evidence base on gender differences in item nonresponse in achievement tests, systematic research on such differences is largely non-existent for student background questionnaires, a critical component of most international large-scale assessments in education often used to contextualise achievement scores in both primary and secondary research. Similarly to the achievement component of these assessments, student background questionnaires are low-stakes, but instead of assessing cognitive ability, the questionnaires gather varied non-cognitive background information on students, students' parents, schooling experiences, attitudes, and beliefs (OECD, 2019). Under these circumstances, competitiveness and risk-taking behaviour are no longer applicable, as there is no objectively correct answer to a typical survey item. Consequently, one cannot expect the trends of gender differences in item nonresponse in the achievement tests to generalise to the questionnaire setting seamlessly.

However, what can be hypothesised based on the evidence from the achievement tests is that item nonresponse in student questionnaires may also not occur completely at random (Rubin, 1976), but rather as a function of gender. If true, item nonresponse may lead to biased parameter estimates, distorting the results and weakening the validity of inferences we draw from the student questionnaire scales (see, e.g., Meinck, Cortes, & Tieck, 2017). Furthermore, we believe that systematically examining differential item nonresponse between genders may advance our understanding of students' response strategies and the quality of the questionnaire, its scales, and items.

1.1 The present study

Given the lack of solid theory and scarce empirical evidence on gender differences in item nonresponse in the low-stakes student questionnaires, the present study sets out to investigate the issue using an exploratory quantitative approach based on the empirical data from the most recent PISA 2018 student background questionnaire (OECD, 2019). The PISA 2018 student background questionnaire consisted of a large and heterogeneous item set: 306 items clustered in 75 scales with a rich variety of response formats and assessing a multitude of content areas. Furthermore, PISA 2018 was administered to large representative samples of 15-year-old students from 80 countries and jurisdictions (henceforth, “countries”), amounting to a total of over 612,000 students. The combined strength of the sample characteristics from both the student and the scale perspectives offers a strategic opportunity to generalise across multiple countries and scales as well as identify unique item nonresponse patterns pertaining to individual countries, scales, response formats, and content areas.

The first research question we aim to address is

> *(RQ1)* To what extent do female and male students differ in their *average item nonresponse propensity* on the PISA 2018 student background questionnaire?

We further hypothesise that given the wide range of the PISA 2018 student background questionnaire scales, the potential gender gap in item nonresponse might be wider on some scales and narrower on others, leading to our second research question

> *(RQ2)* Do *certain scales* on the PISA 2018 student questionnaire *elicit larger or smaller gender differences* than the average gender gap in item nonresponse propensity?

Furthermore, we suspect the potential variation in gender gaps in item nonresponse across scales *(RQ2)* to be related to scale characteristics. The evidence from the item nonresponse literature highlights the scale response format and content as likely relevant moderators of the gender difference in item nonresponse propensity. However, it is less clear what to expect with respect to the directionality and strength of these associations in the student questionnaires (i.e., which format and content may elicit more item nonresponse for which gender). Hence, we extend our second research question to examine whether certain *scale response formats (RQ2a)* and *scale contents (RQ2b)* elicit a *larger or smaller item nonresponse gender difference* than the average item nonresponse gender gap, and if so, for which gender?

To address our research questions, we adopt a mixed modelling framework where we allow for individual differences in item nonresponse across students (accounting, among others, for differences in baseline item nonresponse propensity and item nonresponse trends throughout the questionnaire) and scales (mapping differential tendencies in nonresponse to specific scale formats and contents). Meta-analytical tools such as forest plots and confidence and prediction intervals for pooled estimates are used to summarise our model-based results across 80 countries and 71 scales.

2 Method

2.1 PISA 2018

The overarching objective of PISA is to assess student learning outcomes, including student achievement (measured by cognitive tests) and learning contexts (measured by questionnaires administered to students, parents, teachers, and principals). The latest PISA 2018 student questionnaire was administered to each student approximately two hours into the assessment after the students had completed the achievement tests. The administration mode (computer- or paper-based) did not vary within individual countries and had no bearing on the items administered or the time allocated to respond to the questionnaire (i.e., 35 min; OECD, 2020).

2.1.1 PISA 2018 participants

The study sample included more than 612,000 randomly sampled 15-year-old students from 80 countries and jurisdictions. PISA uses a stratified two-stage sample design, where schools are first sampled proportional to their size, then students are sampled with equal probability within the school. For each country, PISA 2018 prescribed, when feasible, to sample 5200–6300 students (OECD, 2020). The sample sizes ranged from 2016 students in the Moscow Region (Russia) to 35,943 in Spain. The country-specific sample sizes can be found in Fig. 1. Female and male students were equally represented in the student samples of the participating countries.

2.1.2 PISA 2018 student questionnaire

The PISA 2018 student questionnaire included 75 scales totalling 306 items. In this study, the term “scale” refers to item(s) that measure a unique variable or cover a unique (sub)construct, including both multi-item scales (e.g., 6-item scale ST034 assessing sense of belonging to school, part of the General Schooling constructs) and single-item scales (e.g., a 1-item scale ST175 inquiring about the time spent on reading for enjoyment, part of Student Background Reading constructs; Table 1, Appendix 5).

The questionnaire sought background information on the students, their families and households, attitudes and beliefs, learning strategies, and in- and out-of-school experiences (OECD, 2019). Several scales did not have cases of item nonresponse as they were cross-referenced with the student tracking forms filled in by the PISA or school officials (i.e., ST001 "Grade", ST003 "Date of Birth", ST004 "Gender", country-specific part of ST011 "Household Possessions", ST225 "Expected level of completed education"; OECD, 2020). The items belonging to said scales were excluded from further analyses.

Of the remaining 71 scales (293 items), 15 scales consisted of one item (e.g., ST005, ST007), 14 scales of 3 items (e.g., ST019, ST127), 15 scales of 4 items (e.g., ST006, ST008), ten scales of 5 items (e.g., ST023, ST097), eight scales of 6 items (e.g., ST176, ST161), three scales of 7 items (e.g., ST158, ST197), two scales of 8 items (ST012, ST222), two scales of 9 items (ST152, ST186), one scale of 10 items (ST221), and one scale of 13 items (ST011). The items within individual scales followed the same response format and assessed the same content area. Several countries did not administer all the scales (e.g., in Lebanon, only a field-trial version of the student background questionnaire was administered, resulting in substantial deviations from the number of scales administered in other countries; OECD, 2020). Figure 1 lists the number of scales administered in each country.

2.1.3 Covariates

Gender A categorical variable for gender was defined based on the ST004 cross-referenced “Gender” scale. Gender female was coded as 0 (i.e., the reference group), and gender male was coded as 1.

Item position Item position within the questionnaire was considered an additional covariate to account for the likely increase in item nonresponse incidence towards the end of the questionnaire. For a country having administered all items, item position ran from the value 1 to the value 306. Each item retained its original rank position in the country-specific student background questionnaire, that is, before exclusion of the aforementioned cross-referenced items that had cases of item nonresponse but after exclusion of the items that were not administered in that country. In further analyses, item position was centred and re-scaled such that one unit corresponded to 50 items, and negative and positive values corresponded to items before and after the middle of the questionnaire, respectively.

Scale response format The scale response format was defined as a categorical variable with five levels. The *dichotomous* format (8 scales) included scales that required students to choose between two given response options. The *multiple-choice* format (7 scales) was represented by scales permitting the choice of one option out of more than two unordered response options. Also coded as multiple-choice were the items with multiple response options which followed a natural order but could not be meaningfully interpreted as agreement or frequency (e.g., ST021Q01TA; OECD, 2019). The remaining scales with ordered response categories were coded as *Likert agreement* (31 scales) and *Likert frequency* (23 scales). Two scales that did not provide response options were coded as *open-response*. Table 1 (Appendix 5) lists response formats for each individual scale. All items within a scale share the same response format.

Scale content The scale content was defined as a categorical variable with six levels (six content areas) closely corresponding to the overarching PISA 2018 questionnaire framework (OECD, 2019). The PISA 2018 framework consisted of three families of constructs covered by multiple modules comprised of multiple scales. These construct families were (1) *student background* constructs where students were asked about their family background, the education they have received, and their out-of-school learning experiences; (2) *schooling* constructs where educational processes were assessed at the school and classroom level; and (3) *non-cognitive and meta-cognitive* constructs such as dispositions for global competence and overall strategy of awareness (OECD, 2019). Each of the three construct families was further divided into two categories of scales: (a) those that covered *general* topics and (b) those that contained *reading*-related topics (since reading literacy was the major cognitive domain in the PISA 2018). The six category levels of scale content are formed by crossing construct family by category (e.g., student background:

general, student background: reading; see Table 1 in Appendix 5). All items within a scale address the same scale content area.

2.2 Statistical analysis

2.2.1 Preliminary descriptive analysis

Average item nonresponse rates were obtained to gauge the occurrence of item nonresponse per country across scales and per scale across countries. From the country perspective, this implied averaging individual students' item nonresponse rates (i.e., the ratio of the number of items on the questionnaire to which the student did not respond to the number of theoretically valid responses possible on the questionnaire administered within a country) to the country level. From the scale perspective, individual items' nonresponse rates are first averaged by scale in each country and then averaged across countries for each scale. The latter informed the formulation of the working model, to which we return later.

2.2.2 Cross-classified mixed effects model

To relate gender and item nonresponse on the PISA 2018 student background questionnaire, a cross-classified logistic regression model was formulated within a mixed modelling framework (De Boeck & Wilson, 2004; Van den Noortgate et al., 2003). The binary outcome variable $Y_{ps(i)}$ took value 1 when student p did not provide a response to item i of scale s (in case of a valid response, $Y_{ps(i)} = 0$). To account for the data dependence among nonresponses due to the cross-classified persons-by-items-within-scales study design, the probability of nonresponse $\pi_{ps(i)} = \Pr(Y_{ps(i)} = 1)$ on a given item of a given scale by a given person was modelled as a logistic function of the sum of a general intercept β_0, a person-specific deviation θ_p, and a scale-specific deviation β_s.[1] Whereas the general intercept is a fixed effect, the two deviation coefficients were modelled as normally distributed random effects with means of zero and variances $\omega^2_{\theta_p}$ and $\sigma^2_{\beta_s}$, respectively. The deviation coefficients reflected that students varied in their nonresponse propensity (θ_p), and scales varied in the extent to which they elicited nonresponse (β_s).

$$\mathrm{Logit}\left(\pi_{ps(i)}\right) = \beta_0 + \theta_p + \beta_s$$

Given the low-stakes assessment context, a general concern was that the items positioned towards the end of the questionnaire would be more prone to nonresponse due to potentially growing student fatigue and consequent decline in test-taking effort. If not accounted for, such position trends could confound our gender-related findings. Hence, the model incorporated an item-level predictor POSITION$_i$ with a random slope θ_{p1} that varied across students (to reflect potential individual

[1] Given that the items within each scale are of the same format and address the same content area, we make abstraction of item-specific deviations.

differences in position trends). The varying slope θ_{p1} is the sum of the fixed effect θ_1, the average student's item nonresponse propensity as a function of item position, and the random effect δ_{p1}, the individual students' deviations from the average student's item nonresponse propensity as a function of item position. Student's nonresponse propensity was then modelled as $\theta_p = \theta_{p0} + \theta_{p1}\mathrm{POSITION}_i$. Because $\mathrm{POSITION}_i$ is centred, the person-specific intercept θ_{p0} reflected the expected item nonresponse propensity (on the logit scale) of student p on an item from an average scale located in the middle of the questionnaire. The person-varying intercept and slope were allowed to correlate as

$$\begin{bmatrix} \theta_{p0} \\ \theta_{p1} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ \theta_1 \end{bmatrix}, \begin{bmatrix} \omega_0^2 & \omega_{01} \\ & \omega_1^2 \end{bmatrix} \right)$$

Preliminary descriptive analyses further indicated that the items belonging to the ST006 and ST008 scales showed exceptionally high nonresponse rates resulting in peaks in the overall position trends. The peaks in item nonresponse rates were accounted for by including a covariate X_s, coded for the ST006 and ST008 scales as 1 and zero otherwise, as an extra fixed effect θ_2 (see Eq. 1).

To operationalise our research questions, we included a person-level predictor GENDER_p (coded as male = 1, female = 0) and allowed its coefficient β_{s1} to vary by scale. Item nonresponse propensity was then modelled as $\beta_s = \beta_{s0} + \beta_{s1}\mathrm{GENDER}_p$. The coefficient β_{s0} corresponded to the scale-specific nonresponse propensity for the female reference group, and β_{s1} corresponded to the scale-specific gender difference in nonresponse propensity for items of scale s. The scale-varying intercept and slope were allowed to correlate as

$$\begin{bmatrix} \beta_{s0} \\ \beta_{s1} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ & \sigma_1^2 \end{bmatrix} \right)$$

The mean β_1 of this scale-specific gender coefficient reflected the expected difference in item nonresponse propensity on the logit scale between male and female students on an item of an average scale of the PISA 2018 student questionnaire (henceforth, "average gender difference in item nonresponse propensity"). In addition to this average gender difference in item nonresponse propensity, the coefficient $\delta_{s1} = \beta_{s1} - \beta_1$ represented the scale-specific incremental gender difference in item nonresponse propensity. The δ_{s1} parameter reflected what is commonly known in the item response theory framework as differential item functioning (e.g., De Boeck & Cho, 2021), and indicated whether students with the same average item nonresponse propensity but of different gender were more or less likely not to provide a valid response on an item of a specific scale. Worded more colloquially, the overall gender gap in item nonresponse would narrow or widen on a specific scale. Hence, large values of the parameter δ_{s1} would imply a differential item nonresponse gender bias that could be potentially related to scale characteristics, such as response format or content.

Assembled, the terms introduced in the previous paragraphs constitute the working model

$$\text{Logit}\left(\pi_{ps(i)}\right) = \beta_0 + \theta_{p0} + \overbrace{\left(\theta_1 + \delta_{p1}\right)}^{\theta_{p1}} \text{POSITION}_i + \theta_2 \text{X}_s + \beta_{s0} + \overbrace{\left(\beta_1 + \delta_{s1}\right)}^{\beta_{s1}} \text{GENDER}_p, \tag{1}$$

where item nonresponse is modelled as a function of an interaction between person and scale characteristics with a person-level predictor GENDER$_p$ varying by scale s and an item-level predictor POSITION$_i$ varying by student p.

2.2.3 Estimation approach

The cross-classified mixed effects model defined in Eq. 1 was independently estimated for each country, resulting in a total of 80 sets of model estimates. The analysis was performed using the *lme4* package (Bates et al., 2015) in version 4.2.0 of the R software environment (R Core Team, 2020). Full-information maximum likelihood with bound optimisation by quadratic approximation was used for model estimation, with the maximal number of iterations set to 200,000. The procedural steps with an example model script are presented in Appendix 7.

2.2.4 Representation of results

The cross-classified mixed effects models' parameter estimates for (i) the average gender differences in item nonresponse propensity (i.e., the parameter β_1 in Eq. 1; one for each country) and (ii) the scale-specific incremental gender differences in item nonresponse propensity (i.e., the parameter δ_{s1} in Eq. 1; 40–71 estimates per country) were extracted and summarised in forest plots on the basis of random effects meta-analytical models (Borenstein et al., 2009) using the *metafor* package (Viechtbauer, 2010).

To address *(RQ1)*, the 80 β_1 parameter estimates for the average gender differences in item nonresponse propensity were summarised across countries. Individual countries were treated as independent studies or units of the analysis. One pooled estimate $\overline{\beta}_1$ across countries was obtained, accompanied by its confidence and prediction intervals to reflect the precision of the estimated effect in the population and the dispersion around the average in the population, respectively. To address *(RQ2)*, the parameter estimates for the scale-specific incremental gender differences in item nonresponse propensity (δ_{s1}) were summarised across countries for each individual scale, resulting in 71 pooled estimates $\overline{\delta}_{s1}$ (one for each scale s).

To address *(RQ2a-2b)*, a series of subgroup meta-analyses were performed to summarise the parameter estimates for the scale-specific incremental gender differences in item nonresponse propensity (δ_{s1}) for each country as a function of scale format and contents. For each level of the categorical covariates, an estimate δ^*_{s1} was obtained in each country. In other words, per country, we obtained five estimates for different scale response formats and six estimates for different scale contents (5 response formats × 80 countries + 6 content areas × 80 countries = 880 estimates in total). Henceforth, we denote a parameter estimate with an asterisk (e.g., δ^*_{s1}) when the parameter was obtained in the subgroup meta-analyses and pertains to the level of covariate that is being discussed (e.g., the incremental gender differences in item nonresponse propensity on an open-response format scale). The final meta-analyses combined the estimates for each level of both covariates across 80 countries, resulting in one pooled estimate $\overline{\delta}^{*}_{s1}$ for each level of each covariate (i.e., 5 response formats × 1 + 6 content areas × 1 = 11 pooled estimates).

The parameter estimates are presented on the original logit scale in all the figures. For ease of interpretation as effect size measure, both country-wise parameters and pooled parameter estimates were expressed as odds ratios (OR = exp(parameter)) with the variables to which they pertain between parentheses (e.g., OR(POSITION), OR(GENDER)). Depending on the comparison made (male versus female or vice versa), we used the shorthand OR(M) and OR(F) for OR(GENDER = M) and OR(GENDER = F), respectively.

3 Results

In most countries that participated in the PISA 2018 student questionnaire, average item nonresponse rates were between 2 and 7% (Fig. 3, Appendix 1). Near-zero average item nonresponse rates were observed in the Chinese provinces of Beijing, Shanghai, Jiangsu, and Zhejiang (henceforth, "B-S-J-Z (China)") and Macao, whereas in the Dominican Republic, Morocco, and Baku (Azerbaijan), average item nonresponse rates of 20% occurred. When averaging across countries and looking at item nonresponse rates from the individual scales' perspective (Fig. 4, Appendix 1), higher item nonresponse rates could be observed on the scales located towards the end of the questionnaire (e.g., from average nonresponse rates of roughly 1–2% at the beginning of the questionnaire to 10% at the end). However, notice that two scales located at the very beginning of the questionnaire formed an exception to the rule with item nonresponse rates over 13%. The scales in question, ST006 and ST008, inquired about the level of students' parents' education.

The cross-classified mixed effects models' estimates further suggested that, on average across 80 countries, when comparing students of the same gender and with the same item nonresponse propensity, the odds of item nonresponse on the ST006 and ST008 scales' items were 26 times greater than the odds of item nonresponse on the other items with a similar position in the questionnaire (i.e., the median $\theta_2 = 3.26$, OR(X_s) = exp(3.26) = 26.05 and the mean $\theta_2 = 3.80$, OR(X_s) = exp(3.80) = 44.70; Fig. 6, Appendix 3). In most countries, the students were also more likely to omit an item the further in the questionnaire said item was located. On average across 80 countries, the odds of item nonresponse for each subsequent 50 items of the PISA student questionnaire close to doubled (i.e., the median $\theta_1 = 0.34$, OR(POSITION) = exp(0.34) = 1.40 and the mean $\theta_1 = 0.70$, OR(POSITION) = exp(0.70) = 2.01; Fig. 5, Appendix 2). All the results (i.e., model parameters and interpretations of estimates) that we report henceforth were effectively adjusted for person-specific item position trends.

3.1 Average gender differences in item nonresponse propensity

Overwhelmingly, across 80 countries, male students had significantly greater odds of not responding to an item of an average scale in the PISA 2018 student questionnaire compared to female students (see a meta-analysis forest plot of the parameter estimates β_1 from Eq. 1 in Fig. 1). On average across countries, the odds of male students not responding were roughly double ($\overline{\beta}_1 = 0.69[0.60, 0.78]$, $\mathrm{OR(M)} = \exp(0.69) = 1.99$) that of female students. The width of the prediction interval in Fig. 1 implied heterogeneity in effect size across countries (i.e., the wider the prediction interval, the larger differences in effect size can be observed across countries).

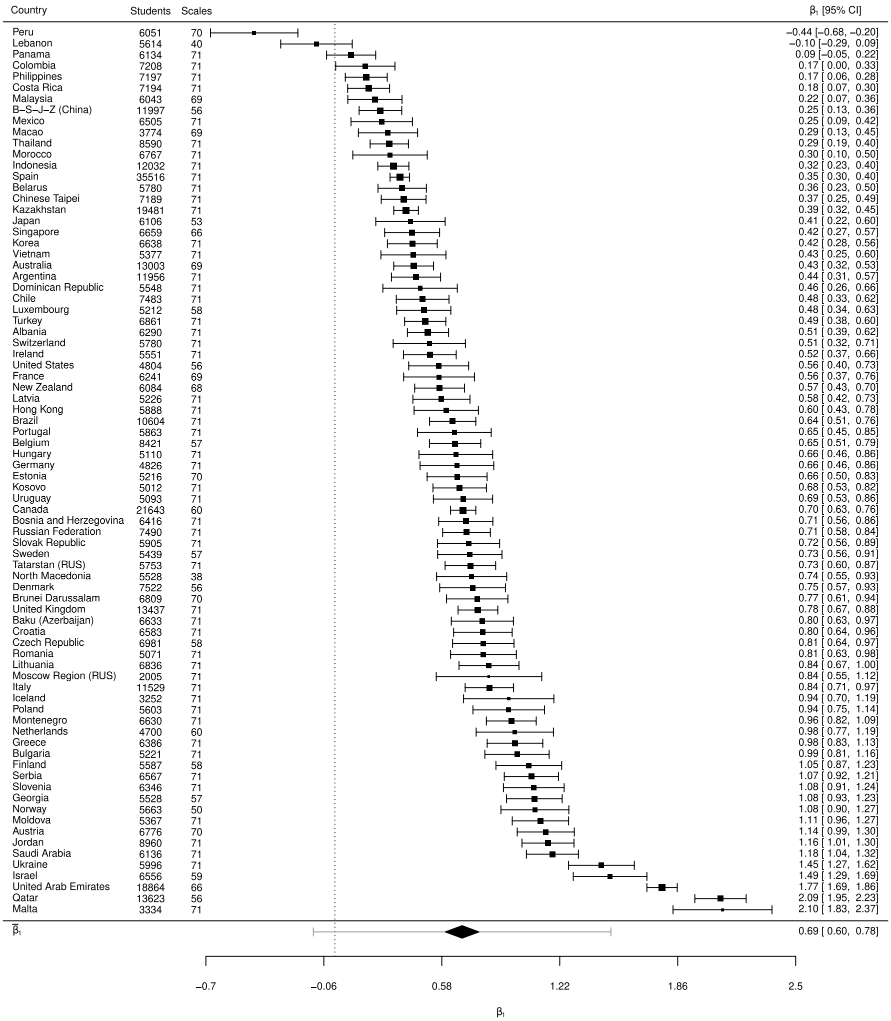
In the Philippines, for instance, the odds of male students not responding to an average item were only slightly greater than that of female students ($\beta_1 = 0.17[0.06, 0.28]$, $\mathrm{OR(M)} = 1.18$; Fig. 1). Naturally, small differences in the odds of item nonresponse between genders were also found in countries with very low average item nonresponse rates (e.g., $\beta_1 = 0.25[0.13, 0.36]$, $\mathrm{OR(M)} = 1.28$ in B-S-J-Z (China) with average item nonresponse rate of 0.3%; Fig. 3, Appendix 1).

Considerably more substantial gender gaps were found, for instance, in Malta ($\beta_1 = 2.10[1.83, 2.37]$, $\mathrm{OR(M)} = 8.19$) and Qatar ($\beta_1 = 2.09[1.95, 2.23]$, $\mathrm{OR(M)} = 8.10$), where the odds of item nonresponse on an item of the average PISA 2018 questionnaire scale for male students were eight times greater than that of the female students. Peru was the only of 80 countries where the odds of not responding were reversed between genders, with the odds of item nonresponse of female students being one and a half times greater than that of male students ($\beta_1 = -0.44[-0.68, -0.20]$, $\mathrm{OR(F)} = 1/\exp(-0.44) = 1.56$; Fig. 1).

3.2 Scale-specific incremental gender differences in item nonresponse propensity

Figure 7 (Appendix 4) presents the scale-specific incremental gender differences in item nonresponse propensity across countries ($\overline{\delta}_{s1}$). Recall that, on average, male students had roughly double the odds of female students not responding to an item from an average scale on the PISA 2018 student questionnaire ($\overline{\beta}_1$ in Fig. 1). If we consider the scale-specific incremental gender differences in item nonresponse propensity in conjunction with the average gender difference in item nonresponse propensity, the overall gender gap in item nonresponse was significantly narrower on 28 scales of the PISA 2018 questionnaire (i.e., $\overline{\delta}_{s1} < 0$; Fig. 7). For instance, the gender gap on the scale ST008 inquiring about students' fathers' qualifications nearly closed such that the odds of item nonresponse on an average item of this scale for female and male students became nearly even ($\left(\overline{\beta}_1 + \overline{\delta}_{s1}\right)\mathrm{GENDER(M)} = 0.69 - 0.48 = 0.21$; $\mathrm{OR(M)}=\exp(0.21) = 1.23$).

On the other hand, the overall gender gap was significantly wider on 26 scales (i.e., $\overline{\delta}_{s1} > 0$; Fig. 7) such that male students had greater odds of not responding to an average item of these scales than female students. For example, the odds of male students not providing a response to an average item of the scale ST186 assessing subjective well-being were 2.5 times that of female students



Country	Students	Scales	β₁ [95% CI]
Peru	6051	70	−0.44 [−0.68, −0.20]
Lebanon	5614	40	−0.10 [−0.29, 0.09]
Panama	6134	71	0.09 [−0.05, 0.22]
Colombia	7208	71	0.17 [0.00, 0.33]
Philippines	7197	71	0.17 [0.06, 0.28]
Costa Rica	7194	71	0.18 [0.07, 0.30]
Malaysia	6043	69	0.22 [0.07, 0.36]
B–S–J–Z (China)	11997	56	0.25 [0.13, 0.36]
Mexico	6505	71	0.25 [0.09, 0.42]
Macao	3774	69	0.29 [0.13, 0.45]
Thailand	8590	71	0.29 [0.19, 0.40]
Morocco	6767	71	0.30 [0.10, 0.50]
Indonesia	12032	71	0.32 [0.23, 0.40]
Spain	35516	71	0.35 [0.30, 0.40]
Belarus	5780	71	0.36 [0.23, 0.50]
Chinese Taipei	7189	71	0.37 [0.25, 0.49]
Kazakhstan	19481	71	0.39 [0.32, 0.45]
Japan	6106	53	0.41 [0.22, 0.60]
Singapore	6659	66	0.42 [0.27, 0.57]
Korea	6638	71	0.42 [0.28, 0.56]
Vietnam	5377	71	0.43 [0.25, 0.60]
Australia	13003	69	0.43 [0.32, 0.53]
Argentina	11956	71	0.44 [0.31, 0.57]
Dominican Republic	5548	71	0.46 [0.26, 0.66]
Chile	7483	71	0.48 [0.33, 0.62]
Luxembourg	5212	58	0.48 [0.34, 0.63]
Turkey	6861	71	0.49 [0.38, 0.60]
Albania	6290	71	0.51 [0.39, 0.62]
Switzerland	5780	71	0.51 [0.32, 0.71]
Ireland	5551	71	0.52 [0.37, 0.66]
United States	4804	56	0.56 [0.40, 0.73]
France	6241	69	0.56 [0.37, 0.76]
New Zealand	6084	68	0.57 [0.43, 0.70]
Latvia	5226	71	0.58 [0.42, 0.73]
Hong Kong	5888	71	0.60 [0.43, 0.78]
Brazil	10604	71	0.64 [0.51, 0.76]
Portugal	5863	71	0.65 [0.45, 0.85]
Belgium	8421	57	0.65 [0.51, 0.79]
Hungary	5110	71	0.66 [0.46, 0.86]
Germany	4826	71	0.66 [0.46, 0.86]
Estonia	5216	70	0.66 [0.50, 0.83]
Kosovo	5012	71	0.68 [0.53, 0.82]
Uruguay	5093	71	0.69 [0.53, 0.86]
Canada	21643	60	0.70 [0.63, 0.76]
Bosnia and Herzegovina	6416	71	0.71 [0.56, 0.86]
Russian Federation	7490	71	0.71 [0.58, 0.84]
Slovak Republic	5905	71	0.72 [0.56, 0.89]
Sweden	5439	57	0.73 [0.56, 0.91]
Tatarstan (RUS)	5753	71	0.73 [0.60, 0.87]
North Macedonia	5528	38	0.74 [0.55, 0.93]
Denmark	7522	56	0.75 [0.57, 0.93]
Brunei Darussalam	6809	70	0.77 [0.61, 0.94]
United Kingdom	13437	71	0.78 [0.67, 0.88]
Baku (Azerbaijan)	6633	71	0.80 [0.63, 0.97]
Croatia	6583	71	0.80 [0.64, 0.96]
Czech Republic	6981	58	0.81 [0.64, 0.97]
Romania	5071	71	0.81 [0.63, 0.98]
Lithuania	6836	71	0.84 [0.67, 1.00]
Moscow Region (RUS)	2005	71	0.84 [0.55, 1.12]
Italy	11529	71	0.84 [0.71, 0.97]
Iceland	3252	71	0.94 [0.70, 1.19]
Poland	5603	71	0.94 [0.75, 1.14]
Montenegro	6630	71	0.96 [0.82, 1.09]
Netherlands	4700	60	0.98 [0.77, 1.19]
Greece	6386	71	0.98 [0.83, 1.13]
Bulgaria	5221	71	0.99 [0.81, 1.16]
Finland	5587	58	1.05 [0.87, 1.23]
Serbia	6567	71	1.07 [0.92, 1.21]
Slovenia	6346	71	1.08 [0.91, 1.24]
Georgia	5528	57	1.08 [0.93, 1.23]
Norway	5663	50	1.08 [0.90, 1.27]
Moldova	5367	71	1.11 [0.96, 1.27]
Austria	6776	70	1.14 [0.99, 1.30]
Jordan	8960	71	1.16 [1.01, 1.30]
Saudi Arabia	6136	71	1.18 [1.04, 1.32]
Ukraine	5996	71	1.45 [1.27, 1.62]
Israel	6556	59	1.49 [1.29, 1.69]
United Arab Emirates	18864	66	1.77 [1.69, 1.86]
Qatar	13623	56	2.09 [1.95, 2.23]
Malta	3334	71	2.10 [1.83, 2.37]
$\overline{\beta}_1$			0.69 [0.60, 0.78]

Fig. 1 Average gender differences in item nonresponse propensity (female students as the reference group). *Note*. Countries' parameter estimates β_1 (see Eq. 1) are presented on the logit scale and correspond to the cross-classified mixed effects models' estimates of the expected differences in item nonresponse propensity on an item of an average scale of the PISA 2018 student background questionnaire for male students, as compared to female students. Positive and negative parameter estimates correspond to greater odds of item nonresponse on an item of an average scale on the PISA 2018 student questionnaire for male and female students, respectively. At the grey dashed line at 0, neither gender has greater odds of item nonresponse. The random effects meta-analytical model results are presented at the bottom of the graph. The black diamond shape corresponds to the 95% confidence interval of $\overline{\beta_1}$, the pooled average across countries, and the bars around it map the corresponding prediction interval as an indication of dispersion around this average.

$\left(\left(\overline{\beta}_1 + \overline{\delta}_{s1}\right)\mathrm{GENDER(M)} = 0.69 + 0.21 = .90; \mathrm{OR(M)} = \exp(0.90) = 2.46\right)$. The adjusted gender gaps for all scales can be derived analogously using the model estimates $\overline{\delta}_{s1}$ provided in Fig. 7 and cross-referencing the results with the scale contents (Table 1, Appendix 5). In the following, we present the findings of the moderator analyses, in which we systematically relate the aforementioned scale-specific incremental gender differences in item nonresponse to the scale format and scale content.

3.2.1 Scale response format

Open-response format The PISA 2018 student background questionnaire included two scales that followed open-response format. On average across 80 countries, male students had significantly greater odds of item nonresponse on an open-response scale, compared to female students with the same item nonresponse propensity ($\overline{\delta}^*_{s1} = 0.18[0.11, 0.24]$, $\mathrm{OR(M)}^* = 1.20$; Fig. 2). A wide prediction interval relative to the confidence interval around the open-response format estimate (i.e., the bars extending from the diamond shape in Fig. 2) implied substantial differences in effect size across countries. For instance, the odds of male students omitting an open-response scale were double the odds of female students with the same item nonresponse propensity in North Macedonia ($\delta^*_{s1} = 0.84[0.42, 1.27]$, $\mathrm{OR(M)}^* = 2.32$);

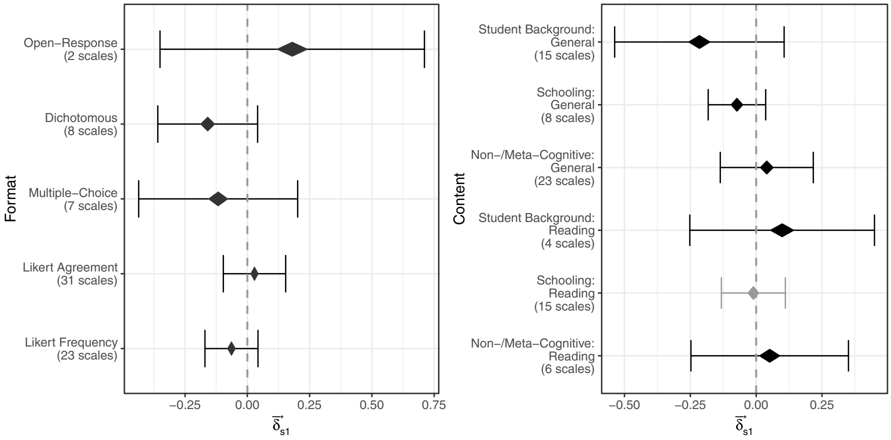


Fig. 2 Scale-specific incremental gender differences in item nonresponse propensity as a function of the scale response format and the scale content (female students as the reference group). *Note.* The results are reported for five scale formats and six content areas. The estimates $\overline{\delta}^*_{s1}$ are presented on the logit scale and correspond to the results of the final meta-analyses that combined the estimates for each level of both covariates (i.e., format and content) across 80 countries, resulting in one pooled estimate $\overline{\delta}^*_{s1}$ for each level of each covariate. The diamond shapes represent the confidence intervals around the pooled estimate for each format and content. The bars around the diamond define the corresponding prediction interval to indicate the dispersion around this average. When the estimate is significantly different from zero, the diamond shape and bars are black, otherwise, grey. When comparing female and male students with the same item nonresponse propensity, positive and negative estimates correspond to greater odds of item nonresponse for male and female students, respectively. At the grey dashed line at 0, neither male nor female students have greater odds of item nonresponse.

nearly triple in Georgia ($\delta^*_{s1} = 0.99[0.84, 1.15]$, $\mathrm{OR(M)}^* = 2.69$); and quadruple in Lebanon ($\delta^*_{s1} = 1.46[1.15;1.77]$, $\mathrm{OR(M)}^* = 4.31$; Fig. 8, Appendix 6). In Denmark, on the other hand, the odds of not responding on an open-response scale were greater for female students, compared to male students with the same item nonresponse propensity ($\delta^*_{s1} = -0.22[-0.43, -0.02]$, $\mathrm{OR(F)}^* = 1.25$; Fig. 8, Appendix 6).

Dichotomous format The PISA 2018 student background questionnaire included eight scales that followed dichotomous response format. On average across 80 countries, female students had significantly greater odds of item nonresponse on a dichotomous scale, compared to male students with the same item nonresponse propensity ($\overline{\delta}^*_{s1} = -0.16[-0.19, -0.13]$, $\mathrm{OR(F)}^* = 1.17$; Fig. 2). The corresponding prediction interval (Fig. 2) suggested there were fewer differences in effect size across countries, than those previously observed for the open-response format. The odds of item nonresponse for female students, compared to male students with the same propensity of item nonresponse, ranged from slightly over one in, for example, Australia ($\delta^*_{s1} = -0.08[-0.15, -0.01]$, $\mathrm{OR(F)}^* = 1.08$) to roughly triple the odds of male students in Qatar ($\delta^*_{s1} = -1.00[-1.58, -0.42]$, $\mathrm{OR(F)}^* = 2.72$; Fig. 8, Appendix 6). The only country where male students had significantly greater odds of omitting a dichotomous scale, compared to the female students with the same item nonresponse propensity, was Peru ($\delta^*_{s1} = 0.18[0.12, 0.24]$, $\mathrm{OR(M)}^* = 1.19$; Fig. 8, Appendix 6).

Multiple-choice format The PISA 2018 student background questionnaire included seven scales that followed multiple-choice response format. On average across 80 countries, female students had significantly greater odds of item nonresponse on a multiple-choice scale, compared to male students with the same item nonresponse propensity ($\overline{\delta}^*_{s1} = -0.12[-0.16, -0.08]$, $\mathrm{OR(F)}^* = 1.13$; Fig. 2). A wide prediction interval (Fig. 2) implied substantial differences in effect size across countries. For example, when comparing female and male students with the same item nonresponse propensity, the odds of female students not responding to a multiple-choice scale were approximately double that of male students in Qatar ($\delta^*_{s1} = -0.75[-1.22, -0.28]$, $\mathrm{OR(F)}^* = 2.12$; Fig. 8, Appendix 6). In contrast, male students were more likely to omit a multiple-choice scale, compared to female students with the same item nonresponse propensity, in Morocco ($\delta^*_{s1} = 0.28[0.16, 0.40]$, $\mathrm{OR(M)}^* = 1.32$; Fig. 8, Appendix 6).

Likert agreement format The PISA 2018 student background questionnaire included 31 scales that followed Likert agreement response format. On average across 80 countries, male students had significantly greater odds of item nonresponse on a Likert agreement scale, compared to female students with the same item nonresponse propensity ($\overline{\delta}^*_{s1} = 0.03[0.01, 0.04]$, $\mathrm{OR} = 1.03$; Fig. 2). The prediction interval (Fig. 2) implied heterogeneity in effect size across countries. For instance, the odds of male students omitting a Likert agreement scale were roughly 1.2 times greater than that of female students in Croatia ($\delta^*_{s1} = 0.22[0.13, 0.30]$, $\mathrm{OR(M)}^* = 1.24$; Fig. 8, Appendix 6). In contrast, female students had greater odds of omitting a Likert agreement scale, compared to male students with the same item nonresponse propensity (e.g., $\delta^*_{s1} = -0.14[-0.22, -0.06]$, $\mathrm{OR(F)}^* = 1.15$ in Lithuania; Fig. 8, Appendix 6).

Likert frequency format The PISA 2018 student background questionnaire included 23 scales that followed Likert frequency response format. On average across 80 countries, female students had significantly greater odds of item nonresponse on a Likert frequency scale, compared to male students with the same item nonresponse propensity ($\overline{\delta}^*_{s1} = -0.06[-0.08, -0.05]$, OR(F)$^* = 1.06$; Fig. 2). The prediction interval (Fig. 2) suggested that the odds of item nonresponse on Likert frequency scales was greater for female students in most countries. In fact, compared to female students with the same item nonresponse propensity, male students had greater odds of omitting a Likert frequency scale only in 17 countries, and none of these estimates was significant (Fig. 8, Appendix 6). In the remaining countries, the significant odds of item nonresponse for female students, compared to male students with the same item nonresponse propensity, ranged from, for example, just over one in the Slovak Republic ($\delta^*_{s1} = -0.09[-0.15, -0.02]$, OR(F)$^* = 1.09$) to roughly one and a half times the odds of male students in the United Arab Emirates ($\delta^*_{s1} = -0.38[-0.56, -0.19]$, OR(F)$^* = 1.46$; Fig. 8, Appendix 6).

3.2.2 Scale content

Student background constructs The PISA 2018 student questionnaire included 15 scales gathering information on students' *general* background (i.e., student background: general; Fig. 2). These scales inquired about students' families (e.g., students' mothers' and fathers' educational and professional lives), socioeconomic status (e.g., home possessions, number of books in students' households), the ethnic background of students and their parents, and students' early educational pathways. On average across 80 countries, when comparing female and male students with the same item nonresponse propensity, female students had significantly greater odds of omitting a general background scale ($\overline{\delta}^*_{s1} = -0.22[-0.25, -0.18]$, OR(F)$^* = 1.5$; Fig. 2). The prediction interval (Fig. 2) implied that the odds of female students omitting general background scales were greater than those of male students with the same item nonresponse propensity in most countries (Fig. 2). For instance, the odds of item nonresponse for female students were nearly triple the odds of male students with the same item nonresponse propensity in Qatar ($\delta^*_{s1} = -1.02[-1.32, -0.71]$, OR(F)$^* = 2.77$; Fig. 9, Appendix 6). The only country where male students had significantly greater odds of item nonresponse on the general background scales, compared to female students with the same item nonresponse propensity, was Peru ($\delta^*_{s1} = 0.16[0.07, 0.25]$, OR(M)$^* = 1.18$).

Student background constructs additionally covered several topics to assess out-of-school *reading* experiences (i.e., student background: reading; Fig. 2). The four scales dedicated to these reading-related topics inquired about the time spent reading for enjoyment and students' involvement in various online reading activities (e.g., reading emails, news, and forums). On average across 80 countries, when comparing female and male students with the same item nonresponse propensity, male students had significantly greater odds of item nonresponse on reading-related background scales ($\overline{\delta}^*_{s1} = 0.10[0.05, 0.14]$, OR$=1.10$; Fig. 2). The corresponding prediction interval (Fig. 2) suggested substantial differences in effect size across countries. The odds of male students omitting scales on reading-related background scales were double that of female students with the same item nonresponse propensity in, for example, Iceland ($\delta^*_{s1} = 0.68[0.49, 0.86]$, $\text{OR(M)}^* = 1.97$; Fig. 9, Appendix 6). On the other hand, female students had greater odds of item nonresponse on the reading-related background scales than male students with the same overall item nonresponse propensity in Chinese Taipei ($\delta^*_{s1} = -0.18[-0.31, -0.05]$, $\text{OR(F)}^* = 1.19$; Fig. 9, Appendix 6).

Schooling constructs The PISA 2018 student questionnaire included eight scales assessing *general* schooling constructs (i.e., Schooling: General; Fig. 2). These scales covered the topics of perceived school climate (e.g., scales inquiring about shared school values and norms, students' sense of belonging, and their experiences with bullying), parental involvement (e.g., parents emotional support), and assessment, evaluation, and accountability (e.g., teacher adaptability of instruction to student's needs, and teacher feedback). On average across 80 countries, when comparing female and male students with the same item nonresponse propensity, female students had significantly greater odds of item nonresponse on general schooling scales ($\overline{\delta}^*_{s1} = -0.07[-0.09, -0.05]$, $\text{OR(F)}^* = 1.07$; Fig. 2). The prediction interval (Fig. 2) implied that the odds of item nonresponse on these scales were greater for female students, compared to male students with the same item nonresponse propensity, in most countries. Compared to female students with the same item nonresponse propensity, male students had greater odds of item nonresponse on a schooling scale only in 15 countries, with two significant effects observed in Georgia ($\delta^*_{s1} = 0.28[0.00, 0.56]$, $\text{OR(M)}^* = 1.32$) and Moscow Region (Russia) ($\delta^*_{s1} = 0.23[0.00, 0.41]$, $\text{OR(M)}^* = 1.25$; Fig. 9, Appendix 6). On the other hand, female students had greater odds of item nonresponse on schooling scales than male students with the same item nonresponse propensity in the remaining countries. For instance, female students had significantly greater odds of item nonresponse on a general schooling scale in Ukraine ($\delta^*_{s1} = -0.43[-0.62, -0.25]$, $\text{OR(F)}^* = 1.54$; Fig. 9, Appendix 6).

Schooling constructs additionally covered several topics on students' *reading* experiences (i.e., schooling: reading; Fig. 2) that addressed teaching practices and school environment for reading as well as students' learning time and curriculum (15 scales in total). On average across 80 countries, no support was found for the effect of this content area on the overall gender gap in item nonresponse (i.e., confidence interval containing zero; Fig. 2). However, the prediction interval (Fig. 2) suggested heterogeneity in effect size across countries, such that the countries were almost evenly divided into those where the odds of item nonresponse on these scales were greater for female students and those where the odds were greater for male students. For instance, when comparing female and male students with the same item nonresponse propensity, male students had significantly greater odds of item nonresponse on a reading-related schooling scale in Jordan ($\delta^*_{s1} = 0.16[0.01, 0.32]$, $\text{OR(M)}^* = 1.18$; Fig. 9, Appendix 6). In contrast, compared to male students with the same item nonresponse propensity, female students had significantly greater odds omitting the scales on their school-reading experiences in Ukraine ($\delta^*_{s1} = -0.37[-0.47, -0.27]$, OR(F)* = 1.45; Fig. 9, Appendix 6).

Non-cognitive and meta-cognitive constructs The PISA 2018 student questionnaire included 23 *general* scales covering non-cognitive and meta-cognitive constructs (i.e., non-/meta-cognitive: general; Fig. 2). These scales gathered information on dispositional and school-focused variables (e.g., students' attitudes towards learning, competitiveness, resilience, and fear of failure) and dispositions for global competence (e.g., students' awareness of global issues and their self-efficacy in discussing said issues, students' attitudes towards immigrants, and interests in other cultures). On average across 80 countries, when comparing female and male students with the same item nonresponse propensity, male students had significantly greater odds of item nonresponse on general non- and meta-cognitive scales ($\overline{\delta}^*_{s1} = 0.04[0.02, 0.06]$, OR(M)* = 1.04; Fig. 2). The prediction interval (Fig. 2) suggested heterogeneity in effect size across countries. For instance, compared to female students with the same item nonresponse propensity, male students had significantly greater odds of omitting general non- and meta-cognitive in Qatar ($\delta^*_{s1} = 0.39[0.30, 0.47]$, OR(M)* = 1.47; Fig. 9, Appendix 6). In contrast, female students had greater odds of item nonresponse, compared to male students with the same item nonresponse propensity, in Latvia ($\delta^*_{s1} = -0.16[-0.25, -0.06]$, OR(F)* = 1.17; Fig. 9, Appendix 6).

Non-cognitive and meta-cognitive constructs additionally included six scales inquiring about students' *reading* attitudes (i.e., non-/meta-cognitive: reading; Fig. 2). These scales assessed students' enjoyment and self-concept of reading, as well as students' perception of the difficulty of the PISA test. On average across 80 countries, when comparing female and male students with the same item nonresponse propensity, male students had significantly greater odds of item nonresponse on the reading-related non- and meta-cognitive scales ($\overline{\delta}^*_{s1} = 0.05[0.01, 0.09]$, OR(M)* = 1.05; Fig. 2). The prediction interval (Fig. 2) pointed to substantial differences in effect size across countries. For example, greater odds of item nonresponse for male students, compared to female students with the same item nonresponse propensity, were observed in Morocco ($\delta^*_{s1} = 0.49[0.34, 0.63]$, OR(M)* = 1.63; Fig. 9, Appendix 6). On the other hand, female students had greater odds of omitting a reading-related non- and meta-cognitive scales, compared to male students with the same item nonresponse propensity, in Vietnam ($\delta^*_{s1} = -0.42[-0.56, -0.29]$, OR(F)* = 1.52; Fig. 9, Appendix 6).

4 Discussion

Gender differences in item nonresponse — when the respondents of one gender are more prone to not respond to specific items than the respondents of the opposite gender — are problematic as they can hinder the quality of the data, reduce representativeness of the sample, and most importantly, bias the results and inferences drawn from the questionnaire scales (see, e.g., Meinck et al., 2017). Educational research on high-stakes achievement testing has generated a large body of empirical evidence showing higher item nonresponse rates in these testing situations for female students compared to their male counterparts (Gafni & Melamed, 1994; Grandy, 1987; Matters & Burnett, 1999). In contrast, compared to male students, female students have been found to omit fewer items in low-stakes achievement settings, due to their generally higher motivation and test-taking effort (Costa et al., 2001; DeMars et al., 2013). However, our knowledge of gender differences in item nonresponse is lacking when it comes to the low-stakes student questionnaires, often used in primary and secondary research to contextualise achievement scores obtained in international large-scale assessments.

The present study examined gender differences in item nonresponse propensity on the latest PISA 2018 student questionnaire across 80 countries and 71 scales, encompassing 612,000 students and 293 items, respectively. For each country, we used a cross-classified mixed effects model to quantify (i) the average gender differences in item nonresponse propensity and (ii) the scale-specific incremental gender differences in item nonresponse propensity, potentially narrowing or widening the overall gender gaps on specific scales. To account for student effort and persistence differences throughout the questionnaire, the model included item position as a predictor. Meta-analytical models were applied to the resulting models' estimates to summarise the abundance of results across countries and scales and further relate the scale-specific incremental gender differences in item nonresponse to two scale characteristics, scale response format and scale content.

Our preliminary descriptive analyses found item nonresponse rates to range from 0.3% in B-S-J-Z (China) to 22% in Baku (Azerbaijan), with a relatively low average across countries of 5%. Average item nonresponse rates on specific scales ranged from 0.4% on scale ST022 (i.e., the language most often spoken at home) to 16.5% on scale ST008 (i.e., father's education), with an average across scales of 5%. Regarding our *RQ1* (i.e., to what extent do female and male students differ in their average item nonresponse propensity on the PISA 2018 student background questionnaire?), our findings further suggest that, on average across countries, the odds of male students not responding to an item on an average PISA 2018 student questionnaire scale were double the odds of female students. The higher propensity of item nonresponse for male students is consistent with the expected trend in the low-stakes setting where gender differences are viewed in connection with test-taking motivation. Test-taking motivation has been commonly related to non-cognitive traits, such as conscientiousness and agreeableness, where the examinees with higher levels of both are expected to put greater effort into responding in low-stakes settings (see, e.g., DeMars et al., 2013). Given that male students are generally believed to be less conscientious and agreeable, as well as more work-avoidant (Costa et al., 2001; DeMars et al., 2013), they might be less motivated to respond and consequently produce higher item nonresponse rates.

Regarding our *RQ2* (i.e., do certain scales on the PISA 2018 student questionnaire elicit larger or smaller gender differences than the average gender gap in item nonresponse propensity?), our findings suggest that, when comparing female and male students with the same item nonresponse propensity, male students had greater odds of omitting open-response scales, whereas female students omitted more dichotomous and multiple-choice scales. We had no explicit hypothesis as to which scale format would elicit more item nonresponse for which gender. However, we did not anticipate our findings in the low-stakes questionnaire to mimic those of the high-stakes achievement tests. In high- and low-stakes achievement settings, female examinees have been shown to omit more multiple-choice items while male examinees omit more open-response items (Gafni & Melamed, 1994; Grandy, 1987; Matters & Burnett, 1999; Okumura, 2014). In high-stakes literature, these trends have been widely linked to differential risk-taking tendencies in competitive testing situations, with males believed to be more likely to take the risk and guess on a multiple-choice item and females being more likely to avoid that risk and omit (provided they do not have a definitively correct answer; Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994). Implicitly, we assumed that with guessing no longer being beneficial to the score received by a questionnaire participant, the trends in omitting scales of these formats might shift between genders (here, we treat dichotomous scales as a special case of multiple-choice, where only two options are given and one has to be selected). Instead, our findings suggested that the propensities of item nonresponse on the multiple-choice and the open-response scales may be similar between high- and low-stakes as well as between achievement tests and questionnaires.

Similar trends observed in future research could suggest that the directionality of various format effects on item nonresponse is a multifaceted feature not unique to competitive settings and not fully accounted for by students' propensity to guess, warranting a further investigation into its driving factors. Most multiple-choice and dichotomous scales on the PISA 2018 student questionnaire, for instance, gather *factual information* about students' backgrounds instead of their *attitudes* (e.g., scales on parents' education, original places of birth, and availability of home resources). Hence, we cautiously speculate that such items could evoke test-taking strategies similar to those in the high-stakes assessments. For example, when female students have no definitive (factually "correct") answer to whether their father holds a postgraduate degree, they might omit an item, whereas male students might likely guess. Arguably, the same logic might apply to the Likert frequency scales (e.g., primarily *factual* scales such as the number of classes per week and frequency of teacher feedback). In most countries, when comparing female and male students with the same item nonresponse propensity, we found female students to have greater odds of item nonresponse on Likert frequency scales, whereas male students omitted more Likert agreement scales (e.g., generally *attitudinal* scales such as perceived emotional support from parents and attitudes towards bullying).

Our results further show that the item nonresponse gender gaps also fluctuate (i.e., narrow or widen) depending on the scale content. When comparing female and male students with the same item nonresponse propensity, female students had greater odds of not responding to scales asking them to provide their background information (e.g., students' parents' educational and professional history, socioeconomic status, and ethnic background). The incremental differences in item nonresponse propensity that we observed on these background scales effectively closed the overall gender gap in item nonresponse in many countries such that female and male students were equally likely not to respond.

We illustrate this finding on two scales, ST006 and ST008, inquiring about the education of students' mothers and fathers, respectively. When comparing students of the same gender and with the same item nonresponse propensity, the odds of item nonresponse on the items of these scales were estimated to be 26 times greater than the odds of item nonresponse on the other items in a similar position within the questionnaire. Our findings suggest that the average gender gap on these scales nearly closed such that the odds of item nonresponse for female and male students became nearly even. We see several issues with these particular scales that could lead to higher than desired nonresponse rates. First, one can assert that the response format is sub-optimal. Per instructions, students had to respond to each item asking whether their parent had a certain level of education (e.g., scale ST006 — Does your mother have any of the following qualifications?; ST006Q02TA — "post-graduate degree"; response options — yes/no). With less attention to spare after two hours of cognitive tests, some students may overlook this instruction and proceed with a more intuitive strategy of choosing "all that apply" and leaving the rest blank. Second and perhaps, a more marring issue arises for students without a present mother or father but in the permanent care of their other immediate, extended, or non-biological family. Such students are left face to face with scale(s) to which they can neither relate nor provide a valid response. A routing option (e.g., a filter item) to identify students' primary caregivers could help combat the high nonresponse rates on these scales. Finally, we believe that it could sometimes be unrealistic to expect a 15-year-old to know their parents' education. Hopfenbeck & Kjærnsli (2016) report an isolated incident showcasing this point in the example of Norway. When asked about the PISA test experience, one female student reported that "she did not quite know what to answer about her parents' education, since she did not know about it" (Hopfenbeck & Kjærnsli, 2016, p. 417). As Hopfenbeck & Kjærnsli (2016) conclude, the common thread in many of the conducted interviews was students' perception of the background questionnaire as problematic, be it due to them not knowing the answer to the items or feeling apprehensive about uncovering their private lives.

In most countries, when comparing female and male students with the same item nonresponse propensity, female students have also been found to have greater odds of item nonresponse on scales covering various schooling constructs such as, for example, students' sense of belonging, experiences with bullying, perceived teacher feedback, and adaptability of instruction. On the other hand, male students were more likely to omit dispositional scales than female students with the same item nonresponse propensity. Dispositional scales gathered information on various non-cognitive and meta-cognitive constructs, such as students' attitudes towards learning, competitiveness, resilience, fear of failure, and dispositions for global competence. In-depth exploration of gender differences in item nonresponse on the schooling and non-/meta-cognitive scales, but at a finer-grained level of topic differentiation, could present a promising area for inquiry and nuance our understanding of the students' test-taking behaviours. For instance, future research could explore gender differences in item nonresponse against the backdrop of intercultural sensitive topics (e.g., attitudes towards immigrants, respect for other cultures, awareness of global issues).

We make three additional remarks regarding the effects of the scale response format and content on gender differences in item nonresponse. The first one relates to the substantial differences in effect size across countries. Although some gender differences might be relatively small or non-existent, when averaged across countries, they can be more pronounced in certain countries or regions with similar cultural or linguistic backgrounds. For example, compared to male students with the same item nonresponse propensity, the odds of item nonresponse of female students (when averaged across countries) were just slightly over one on the multiple-choice, dichotomous, and student background scales. However, high differential format and content bias (for the same sequence of the multiple-choice, dichotomous, and student background scales) was found in Qatar and the United Arab Emirates (where the odds of item nonresponse for female students were double or triple the odds of male students with the same item nonresponse propensity).

The second note relates to some content and format overlap across scales. For example, most non-/meta-cognitive scales followed the Likert agreement or frequency response format. Given our quasi-experimental study design, we could not experimentally manipulate the questionnaire, and consequently, we could not fully disentangle the scale format or content. The third remark concerns the opportunities for further advancement and cross-validation of our findings regarding gender differences in item nonresponse on the student questionnaire scales that were additionally administered as parts of the PISA 2018 parent or teacher questionnaires (OECD, 2019). The response data on the scales identical to the student questionnaire but gathered from parents and teachers and linked back to the respective students could help identify complementary trends in student nonresponse behaviour. For example, one might be interested in determining if students who omit the items on their parents' education predominantly have parents without higher or completed secondary education.

Lastly, this study aimed to identify the presence of gender gaps in item nonresponse on the PISA 2018 student questionnaire and we linked those gaps to potential covariates such as scale format and scale content. However, there are a number of potential topics for future research on item nonresponse that could build on the present study to ultimately enhance our knowledge of the nature, the extent, and the consequences of the item nonresponse phenomenon. For instance, other factors that may contribute to the width of gender gaps in item nonresponse may warrant further investigation. Future research could explore, among many other factors, the association between students' reading abilities and item nonresponse at the individual response level (e.g., at the country level, there was a strong positive association between the gender gaps in reading achievement and item nonresponse; $r = 0.77$). Furthermore, the impact of item nonresponse on scale construction and its potential to lead to biased parameter estimates was not addressed in this study and could present promising opportunities for future inquiry.

5 Conclusion

The current study identified systematic gender differences in item nonresponse propensities on the PISA 2018 student background questionnaire. Therefore, we argue that in order to arrive at valid inferences based on the PISA 2018 questionnaire data, the presence of such differences should be considered in future studies when modelling the missing data mechanism (Rubin, 1976). On average across countries, male students had greater odds of item nonresponse than female students, consistent with the expected trend in the low-stakes setting, where male students are believed to be less motivated and exert less test-taking effort than female students (DeMars et al., 2013). However, we show that gender differences in item nonresponse are not merely a function of the stakes involved for individual students. Our results suggest that gender differences in item nonresponse may be a more complex phenomenon that is context-dependent and not necessarily stable across countries and scales' formats and contents. To that end, we argue that differences in item nonresponse patterns are a source of additional information about the test-taking behaviours of students as well as the quality of the items and the questionnaire as a whole.

Appendix 1. Country- and scale-wise item nonresponse rates

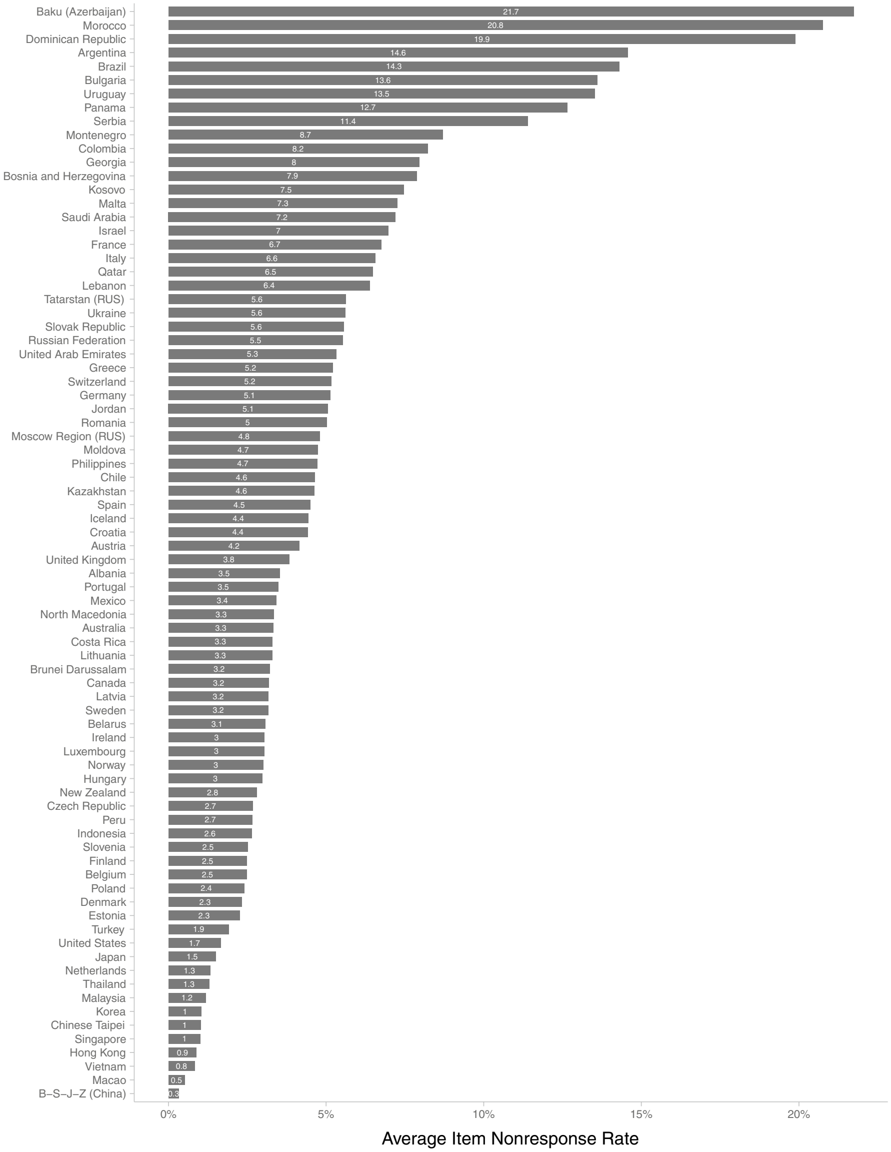


Fig. 3 Country-wise average item nonresponse rates. *Note*. Each country's average item nonresponse rate is calculated by averaging individual students' item nonresponse rates (i.e., the ratio of the number of items on the questionnaire to which the student did not provide a response to the number of theoretically valid responses possible on the questionnaire administered within a country) to the country level.

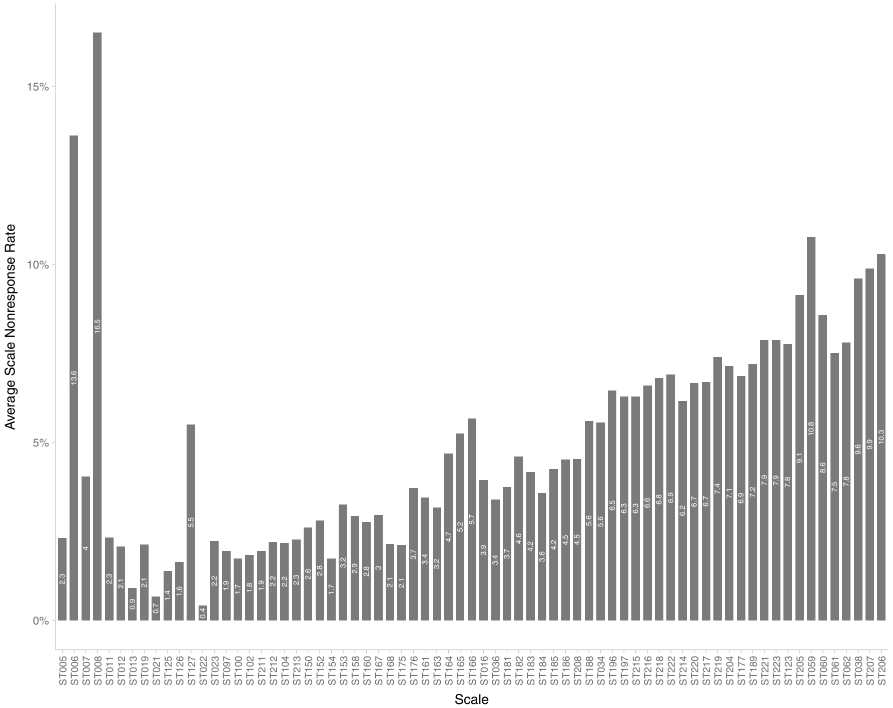


Fig. 4 Scale-wise average nonresponse rates. *Note*. Each scale's nonresponse rate is calculated by first averaging individual items' nonresponse rates by scale in each country and then averaging across countries for each scale. The items are arranged in the sequence as they appear on the questionnaire.

Appendix 2. Item nonresponse propensity as a function of item position

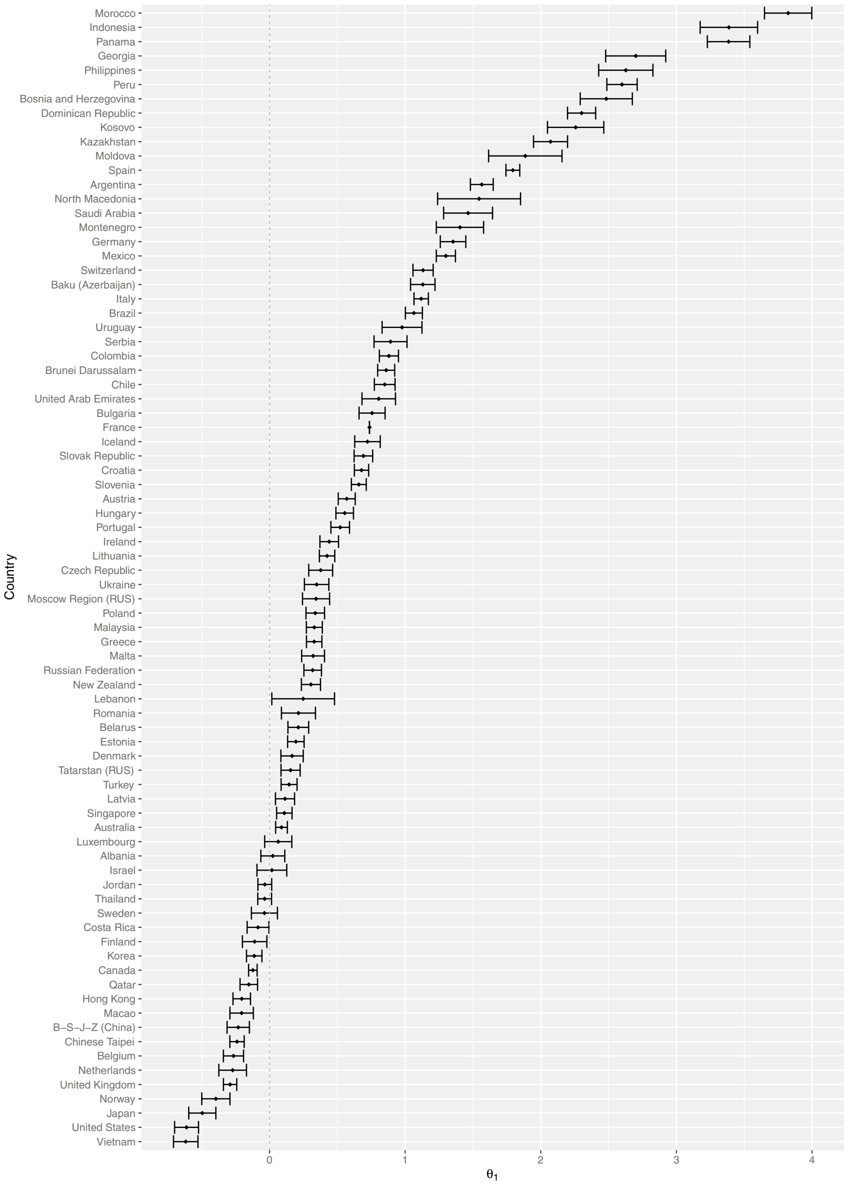


Fig. 5 Odds of item nonresponse propensity as a function of item position. *Note.* The parameter estimates are presented on the logit scale and correspond to the θ_1 parameter estimates of the cross-classified mixed effects models (Eq. 1). One unit on the item position predictor corresponds to 50 items. For students of the same gender and with the same item nonresponse propensity, positive and negative parameter values correspond to higher and lower odds of item nonresponse as they progress through the questionnaire, respectively. At the grey dashed line, neither outcome (nonresponse or response) is more likely than the other ($\theta_1 = 0$).

Appendix 3. Item nonresponse propensity on scales ST006 and ST008

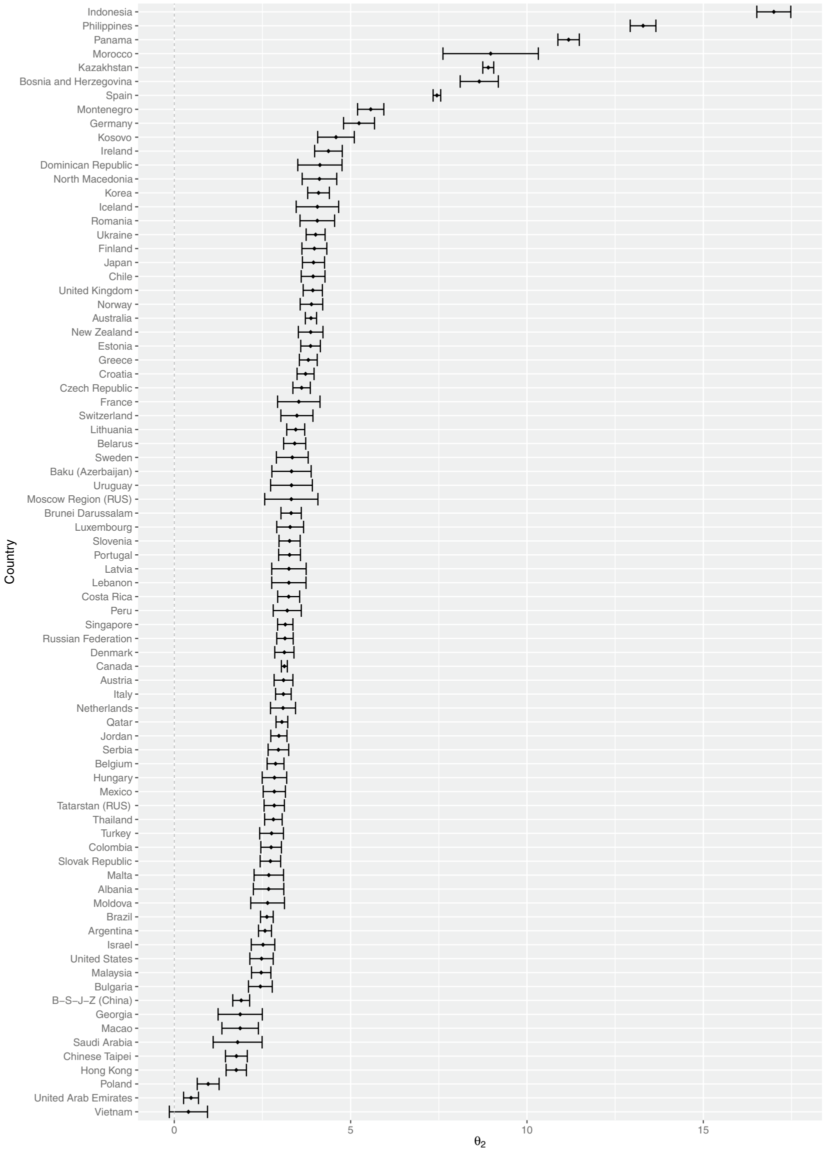


Fig. 6 Odds of item nonresponse propensity on scales ST006 and ST008. *Note.* The parameter estimates are presented on the logit scale and correspond to the θ_2 parameter estimates of the cross-classified mixed effects models (Eq. 1). The scales ST006 and ST008 consisted of four items each and inquired about students' parents' qualifications (i.e., scale ST006: Does your mother have any of the following qualifications? ST006Q01TA < ISCED level 6 > (doctorate), ST006Q02TA < ISCED level 5A > (post-graduate), ST006Q03TA < ISCED level 5B > (vocational tertiary), ST006Q04TA < ISCED level 4 > (non-tertiary post-secondary)). A yes/no response had to be selected for each item. For students of the same gender and with the same item nonresponse propensity, positive and negative parameter estimates correspond to higher and lower odds of item nonresponse on an average item of the ST006 and ST008 scales, respectively, as compared to the odds of item nonresponse on the other items with a similar position on the questionnaire. At the grey dashed line, neither outcome (nonresponse or response) is more likely than the other ($\theta_2 = 0$).

Appendix 4. Scale-specific incremental gender differences in item nonresponse propensity

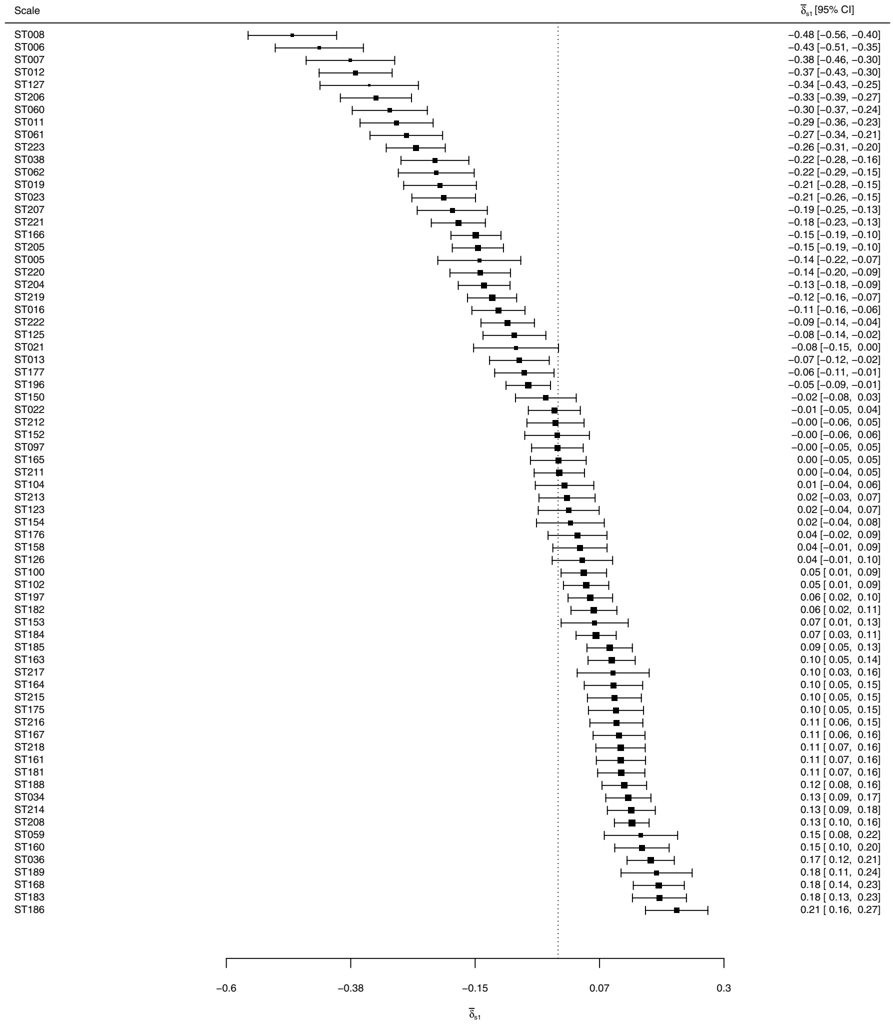


Fig. 7 Scale-specific incremental gender differences in item nonresponse propensity (female students as the reference group). *Note.* The estimates $\overline{\delta}_{s1}$ are presented on the logit scale and correspond to the results of the meta-analyses that summarised the parameter estimates for the scale-specific incremental gender differences in item nonresponse propensity (δ_{s1}, Eq. 1) across countries for each individual scale, resulting in 71 pooled estimates $\overline{\delta}_{s1}$ (one for each scale s). Positive and negative estimates correspond to greater odds of item nonresponse on the corresponding PISA 2018 student background questionnaire scale for male and female students, respectively. At the grey dashed line at 0, neither gender has greater odds of item nonresponse.

Appendix 5. PISA 2018 student questionnaire scales characteristics

Table 1 PISA 2018 student questionnaire scales characteristic (number of items per scale, scale response format, target construct, category of scale, topic)

Scale	Items	Format	Construct	Category	Topic
ST005	1	Multiple-choice	Student background	General	Mother’s education
ST006	4	Dichotomous	Student background	General	Mother’s qualifications
ST007	1	Multiple-choice	Student background	General	Father’s education
ST008	4	Dichotomous	Student background	General	Father’s qualifications
ST011	13	Dichotomous	Student background	General	Availability of household items
ST012	8	Likert frequency	Student background	General	Amount of home possessions
ST013	1	Likert frequency	Student background	General	Amount of books at home
ST019	3	Multiple-choice	Student background	General	Parents’ and student’s place of birth
ST021	1	Multiple-choice	Student background	General	Age of arrival to the country of test
ST125	1	Multiple-choice	Student background	General	Early educational pathways < ISCED 0 >
ST126	1	Multiple-choice	Student background	General	Early educational pathways < ISCED 1 >
ST127	3	Likert frequency	Student background	General	Grade repetition
ST022	1	Multiple-choice	Student background	General	Language spoken at home
ST023	5	Likert frequency	Student background	General	Language spoken with family
ST097	5	Likert frequency	Schooling	Reading	Disciplinary climate
ST100	4	Likert frequency	Schooling	Reading	Teacher support in test language
ST102	4	Likert frequency	Schooling	Reading	Teacher instruction
ST211*	3	Likert agreement	Schooling	Reading	Teacher responsiveness
ST212	3	Likert frequency	Schooling	General	Adaptivity of instruction
ST104	3	Likert frequency	Schooling	General	Perceived teacher feedback
ST213	4	Likert agreement	Schooling	Reading	Teacher interest in teaching
ST150*	4	Likert frequency	Schooling	Reading	Reading practices at school

Table 1 (continued)

Scale	Items	Format	Construct	Category	Topic
ST152	4	Likert frequency	Schooling	Reading	Teacher stimulation of reading
ST154*	1	Likert frequency	Schooling	Reading	Reading practices at school
ST153*	9	Dichotomous	Schooling	Reading	Teacher-directed reading tasks
ST158*	7	Dichotomous	Schooling	Reading	Use of internet at school
ST160	5	Likert agreement	Non-/meta-cognitive	Reading	Enjoyment of reading
ST167*	5	Likert frequency	Student background	Reading	Out-of-school reading experiences
ST168*	1	Likert frequency	Student background	Reading	Frequencies and practices for out-of-school reading
ST175*	1	Likert frequency	Student background	Reading	Time spent on reading for enjoyment
ST176*	6	Likert frequency	Student background	Reading	Time spent on online reading activities
ST161	6	Likert agreement	Non-/meta-cognitive	Reading	Perception of competence and difficulty when performing reading tasks
ST163	3	Likert agreement	Non-/meta-cognitive	Reading	Perception of difficulty of the PISA test
ST164	6	Likert agreement	Non-/meta-cognitive	Reading	Understanding and remembering
ST165	5	Likert agreement	Non-/meta-cognitive	Reading	Summarising
ST166	5	Likert agreement	Non-/meta-cognitive	Reading	Assessing credibility
ST016*	1	Likert agreement	Non-/meta-cognitive	General	Satisfaction with life
ST036	3	Likert agreement	Non-/meta-cognitive	General	Attitudes toward learning activities
ST181	3	Likert agreement	Non-/meta-cognitive	General	Competitiveness
ST182	4	Likert agreement	Non-/meta-cognitive	General	Working motive and mastery
ST183	3	Likert agreement	Non-/meta-cognitive	General	General fear of failure
ST184*	1	Likert agreement	Non-/meta-cognitive	General	Intelligence view
ST185	3	Likert agreement	Non-/meta-cognitive	General	Sense of meaning and purpose in life
ST186	9	Likert frequency	Non-/meta-cognitive	General	Subjective well-being

Table 1 (continued)

Scale	Items	Format	Construct	Category	Topic
ST208	3	Likert agreement	Non-/meta-cognitive	General	Mastery-approach orientation of achievement goals
ST188	5	Likert agreement	Non-/meta-cognitive	General	Resilience
ST034	6	Likert agreement	Schooling	General	Sense of belonging to school
ST196	6	Likert agreement	Non-/meta-cognitive	General	Self-efficacy explaining complex global issues
ST197	7	Likert agreement	Non-/meta-cognitive	General	Awareness of global issues
ST215	5	Likert agreement	Non-/meta-cognitive	General	Perspective taking
ST216	6	Likert agreement	Non-/meta-cognitive	General	Adaptability in dealing with challenging situations
ST218	7	Likert agreement	Non-/meta-cognitive	General	Intercultural communicative awareness
ST222*	8	Dichotomous	Non-/meta-cognitive	General	Activities dealing with complex global issues
ST214	4	Likert agreement	Non-/meta-cognitive	General	Learning about other cultures
ST220*	4	Dichotomous	Non-/meta-cognitive	General	Contact with people from other countries
ST217	5	Likert agreement	Non-/meta-cognitive	General	Respect for people from other cultures
ST219	6	Likert agreement	Non-/meta-cognitive	General	Sense of global-mindedness
ST204	4	Likert agreement	Non-/meta-cognitive	General	Attitudes toward equal rights for immigrants
ST177*	3	Likert frequency	Student background	General	Amount of shared languages with parents
ST189*	1	Open-response	Schooling	Reading	Number of foreign languages at school
ST221*	10	Dichotomous	Non-/meta-cognitive	General	Learning about complex global issues at school
ST223	4	Likert frequency	Non-/meta-cognitive	General	Multicultural school climate
ST123	3	Likert agreement	Schooling	General	Perceived emotional support from parents
ST205	4	Likert agreement	Schooling	General	Climate of competition in school
ST059	4	Open-response	Schooling	Reading	Learning time in test language
ST060	1	Likert frequency	Schooling	Reading	Number of class periods per week

Table 1 (continued)

Scale	Items	Format	Construct	Category	Topic
ST061	1	Likert frequency	Schooling	Reading	Average minutes in a class period
ST062*	3	Likert frequency	Schooling	Reading	Skipping classes
ST038	6	Likert frequency	Schooling	General	Experience of being bullied
ST207*	5	Likert agreement	Schooling	General	Attitudes toward bullying
ST206	4	Likert agreement	Schooling	General	Climate of cooperation in school

The *dichotomous* format included scales that required students to choose between two given response options. The *multiple-choice* format was represented by scales permitting the choice of one option out of more than two unordered response options. Also coded as multiple-choice were the items with multiple response options, which despite following a natural order, could not be meaningfully interpreted as agreement or frequency (e.g., ST021Q01TA). The remaining scales with ordered response categories were coded as *Likert agreement* and *Likert frequency*. Two scales that did not provide response options were coded as *open-response*. The scales were coded into six content areas corresponding to the intersections of the construct and topic columns of this table (e.g., student background: general, student background: reading). The PISA 2018 documentation did not connect 18 scales to a specific construct, and these scales were assigned one of the six content areas, corresponding the closest to the topics they covered (marked by an asterisk)

Appendix 6. Scale-specific incremental gender differences in item nonresponse propensity as a function of the scale response format and the scale content by country

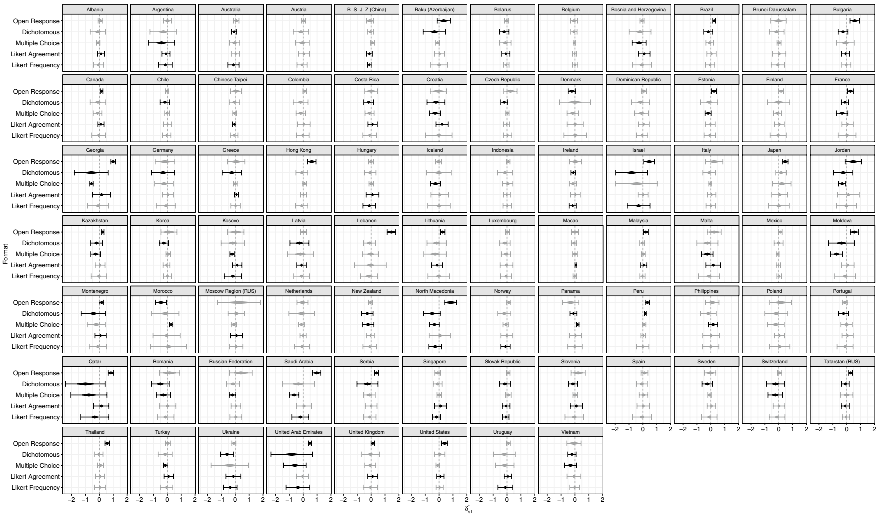


Fig. 8 Scale-specific incremental gender differences in item nonresponse propensity as a function of the scale response format (female students as the reference group). *Note.* The results are reported for five response formats. The estimates δ^*_{s1} are presented on the logit scale and correspond to the results of the subgroup meta-analyses that summarised the parameter estimates for the scale-specific incremental gender differences in item nonresponse propensity (δ_{s1} in Eq. 1) for each country as a function of scale format. The diamond shapes represent the confidence intervals around the estimate for each format within each country, and the bars around the diamond define the corresponding prediction interval for a randomly sampled scale from all scales of the same format. When the estimate is significantly different from zero, the diamond shape and bars are black; otherwise, grey. When comparing female and male students with the same item nonresponse propensity, positive and negative estimates correspond to greater odds of item nonresponse for male and female students, respectively. At the grey dashed line at 0, neither male nor female students have greater odds of item nonresponse.

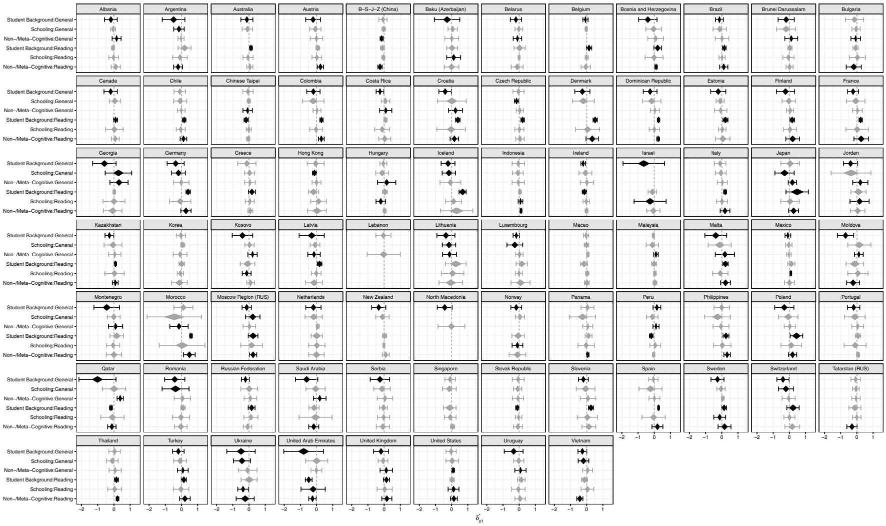


Fig. 9 Scale-specific incremental gender differences as a function of the scale content (female students as the reference group). *Note.* The results are reported for six content areas. The estimates δ^*_{s1} are presented on the logit scale and correspond to the results of the subgroup meta-analyses that summarised the parameter estimates for the scale-specific incremental gender differences in item nonresponse propensity (δ_{s1} in Eq. 1) for each country as a function of scale content. The diamond shapes represent the confidence intervals around the estimate for each format within each country, and the bars around the diamond define the corresponding prediction interval for a randomly sampled scale from all scales covering the same content. When the estimate is significantly different from zero, the diamond shape and bars are black; otherwise, grey. When comparing female and male students with the same item nonresponse propensity, positive and negative estimates correspond to greater odds of item nonresponse for male and female students, respectively. At the grey dashed line at 0, neither male nor female students have greater odds of item nonresponse.

Appendix 7. Procedural steps of the analysis

The procedure for conducting the present study is as follows.

Step 1. Data management. The data for the student questionnaire is available at https://www.oecd.org/pisa/data/2018database/. The steps undertaken in this study are:

1. Re-code “No Response” PISA missing values into 1, “Valid Response” into 0, the rest into NA;
2. Split the main dataset into 80 separate datasets (by country);
3. For each dataset, perform the following operations:

   (a) Subset the dataset to the items that the country administered;
   (b) Transform data into long format;
   (c) Create a continuous variable POSITION counting from 1 to n-th item the country administered; the variable is then re-scaled to count 50 items per unit and start at 0 as $\frac{\text{POSITION}-1}{50}$ and centred;

(d) Remove items that were cross-referenced by the school officials with exception of the GENDER item which is retained to be used as a predictor in the cross-classified mixed effects model;

(e) Create SCALE, FORMAT, and CONTENT variables to denote to which scale each item belongs, which format the scale follows, and which content the scale covers, respectively (Table 1, Appendix 5);

(f) Create a binary X variable such that $X = 1$ if SCALE = = “ST006” or SCALE = = “ST008”, 0 otherwise.

Step 2. Cross-classified mixed effects model. The cross-classified mixed effects model is fitted using *lme4* package to each country’s dataset separately (a total of 80 models; Bates et al., 2015):

```
library(lme4)
m res < −glmer(NONRESPONSE ∼ 1 + (1 + POSITION|STUDENTID))+
                    (1 + GENDER|SCALE) + POSITION + X + GENDER,
                              family = binomial, data = dat,
  control = glmerControl(optimizer = ″bobyqa″, optCtrl = list(maxfun = 2e5)))
```

The following parameter estimates are extracted from the fitted models for each country:

1. Fixed effect parameter estimates with their respective standard errors: general intercept (β_0), POSITION slope (θ_1), GENDER slope (β_1), X slope (θ_2; Eq. 1);
2. Random effect parameter estimates: random slopes of GENDER by SCALE (δ_{s1}; Eq. 1).

Step 3. Representation of results. The effect parameters of interest are combined across countries using random effects meta-analytical models in the metafor package (Viechtbauer, 2010):

```
library(metafor)
meta <− rma(yi, vi, data = dat)
#′yi′ − the parameter estimates
#′vi′ − their respective standard error squared
```

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital)

Data Availability The data presented in this study is public and is available on OECD official page: https://www.oecd.org/pisa/data/2018database/.

Declarations

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere. Appropriate ethical guidelines were followed in the conduct of the research and we have no conflicts of interest to declare. No funding was received to assist with the preparation of this manuscript. All data analysed in this study are readily available at https://www.oecd.org/pisa/data/2018database/.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23–35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. (2009). *Introduction to meta-analysis*. John Wiley and Sons, Ltd.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

De Boeck, P., & Cho, S.-J. (2021). Not all DIF is shaped similarly. *Psychometrika, 86*(3), 712–716. https://doi.org/10.1007/s11336-021-09772-3

De Boeck, P., & Wilson, M., (Eds.). (2004). *Explanatory item response models*. Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

DeMars, C., Bashkov, B., & Socha, A. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment, 8*, 69–82.

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation, 20*(3), 309–319. https://doi.org/10.1016/0191-491X(94)90018-3

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring* (Tech. Rep. No. 87-38). Educational Testing Service.

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal, 27*(3), 406–422. https://doi.org/10.1080/09585176.2016.1156004

Jakwerth, P. R., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions* (Tech. Rep.). American Institutes for Research.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement, 54*(4), 397–419. https://doi.org/10.1111/jedm.12154

Matters, G., & Burnett, P. C. (1999). Multiple-choice versus short-response items: Differences in omit behaviour. *Australian Journal of Education, 43*(2), 117–128. https://doi.org/10.1177/000494419904300202

Meinck, S., Cortes, D., & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: A theoretical study. *Large-scale Assessments in Education, 5*(3), 1–21. https://doi.org/10.1186/s40536-017-0038-6

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. TIMSS and PIRLS International Study Center.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives, 24*(2), 129–144. https://doi.org/10.1257/jep.24.2.129

OECD. (2019). PISA 2018 background questionnaires. In *PISA 2018. Assessment And Analytical Framework*. OECD Publishing. https://doi.org/10.1787/67e1518f-en

OECD. (2020). *PISA 2018 technical report, PISA*. OECD Publishing.

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educational and Psychological Measurement, 74*(4), 611–626. https://doi.org/10.1177/0013164413516976

Papanastasiou, E. C. (2020). Can non-responses speak louder than words? Examining patterns of item non-response in TIMSS 2015. *International Journal of Quantitative Research in Education, 5*(2), 157–172. https://doi.org/10.1504/IJQRE.2020.10033505

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386. https://doi.org/10.3102/10769986028004369

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3). https://doi.org/10.18637/jss.v036.i03

Von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*(1), 41–65. https://doi.org/10.1207/s15324818ame1901_3

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1