 doi:10.20944/preprints202110.0107.v1

Article

On the Treatment of Missing Item Responses in Educational Large-scale Assessment Data: The Case of PISA 2018 Mathematics

Alexander Robitzsch [1,2]*

[1] IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany
[2] Centre for International Student Assessment (ZIB), Olshausenstraße, 24118 Kiel, Germany
* Correspondence: robitzsch@leibniz-ipn.de

Abstract: Missing item responses are prevalent in educational large-scale assessment studies like the programme for international student assessment (PISA). The current operational practice scores missing item responses as wrong, but several psychometricians advocated a model-based treatment based on latent ignorability assumption. In this approach, item responses and response indicators are jointly modeled conditional on a latent ability and a latent response propensity variable. Alternatively, imputation-based approaches can be used. The latent ignorability assumption is weakened in the Mislevy-Wu model that characterizes a nonignorable missingness mechanism and allows the missingness of an item to depend on the item itself. The scoring of missing item responses as wrong and the latent ignorable model are submodels of the Mislevy-Wu model. This article uses the PISA 2018 mathematics dataset to investigate the consequences of different missing data treatments on country means. Obtained country means can substantially differ for the different scaling models. In contrast to previous statements in the literature, the scoring of missing item responses as incorrect provided a better model fit than a latent ignorable model for most countries. Furthermore, the dependence of the missingness of an item from the item itself after conditioning on the latent response propensity was much more pronounced for constructed-response items than for multiple-choice items. As a consequence, scaling models that presuppose latent ignorability should be refused from two perspectives. First, the Mislevy-Wu model is preferred over the latent ignorable model for reasons of model fit. Second, we argue that model fit should only play a minor role in choosing psychometric models in large-scale assessment studies because validity aspects are most relevant. Missing data treatments that countries can simply manipulate (and, hence, their students) result in unfair country comparisons.

Keywords: missing item responses, multiple imputation, item response model, PISA, country comparisons, Mislevy-Wu model, latent ignorability, nonignorable item responses

1. Introduction

It has frequently been argued that measured student performance in educational large-scale assessment (LSA; [1,2]) studies is affected by test-taking strategies. In a recent paper that was published in the highly-ranked *Science* journal, Pohl et al. [3] argued that "current reporting practices, however, confound differences in test-taking behavior (such as working speed and item nonresponse) with differences in competencies (ability). Furthermore, they do so in a different way for different examinees, threatening the fairness of country comparisons." [3]. Hence, the reported student performance (or, equivalently, student ability) would be confounded by a "true" ability and test-taking strategies. Importantly, the authors question the validity of country comparisons that are currently reported in LSA studies and argue for an approach that separates test-taking behavior (i.e., item response propensity and working speed) from a purified ability measure. The core of the Pohl et al. [3] approach is on how to model missing item responses. In this article, we systematically investigate the consequences of different treatments of missing item responses in the programme for international student assessment (PISA) study conducted in 2018.

 doi:10.20944/preprints202110.0107.v1

While the treatment of missing data in statistical analyses in social sciences is now widely used [4,5], in recent literature, there are recommendations for treating missing item responses in item response theory (IRT; [6]) models in LSA studies [7,8]. Typically, the treatment of item responses can be distinguished between calibration (computation of item parameters) and scaling (computation of ability distributions).

It is essential to distinguish the type of missing item responses. Missing item responses at the end of the test are referred to as not reached items while missing items within the test are denoted as omitted items [9]. Since the PISA 2015 study, not reached items are no longer scored as incorrect, and the proportion of not reached items is used as a predictor in the latent background model [10].

Some psychometricians repeatedly argued that missing item responses should never be scored as wrong because such a treatment would produce biased item parameter estimates and unfair country rankings [3,7,8,11,12]. In contrast, model-based treatments of missing item responses that rely on latent ignorability [3,7,8,13] are advocated. Missing item responses can be ignored in this approach when including response indicators and a latent response propensity [14,15]. Importantly, the missingness process is summarized by the latent response variable. As an alternative, multiple imputation at the level of items can be employed to handle missing item responses properly [16,17]. However, the scoring of missing item responses as wrong has been defended for validity reasons [18–20]. Moreover, simulation studies cannot inform about the proper treatment of missing item responses [19,21].

Although the proposals of using alternative scaling models for abilities in LSA studies like PISA have been made, previous work either did not report country means in the metric of interest [7] such that consequences cannot be interpreted, or constituted only a toy analysis consisting only a few countries [3] that did enable a generalization to operational practice. Therefore, this article tries to compare different scaling models that rely on different treatments of missing item responses. We use the PISA 2018 mathematics dataset as a showcase. We particularly contrast the scoring of missing item responses as wrong with model-based approaches that rely on latent ignorability [3,7,8] and a more flexible Mislevy-Wu model [22,23] containing the former two models as submodels. In the framework of the Mislevy-Wu model, it is tested whether the scoring of missing item responses as incorrect or treating them as latent ignorable are preferred in terms of model fit. Moreover, it is studied whether the probability of responding to an item depends on the item response itself (i.e., nonignorable missingness, [5]). In the most general model, the missingness process is assumed to be item format specific. Finally, we investigate the variability across means from different models for a country.

The rest of the article is structured as follows. In Section 2, the sample of persons and items in PISA 2018 is characterized. In Section 3, the different scaling models and the linking procedure are described. In Section 4, the results are presented. Finally, the paper closes with a discussion in Section 5.

2. Sample

The mathematics test in PISA 2018 [24] was used to investigate different treatments of missing item responses. We included 45 countries that did receive the main test. These country did not receive test booklets with items of lower difficulty that were included for low-performing countries. In total, 70 items were included in our analysis. Seven out of 70 items were partial credit items with a maximum score of two. For simplicity, these partial credit items were dichotomized (i.e., dichotomously scored as correct with a score of two for the partial credit item). In total, 27 items had the multiple-choice (MC) format, and 43 items had constructed-response (CR) format. For 18 MC items, the guessing probability was 1/4, for 4 MC items, it was 1/8, and for 5 MC items, it was 1/16.

Students from booklets 1 to 12 were selected. Those booklets had mathematics items included in two out of four item clusters. Mathematics items appeared either at the first and second or the third and fourth positions in the test.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

In Table 1, descriptive statistics for the sample used in our analysis are presented. In total, 167,092 students from these 45 countries were included in the analysis. On average, M = 3713.2 students were available in each country. The average number of students per item within each country ranged between 415.8 (MLT, Malta) and 4408.3 (ESP, Spain). On average, M = 1120.3 students per item were available at the country level.

The average proportion of missing item responses in the dataset was 8.4% (SD = 3.3%) and ranged between 1.2% (MYS, Malaysia) and 18.8% (BIH; Bosnia and Herzegovina). The proportion of not reached item responses was on average 2.4% (SD = 1.0%) with the maximum of 5.9% (SWE, Sweden). Interestingly, the missing data proportions and the country means were only moderately correlated (Cor = −.48). Missing proportions for CR items were substantially larger (M = 12.3%, SD = 4.8%, Min = 1.5%, Max = 27.9%) than for MC items (M = 2.3%, SD = 1.0%, Min = 0.7%, Max = 5.4%).

3. Analysis

As stated above, all polytomous items were dichotomously scored for simplicity. Let X_{pi} denote the dichotomous item responses and the R_{pi} response indicators for person p and item i. The response indicator takes the value one if X_{pi} is observed. Consistent with the operational practice since PISA 2015, the two-parameter logistic (2PL) model [25] is used for scaling item responses [10,24]. The item response function is given as

$$\mathrm{P}(X_{pi}=1|\theta_p)=\Psi(a_i(\theta_p-b_i)) \quad , \tag{1}$$

where Ψ denotes the logistic distribution function. The item parameters a_i and b_i are item discriminations and difficulties, respectively. It holds that $1-\Psi(x)=\Psi(-x)$. The latent ability θ_p follows a normal distribution. If all item parameters are estimated, the mean of the ability distribution is fixed to zero and the standard deviation is fixed to one. The one-parameter logistic (1PL, [26]) model is obtained if all item discriminations are set equal to each other.

In our analysis, the scalings are carried out separately for each country c. That is, one obtains country-specific item parameters a_{ic} and b_{ic}:

$$\mathrm{P}(X_{pci}=1|\theta_{pc})=\Psi(a_{ic}(\theta_{pc}-b_{ic})) \quad , \quad \theta_{pc}\sim \mathrm{N}(0,1) \quad . \tag{2}$$

Sampling weights were always when applying the scaling model (2) to the PISA 2018 dataset. To enable the comparability of the ability distribution across countries, the obtained item discriminations a_{ic} and item difficulties b_{ic} are transformed on a common in a subsequent linking step (see Section 3.2) for details.

3.1. Different Scaling Models for Handling Missing Item Responses

In this subsection, we describe the different scaling models used for determining country means. These models differ concerning the missingness mechanism assumptions of missing item responses.

3.1.1. Scoring Missing Item Responses as Wrong

In a reference model, we scored all missing item responses (omitted and not reached items) as wrong (model UW). In the literature, it is frequently argued that missing item responses should never be scored as incorrect [3,7,11,27]. However, we think that the arguments against the incorrect scoring are flawed, and simulation studies cannot show the inadequacy of the UW model (see [19–21]).

3.1.2. Scoring Missing Item Responses as Partially Correct

Missing responses for MC items can be scored as partially correct. The main idea is that a student could guess the MC item if s/he does not know the answer. If an item i has K_i alternatives, a random guess of an item option would provide a correct response with probability $1/K_i$. In IRT estimation, one can weigh probabilities $\mathrm{P}(X_{pi}=1)$ with $1/K_i$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

Table 1. Descriptive Statistics of the PISA 2018 Mathematics Sample

Country	N	I	N_{item}	M_{OECD}	SD_{OECD}	M_{stand}	%NA	%NR	$\%NA_{CR}$	$\%NA_{MC}$
ALB	2609	69	787.0	438.0	83.4	446.0	8.0	1.9	11.4	2.6
AUS	7705	70	2367.1	491.7	92.9	501.8	7.3	2.4	10.3	2.5
AUT	3731	70	1133.7	499.1	92.7	509.6	8.4	1.8	12.5	2.0
BEL	4696	70	1393.0	507.8	95.6	518.6	8.3	2.6	11.9	2.5
BIH	3512	70	1071.0	406.5	82.0	413.1	**18.8**	**3.9**	27.9	4.2
BLR	3141	70	967.8	470.7	92.4	480.0	7.8	2.4	11.4	2.1
BRN	2812	69	845.0	430.6	91.3	438.2	6.1	1.7	8.8	1.8
CAN	9782	70	2786.3	511.7	92.4	522.7	5.8	2.2	8.2	2.1
CHE	3141	70	964.5	514.5	93.4	525.6	8.2	2.5	11.9	2.2
CZE	3798	70	1164.0	498.5	93.4	509.0	9.2	2.0	13.9	1.9
DEU	3000	70	908.6	499.0	95.9	509.5	9.6	2.5	14.0	2.4
DNK	4354	70	1250.1	510.7	81.3	521.7	5.9	2.0	8.6	1.7
ESP	14768	70	4408.3	481.7	88.3	491.5	**10.6**	2.9	15.5	2.7
EST	2880	70	890.9	523.8	81.6	535.3	6.6	2.0	9.6	1.8
FIN	3056	70	935.0	505.7	83.3	516.4	8.9	3.0	12.8	2.7
FRA	3405	70	1046.8	495.5	92.2	505.8	**10.1**	3.0	14.8	2.6
GBR	7063	70	2174.0	502.1	92.9	512.7	8.2	2.5	11.8	2.5
GRC	2634	70	790.4	451.1	89.5	459.6	**10.7**	2.7	15.7	2.6
HKG	2484	70	748.0	551.0	92.5	563.5	**3.9**	0.8	5.8	0.8
HRV	2683	70	805.2	464.5	87.1	473.6	**11.8**	2.7	17.6	2.5
HUN	2785	70	857.4	482.3	91.2	492.0	8.6	2.0	13.0	1.7
IRL	3031	70	935.5	500.2	78.1	510.7	5.8	1.3	8.7	1.2
ISL	1807	70	545.1	493.8	90.8	504.1	9.7	**4.4**	12.9	4.5
ISR	2825	70	846.6	464.0	107.5	473.0	**12.1**	**4.5**	16.9	4.5
ITA	6401	70	1978.9	485.9	94.0	495.8	**12.4**	2.8	18.9	2.1
JPN	3302	70	1018.6	527.4	87.1	539.1	8.4	1.9	12.9	1.4
KOR	2741	70	823.1	525.9	100.4	537.5	6.4	1.7	9.4	1.6
LTU	2824	70	846.3	480.1	90.0	489.8	7.4	1.5	11.4	1.1
LUX	2827	70	872.0	481.3	98.6	491.0	**10.4**	2.8	15.3	2.7
LVA	2190	70	656.4	498.5	80.5	509.0	6.4	1.7	9.7	1.1
MLT	1383	69	415.8	469.5	101.6	478.8	9.8	**3.9**	13.5	3.6
MNE	3595	70	1109.7	430.8	83.0	438.4	**17.3**	**3.8**	25.9	3.5
MYS	3284	70	1000.8	440.2	82.0	448.2	**1.2**	0.6	1.5	0.7
NLD	2939	70	742.6	518.4	92.9	529.7	**4.4**	1.1	6.7	0.9
NOR	3141	70	969.5	502.4	90.3	513.0	**10.7**	**3.7**	15.1	3.7
NZL	3309	70	1021.2	495.6	93.0	506.0	8.1	2.2	11.7	2.3
POL	3022	70	932.6	515.8	90.5	526.9	7.1	1.9	10.7	1.3
PRT	3202	70	987.6	493.1	96.2	503.3	**10.6**	2.8	15.8	2.3
RUS	3131	70	939.3	487.8	87.4	497.8	7.9	2.2	11.6	2.1
SGP	2732	70	822.3	570.3	93.3	583.6	**2.7**	0.8	3.8	0.8
SVK	2514	70	727.9	484.6	100.1	494.5	8.0	1.8	11.9	1.7
SVN	3519	70	1054.7	509.5	88.7	520.4	7.1	1.5	10.7	1.4
SWE	2982	70	918.7	502.8	90.3	513.4	**12.7**	**5.9**	17.3	5.4
TUR	3723	70	1147.8	453.4	87.4	462.0	6.7	1.6	9.7	1.8
USA	2629	70	804.9	478.0	92.4	487.6	**4.0**	2.0	5.2	1.9

Note. N = number of students; I = number of items; N_{item} = average number of students per item; M_{OECD} = officially reported country mean by OECD [24]; M_{OECD} = officially reported country standard deviation by OECD [24]; M_{stand} = standardized country mean ($M = 500$ and $SD = 100$ in total population); %NA = proportion of item responses with missing data; %NR = proportion of item responses that are not reached; $\%NA_{CR}$ = proportion of constructed-response item responses with missing data; $\%NA_{MC}$ = proportion of multiple-choice item responses with missing data; Missing item response rates larger than 10.0% and smaller than 5.0% are printed in bold. Missing rates for not reached responses larger than 3.0% are printed in bold. See Appendix A for country labels.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

and $P(X_{pi} = 0)$ with $1 - 1/K_i$ [28]. This weighing implements a scoring of a missing MC item as partially correct (model UP). The maximum likelihood estimation is replaced by a pseudo-likelihood estimation that allows non-integer item responses [28]. The estimation was conducted in the R [29] package sirt [30].

Pseudo-likelihood estimation of IRT models that allow non-integer item responses is not widely implemented in IRT software. However, the partially correct scoring can be alternatively implemented by employing a multiple imputation approach of item responses. For every missing item response of item i, a correct item response is imputed with probability $1/K_i$. In our analysis, we created 10 imputed datasets to reduce the simulation error associated with the imputation. We stack the 10 multiply imputed datasets into one long dataset and applied the 2PL scaling model (see Equation (2)) for the stacked dataset (see [31–33]). The stacking approach does not result in biased item parameter estimates [32], but resampling procedures are required for obtaining correct standard errors [31]. In this article, we mainly focused on differences between results from different models and did not implement resampling procedures for computing standard errors.

Missing item responses for CR items are scored as incorrect in the partially correct scoring approach because unknown answers cannot be simply guessed by students in this situation.

3.1.3. Treating Not Reached Items as Ignorable

Since PISA 2015, not reached items are no longer scored as wrong [10]. To investigate this scaling method, we ignored not reached items in the scaling model but scored omitted items as incorrect (model UN1). We also implemented the operational practice since PISA 2015 [10] that includes the proportion of not reached item response as a covariate in the latent background model (model UN2; [9,34]). This second model is equivalent to latent ignorability when the response indicators for not reached items follow a 1PL model.

3.1.4. Treating Missing Item Responses as Ignorable

In model UO1, all missing item responses are ignored in the scaling model. The student ability θ is extracted based on the observed item responses only. The method is valid if missing item responses can be regarded as ignorable [12]. In this case, the probability of omitting items only depends on observed items and not the unobserved item responses.

3.1.5. Treating Missing Item Responses as Latent Ignorable

A weak variant of nonignorable missing data is latent ignorability [13,35–43]. Observed item responses X_{pi} and response indicators R_{pi} are jointly modeled conditional on the latent ability θ_p and the latent response propensity ξ_p. The probability of responding to an item is given by (model MO2; [7,14,44–46])

$$P(R_{pi} = 1|\theta_p, \xi_p) = \Psi(\xi_p - \beta_i) \quad . \tag{3}$$

The 2PL model holds for item responses (see Equation (1)). A joint bivariate distribution (θ_p, ξ_p) is modeled. In this study, a bivariate normal distribution is assumed, where $SD(\theta_p)$ is fixed to one, and $SD(\xi_p)$, as well as $Cor(\theta_p, \xi_p)$, are estimated (see [47] for more complex distributions). The model UO1 (see Section 3.1.4) that presupposes ignorability (instead of latent ignorability) can be tested as a nested model within model MO2 by setting $Cor(\theta_p, \xi_p) = 0$. This model is referred to as model MO1.

The model for response indicators R_{pi} in Equation (3) is a 1PL model. Hence, the sum score $R_{p\bullet} = \sum_{i=1}^{I} R_{pi}$ is a sufficient statistic for ξ_p. Instead of estimating a joint distribution (θ_p, ξ_p), a conditional distribution $\theta_p|R_{p\bullet}$ can be specified in a latent background model. In our study, the proportion of missing item responses is used as a predictor in the latent background model (model UO2, [9]).

The models MO1 and MO2 are also used for generating multiply imputed datasets. Conditional and θ_p, missing item responses are imputed according to the response proba-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

bility from the 2PL model (see Equation (1)). The stacked imputed dataset (see Section 3.1.2) can be scaled with the unidimensional 2PL model. If models MO1 or MO2 would be the true data-generating models, results from multiple imputation (i.e., IO1 and IO2) would coincide with model-based treatments (i.e., MO1 and MO2). However, results can differ in the case of misspecified models [48,49].

3.1.6. Mislevy-Wu Model for Nonignorable Item Responses

Latent ignorability characterizes only a fragile nonignorable missing data process. It might be more plausible that the probability of responding to an item depends on the observed or unobserved item response itself [50–54]. The so-called Mislevy-Wu model [22,55] extends the model MO2 (Equation (3)) that assumes latent ignorability to

$$\mathrm{P}(R_{pi} = 1 | X_{pi}, \theta_p, \xi_p) = \Psi(\xi_p - \beta_i - \delta_i X_{pi}) \quad . \tag{4}$$

In this model, the probability of responding to an item depends on the latent response propensity ξ_p and the item response X_{pi} itself (see [18,19,30,56–58]). Model MM1 is defined by assuming a common δ_i parameter for all items. In model MM2, two δ parameters are estimated for item formats CR and MC. For both models, multiply imputed datasets were also created based on conditional distributions $\mathrm{P}(X_{pi}|R_{pi}, \theta_p, \xi_p)$. The scaling models based on stacked imputed datasets are referred to as IM1 and IM2.

The most salient property of the models MM1 and MM2 is that the model treating missing item responses as wrong (model UW) can be tested by setting $\delta_i = -10$, resulting in a response probability of approximately one ([23]; see also [59]). This model is referred to as model MW and the corresponding scaling model based on imputations as IW. Moreover, the model MO2 assuming latent ignorability is obtained by setting $\delta_i = 0$ for all items i. It has been shown that model selection among models MW, MO2, and MM1 can be satisfactorily conducted utilizing information criteria [23].

3.1.7. Imputation Models Based on Fully Conditional Specification

The imputation models discussed above are based on unidimensional or two-dimensional IRT models. Posing such a dimensionality reduction might result in invalid imputations because almost all IRT models in large-scale assessment studies are misspecified [20]. Hence, two alternative imputation models for missing item responses were considered that relied on fully conditional specification (FCS; [32]) implemented in the R package mice [60].

Previous research indicated that item parameters are affected by position effects [61–68]. Hence, the FCS imputation is separately conducted for each test booklet. In the imputation model IF1, only item responses were included. Linear regression with predictive mean matching (PMM; [32]) was used as the imputation model. In each iteration and for each imputation model, the predictors (item responses except the item that is imputed) are transformed using ten factors obtained from partial least squares regression to avoid the curse of dimensionality due to estimating too many parameters in the regression models [69,70].

In model IF2, response indicators were additionally included [71]. In contrast to the Mislevy-Wu model, for imputing item response X_{pi}, the set of predictors X_{pj}, R_{pj} $(j \neq i)$ were used. Hence, the probability of responding to an item is not allowed to depend on the item itself. This assumption might be less plausible than assuming the response model in Equation (4). Like in model IF1, ten factors from partial least squares regression were used for reducing the dimension of the covariate space in the conditional imputation models and PMM was utilized.

Like for all multiple imputations in our study, 10 imputed datasets were created, and the 2PL scaling model is applied to the stacked dataset involving all imputed datasets (see Section 3.1.2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

3.2. Linking Procedure

The scaling models described above resulted in country-specific item discriminations a_{ic} and item difficulties b_{ic}. To enable a comparison of country means, the corresponding ability distributions can be obtained by linking approaches that establish a common ability metric [72,73]. In this article, Haberman linking [74] in its original proposal is used (see also [75,76]). The outcome of the linking procedure are country means and standard deviations. To enable a comparisons across the 19 specified different scaling models, the ability distributions were linearly transformed such that the total population involving all students in all countries in our study has a mean M = 500 and a standard deviation SD = 100.

3.3. Model Comparisons

It is of particular interest whether the Mislevy-Wu model (MM1 and MM2) outperforms other treatments of missing item responses such as the scoring as wrong (model MW) and latent ignorable (models MO1 and MO2). The Bayesian information criterion (BIC) is used for conducting model comparisons ([23]; see also [24,77–79] for similar model comparisons in PISA, but [80–82] for improved information criteria in complex surveys). Moreover, the Gilula-Haberman penalty (GHP; [83,84]) is used as an effect size that is relatively independent of the sample size and the number of items. A difference in GHP larger than 0.001 is declared a notable difference in model fit [84,85].

4. Results

4.1. Similarity of Scaling Models

Each of the 19 scaling models provided a set of country means. For each country, the absolute difference of the means stemming from two models can be computed. Table 2 summarizes the average absolute differences. Scaling models that resulted in an average absolute difference of at most 1.0 can be considered similar.

Table 2 indicates that the methods that treat missing item responses as wrong (UW, MW, IW) or treat MC items as partially correct (UP, IP) resulted in similar country mean estimates. Both methods that did not score not reached item responses as wrong (UN1, UN2) resulted in relatively similar estimates. The models that rely on ignorability (UO1, MO1, IO1) or latent ignorability (MO2, UO2, IO2) provided similar estimates. In line with previous research [12], the inclusion of the latent response propensity ξ did not result in strongly different estimates of country means compared to models that ignore missing item responses. The specifications of the Mislevy-Wu model (MM1, IM1, MM2, IM2) resulted in similar country means. Interestingly, country means from the Mislevy-Wu model were more similar to the treatment of missing item responses as incorrect than those that relied on ignorability or latent ignorability. Finally, the scaling model based on FCS imputation involving only item responses (IF1) was similar to the models assuming (latent) ignorability (UO1, MO1, IO1, MO2, UO2, IO2). FCS imputation involving item responses and response indicators different from the imputed item (IF2) were neither similar to the ignorability-based treatment nor the incorrect scoring method or the Mislevy-Wu model. This finding could be explained by the fact that the imputation method IF2 is based on strongly opposing assumptions of the missingness mechanism than the Mislevy-Wu model.

4.2. Model Comparisons

From Table 3, we can see that for the majority of countries (35 out of 45), the IRT model treating missing item responses as incorrect (model MW) provided a better model fit in terms of BIC than modeling it with a latent propensity (model MO2). For 39 out of 45 countries, the Mislevy-Wu model with item-format specific ρ parameters (model MM2) was preferred. In 5 out of 45 countries, the Mislevy-Wu model with one common ρ parameter (MM1) was the best-fitting model. Only in one country (MYS), the model treating missing item responses as wrong had the best model fit.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

Table 2. Average absolute differences in country means of different treatments of missing item responses

	UW	MW	IW	UP	IP	UN1	UN2	UO1	MO1	IO1	MO2	UO2	IO2	MM1	IM1	MM2	IM2	IF1	IF2
UW	—	**0.3**	**0.0**	**0.7**	**0.8**	1.9	1.7	3.0	3.0	3.0	2.6	2.8	2.6	1.4	1.5	1.6	1.5	3.0	2.8
MW	**0.3**	—	**0.3**	**0.9**	**0.9**	2.0	1.7	3.0	3.0	3.0	2.7	2.8	2.6	1.4	1.6	1.6	1.6	3.1	2.8
IW	**0.0**	**0.3**	—	**0.7**	**0.8**	1.9	1.7	3.0	3.0	3.0	2.6	2.8	2.6	1.4	1.5	1.6	1.5	3.0	2.8
UP	**0.7**	**0.9**	**0.7**	—	**0.3**	1.4	1.5	2.7	2.7	2.7	2.4	2.4	2.3	1.1	1.2	1.3	1.2	2.7	2.6
IP	**0.8**	**0.9**	**0.8**	**0.3**	—	1.5	1.5	2.7	2.7	2.7	2.4	2.4	2.3	1.2	1.2	1.4	1.3	2.7	2.6
UN1	1.9	2.0	1.9	1.4	1.5	—	**1.0**	2.1	2.1	2.1	2.0	1.8	1.9	**1.0**	**1.0**	**0.9**	**0.9**	2.2	2.6
UN2	1.7	1.7	1.7	1.5	1.5	**1.0**	—	2.4	2.5	2.5	2.0	2.2	2.0	1.4	1.4	1.2	1.3	2.6	2.7
UO1	3.0	3.0	3.0	2.7	2.7	2.1	2.4	—	**0.0**	**0.2**	**0.7**	**0.3**	**0.6**	2.5	2.5	2.2	2.3	**0.7**	1.9
MO1	3.0	3.0	3.0	2.7	2.7	2.1	2.5	**0.0**	—	**0.2**	**0.7**	**0.4**	**0.7**	2.5	2.5	2.2	2.3	**0.7**	1.9
IO1	3.0	3.0	3.0	2.7	2.7	2.1	2.5	**0.2**	**0.2**	—	**0.7**	**0.4**	**0.7**	2.6	2.5	2.3	2.4	**0.8**	1.9
MO2	2.6	2.7	2.6	2.4	2.4	2.0	2.0	**0.7**	**0.7**	**0.7**	—	**0.6**	**0.4**	2.2	2.3	1.8	2.0	**1.0**	1.8
UO2	2.8	2.8	2.8	2.4	2.4	1.8	2.2	**0.3**	**0.4**	**0.4**	**0.6**	—	**0.5**	2.3	2.2	2.0	2.1	**0.8**	1.8
IO2	2.6	2.6	2.6	2.3	2.3	1.9	2.0	**0.6**	**0.7**	**0.7**	**0.4**	**0.5**	—	2.2	2.2	1.8	2.0	**1.0**	1.8
MM1	1.4	1.4	1.4	1.1	1.2	**1.0**	1.4	2.5	2.5	2.6	2.2	2.3	2.2	—	**0.4**	**0.6**	**0.5**	2.6	2.7
IM1	1.5	1.6	1.5	1.2	1.2	**1.0**	1.4	2.5	2.5	2.5	2.3	2.2	2.2	**0.4**	—	**0.8**	**0.7**	2.6	2.6
MM2	1.6	1.6	1.6	1.3	1.4	**0.9**	1.2	2.2	2.2	2.3	1.8	2.0	1.8	**0.6**	**0.8**	—	**0.4**	2.3	2.5
IM2	1.5	1.6	1.5	1.2	1.3	**0.9**	1.3	2.3	2.3	2.4	2.0	2.1	2.0	**0.5**	**0.7**	**0.4**	—	2.4	2.5
IF1	3.0	3.1	3.0	2.7	2.7	2.2	2.6	**0.7**	**0.7**	**0.8**	**1.0**	**0.8**	**1.0**	2.6	2.6	2.3	2.4	—	1.9
IF2	2.8	2.8	2.8	2.6	2.6	2.6	2.7	1.9	1.9	1.9	1.8	1.8	1.8	2.7	2.6	2.5	2.5	1.9	—

Note. UW = scoring as wrong (Sect. 3.1.1) ; MW = model-based treatment as wrong (Sect. 3.1.6, Eq. (4) with $\rho_i = -10$); IW = imputed as wrong (Sect. 3.1.6, Eq. (4) with $\rho_i = -10$); UP = MC items scored as partially correct (Sect. 3.1.2); IP = MC items imputed as partially correct (Sect. 3.1.2); UN1 = ignoring not reached items (Sect. 3.1.3); UN2 = including proportion of not reached items in background model (Sect. 3.1.3); UO1 = ignoring missing items (Sect. 3.1.4); MO1 = model-based ignorability (Sect. 3.1.5, Eq. (3) with $\text{Cor}(\theta, \xi) = 0$); IO1 = imputed under ignorability (Sect. 3.1.5, Eq. (3) with $\text{Cor}(\theta, \xi) = 0$); MO2 = model-based latent ignorability (Sect. 3.1.5, Eq. (3)); UO2 = including proportion of missing items in background model (Sect. 3.1.5); IO2 = imputed under latent ignorability (Sect. 3.1.5, Eq. (3)); MM1 = Mislevy-Wu model with common δ parameter (Sect. 3.1.6, Eq. (4)); IM1 = imputed under Mislevy-Wu model with common δ parameter (Sect. 3.1.6, Eq. (4)); MM2 = Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); IM2 = imputed under Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); IF1 = FCS imputation based on item responses (Sect. 3.1.7); IF2 = FCS imputation based on item responses and response indicators (Sect. 3.1.7); Mean absolute differences smaller or equal than 1.0 are printed in bold.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

Table 3. Model comparisons based on the Bayesian information crierion (BIC) and the Gilula-Haberman penalty (GHP)

	BIC					GHP					
Country	MW	MO1	MO2	MM1	MM2	MW	MO1	MO2	MM1	MM2	Diff
ALB	63663	63754	63600	**63579**	63586	0.6423	0.6433	0.6416	0.6414	0.6414	0.0003
AUS	193304	194008	193316	193145	**193105**	0.6321	0.6344	0.6321	0.6315	0.6314	0.0007
AUT	97019	97685	97174	97007	**96993**	0.6618	0.6664	0.6628	0.6616	0.6615	**0.0013**
BEL	118264	119131	118426	118236	**118186**	0.6665	0.6715	0.6675	0.6664	0.6660	**0.0014**
BIH	98447	98779	98534	98371	**98359**	0.7101	0.7125	0.7107	0.7095	0.7093	**0.0014**
BLR	82460	82729	82564	82455	**82396**	0.6509	0.6531	0.6517	0.6508	0.6503	**0.0014**
BRN	62751	62864	62756	**62715**	62719	0.5925	0.5936	0.5925	0.5921	0.5921	0.0005
CAN	213551	214215	213549	213382	**213268**	0.6316	0.6336	0.6316	0.6311	0.6307	0.0009
CHE	84792	85329	84940	84777	**84743**	0.6724	0.6768	0.6736	0.6722	0.6719	**0.0017**
CZE	102441	102838	102508	102382	**102301**	0.6780	0.6807	0.6784	0.6776	0.6770	**0.0015**
DEU	79134	79714	79219	79118	**79102**	0.6729	0.6779	0.6736	0.6727	0.6725	**0.0011**
DNK	97368	97632	97328	**97270**	97277	0.6232	0.6249	0.6229	0.6225	0.6225	0.0004
ESP	377203	378998	377528	377027	**376832**	0.6844	0.6877	0.6850	0.6841	0.6837	**0.0013**
EST	74697	74921	74716	74639	**74623**	0.6384	0.6404	0.6386	0.6379	0.6377	0.0009
FIN	80421	80504	80386	80315	**80228**	0.6602	0.6609	0.6599	0.6592	0.6585	**0.0014**
FRA	92877	93593	93019	92868	**92833**	0.6820	0.6874	0.6830	0.6819	0.6816	**0.0015**
GBR	181680	182770	181704	181518	**181471**	0.6457	0.6496	0.6458	0.6451	0.6449	0.0009
GRC	68339	68606	68485	68317	**68269**	0.6814	0.6841	0.6829	0.6811	0.6805	**0.0023**
HKG	57050	57459	57113	57054	**57048**	0.5965	0.6009	0.5972	0.5965	0.5964	0.0008
HRV	70685	71044	70791	70679	**70669**	0.6927	0.6963	0.6937	0.6926	0.6924	**0.0013**
HUN	72125	72492	72187	72080	**72060**	0.6437	0.6470	0.6442	0.6432	0.6430	**0.0013**
IRL	77409	77712	77432	77381	**77369**	0.6323	0.6349	0.6325	0.6320	0.6319	0.0006
ISL	48098	48071	48043	48006	**47965**	0.6782	0.6779	0.6774	0.6768	0.6761	**0.0013**
ISR	62551	62964	62675	62531	**62520**	0.6771	0.6817	0.6785	0.6768	0.6766	**0.0018**
ITA	179041	180275	179253	178956	**178914**	0.6951	0.6999	0.6959	0.6947	0.6945	**0.0014**
JPN	87938	88375	87998	87917	**87858**	0.6606	0.6639	0.6610	0.6604	0.6599	**0.0012**
KOR	65114	65613	65110	65067	**65066**	0.6229	0.6278	0.6229	0.6224	0.6223	0.0005
LTU	68816	69098	68893	68797	**68788**	0.6411	0.6439	0.6419	0.6409	0.6408	**0.0011**
LUX	79066	79552	79236	79051	**79033**	0.6933	0.6976	0.6948	0.6931	0.6929	**0.0019**
LVA	53764	53922	53754	53731	**53728**	0.6441	0.6461	0.6439	0.6436	0.6435	0.0005
MLT	33418	33625	33404	**33370**	33371	0.6325	0.6367	0.6323	0.6315	0.6314	0.0008
MNE	103907	104412	104044	103857	**103833**	0.7174	0.7210	0.7183	0.7170	0.7168	**0.0016**
MYS	**66244**	66271	66256	66246	66253	0.5042	0.5045	0.5043	0.5042	0.5042	0.0001
NLD	50077	50286	50125	50063	**50055**	0.5869	0.5895	0.5875	0.5867	0.5865	**0.0010**
NOR	86955	87260	87005	86842	**86802**	0.6859	0.6884	0.6863	0.6850	0.6846	**0.0017**
NZL	87003	87519	87077	86965	**86951**	0.6514	0.6554	0.6520	0.6511	0.6509	**0.0010**
POL	78675	78987	78675	78616	**78599**	0.6441	0.6468	0.6441	0.6436	0.6434	0.0007
PRT	89473	89900	89627	89457	**89322**	0.6933	0.6967	0.6945	0.6931	0.6920	**0.0025**
RUS	78318	78563	78384	78290	**78262**	0.6588	0.6610	0.6594	0.6586	0.6583	**0.0011**
SGP	58480	58724	58515	**58466**	58466	0.5576	0.5600	0.5579	0.5574	0.5573	0.0006
SVK	59699	59958	59788	59692	**59671**	0.6593	0.6622	0.6602	0.6591	0.6588	**0.0014**
SVN	88287	88818	88451	88292	**88245**	0.6518	0.6558	0.6530	0.6518	0.6514	**0.0016**
SWE	86292	86416	86272	86145	**86037**	0.7188	0.7199	0.7187	0.7175	0.7166	**0.0021**
TUR	96064	96326	96230	96041	**96032**	0.6412	0.6430	0.6423	0.6410	0.6409	**0.0014**
USA	61234	61223	61167	61154	**61147**	0.5806	0.5806	0.5800	0.5798	0.5797	0.0003

Note. MW = model-based treatment as wrong (Sect. 3.1.6, Eq. (4) with $\rho_i = -10$); MO1 = model-based ignorability (Sect. 3.1.5, Eq. (3) with $\text{Cor}(\theta, \xi) = 0$); MO2 = model-based latent ignorability (Sect. 3.1.5, Eq. (3)); MM1 = Mislevy-Wu model with common δ parameter (Sect. 3.1.6, Eq. (4)); MM2 = Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); BIC values for best-performing model printed in bold. GHP differences larger than 0.001 printed in bold. See Appendix A for country labels.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

For 29 out of 45 countries, the proposed Mislevy-Wu model outperformed the suggested model with a latent response propensity in terms of a GHP difference of at least .001. Overall, these findings indicated that the models assuming ignorability or latent ignorability performed worse in terms of model fit compared to scaling models that acknowledge the dependence of responding to an item from the true but occasionally unobserved item response.

4.3. Country-Specific Model Parameters for Latent Ignorable and Mislevy-Wu Model

Now, we present findings of model parameters characterizing the missingness mechanism from the model MO2 relying on latent ignorability and the Mislevy-Wu model MM2. The parameters are shown in Table 4. The SD of the latent response propensity $\mathrm{SD}(\xi)$ was somewhat lower in the Mislevy-Wu model (MM2, with a median Med $= 1.98$) than the model assuming latent ignorability (MO2, Med $= 1.93$). Moreover, by additionally including the latent item response as a predictor for the response indicator, the correlation $\mathrm{Cor}(\theta, \xi)$ between the latent ability θ and response propensity ξ was slightly lower in model MM2 (Med $= .43$) than MO2 (Med $= .46$). Most importantly, the missingness mechanism strongly differed between CR and MC items. The median δ parameter in model MM2 for CR items was -2.61, indicating that students that did not know the item had a higher probability of omitting the item even after controlling for the latent response propensity ξ. In contrast, the median δ parameter was -0.48. Hence, there was a smaller influence of (latently) knowing the item with the response indicators. However, it was different from zero for most countries, indicating that the model MO2 assuming latent ignorability did not adequately explain the missingness mechanism. Overall, it can be seen that model parameters strongly vary across countries. Hence, it can be concluded that assuming different missingness mechanisms for countries could have non-negligible consequences for country rankings (see [86]).

4.4. Country Means Obtained From Different Scaling Models

For comparing country means, 11 out of 19 specified scaling models were selected to contrast the dissimilarity of country mean estimates. Table 5 shows the country means of these 11 different treatments of missing item responses. The country rank (column "$\mathrm{rk}_{\mathrm{UW}}$") serves as the reference for the comparison among methods. Moreover, the interval of country ranks obtained from the different methods are shown in column "$\mathrm{rk}_{\mathrm{Int}}$". The average maximum difference in country ranks was 2.4 (SD $= 1.8$) and ranged between 0 (SGP, HKG, EST, DEU, LUX, BIH) and 8 (IRL). The range in country means (i.e., the difference of the largest and smallest country mean of the 11 methods) was noticeable (M $=$ 5.0) and showed strong variability between countries (SD $= 2.8$, Min $= 1.5$, Max $= 12.5$). Interestingly, large range values were obtained for countries with missing proportions that were strongly below and above the average missing proportion. For example, Ireland (IRL) had a relatively low missing rate of 5.8% and reached rank 15 with method UW (M $= 505.2$) that treated missing item responses as incorrect. Methods that ignore missing item responses resulted in a lower country mean (UO1: M $= 499.9$; MO2: M $= 500.7$; IO2: M $= 500.0$). In contrast, the Mislevy-Wu model (MM2 and IO2)—which also takes the relation of the response indicator and the true item response into account—resulted in higher country means (MM2: M $= 505.1$; IO2: M $= 504.9$). Across the 11 estimation methods, Ireland reached ranks between 15 and 23 which can be considered a large variability. Moreover, the range of country means for Ireland was 8.2, which is two to three times higher than standard errors for country means due to the sampling of students in PISA. Italy (ITA, rank 26; M $= 492.0$) that had a relatively high missing rate of 12.4% profit by ignoring missing item responses assuming latent ignorability (UO1: M $= 494.7$; MO2: M $= 494.4$; IO2: M $= 494.0$). However, the Mislevy-Wu model produced considerably lower scores (MO2: M $= 490.1$; IO2: M $= 489.9$). An interesting case is Sweden (SWE, rank 25) that had a high missing proportion rate of 12.7%, but almost half of missing item responses (i.e., 5.9%) stemmed from not reached responses. This not reached proportion

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

Table 4. Model parameters from the latent ignorable model (MO2) and the Mislevy-Wu Model (MM2)

	MO2		MM2			
Country	$SD(\xi)$	$Cor(\theta,\xi)$	$SD(\xi)$	$Cor(\theta,\xi)$	δ_{CR}	δ_{MC}
ALB	2.50	.42	2.47	.44	−1.23	−0.91
AUS	2.59	.46	2.52	.46	−2.31	−0.71
AUT	1.90	.54	1.79	.49	−3.42	−1.01
BEL	1.92	.56	1.83	.51	−3.10	−0.43
BIH	1.87	.40	1.82	.43	−2.12	−0.53
BLR	1.81	.35	1.79	.29	−2.95	0.43
BRN	2.21	.33	2.17	.33	−2.08	−1.08
CAN	2.30	.44	2.26	.41	−2.37	−0.09
CHE	1.91	.50	1.83	.44	−3.12	−0.46
CZE	1.73	.43	1.68	.35	−2.46	0.46
DEU	1.91	.57	1.80	.53	−2.63	−0.48
DNK	2.25	.43	2.19	.43	−1.73	−1.32
ESP	1.83	.47	1.77	.45	−2.45	−0.01
EST	2.10	.41	2.06	.36	−2.43	−0.35
FIN	1.99	.31	2.00	.28	−2.22	0.57
FRA	1.85	.57	1.74	.52	−3.19	−0.52
GBR	2.48	.57	2.38	.56	−2.26	−0.41
GRC	1.80	.33	1.78	.30	−3.59	−0.24
HKG	2.34	.60	2.22	.52	−4.07	−0.67
HRV	1.89	.46	1.83	.45	−2.99	−0.64
HUN	2.17	.49	2.11	.45	−2.48	−0.15
IRL	1.97	.47	1.91	.44	−2.23	−0.01
ISL	2.35	.22	2.36	.23	−2.00	0.06
ISR	2.36	.50	2.26	.49	−3.04	−1.28
ITA	1.75	.54	1.65	.49	−2.69	−0.49
JPN	1.92	.49	1.84	.43	−2.67	0.45
KOR	2.61	.64	2.49	.62	−2.15	−0.80
LTU	1.89	.42	1.84	.36	−3.20	−0.69
LUX	1.76	.47	1.68	.41	−3.01	−0.73
LVA	1.98	.44	1.93	.41	−1.86	−0.08
MLT	2.94	.61	2.86	.62	−2.03	−0.82
MNE	1.86	.47	1.81	.49	−2.61	−0.57
MYS	2.42	.18	2.43	.15	−1.94	−2.76
NLD	2.37	.45	2.32	.40	−3.07	−0.61
NOR	2.11	.42	2.05	.41	−2.64	−0.64
NZL	2.19	.53	2.09	.50	−2.56	−0.60
POL	2.05	.48	1.99	.42	−2.12	−0.08
PRT	1.76	.42	1.72	.34	−2.72	1.20
RUS	2.00	.38	1.97	.35	−2.79	−0.28
SGP	2.51	.50	2.43	.44	−2.80	−1.11
SVK	1.93	.41	1.88	.36	−3.15	−0.23
SVN	1.85	.49	1.77	.42	−9.99	−0.34
SWE	1.90	.32	1.89	.30	−2.24	0.01
TUR	1.71	.26	1.68	.18	−4.07	−1.40
USA	2.72	.26	2.70	.26	−1.54	−0.28

Note. MO2 = model-based latent ignorability (Sect. 3.1.5, Eq. (3)); MM2 = Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); $SD(\xi)$ = standard deviation of latent propensity variable ξ; $Cor(\theta,\xi)$ = correlation of latent ability θ with latent propensity variable ξ; δ_{CR} = common δ parameter for constructed response items (see Equation xxx); δ_{MC} = common δ parameter for multiple-choice items (see Equation (4)). See Appendix A for country labels.

 doi:10.20944/preprints202110.0107.v1

was the highest among all countries in our study. Sweden had rank 25 when treating missing item responses as incorrect (UW: M = 491.8), but strongly profits in models that ignore the not reached items (UN1: M = 499.1) or treated the proportion of not reached items as a predictor in the latent background model (UN2: M = 499.7). If also omitted items would be treated as (latent) ignorable, the country mean for Sweden further increased (UO1: M = 501.3; MO2: M = 501.1; IO2: M = 501.3). In contrast to many other countries, the country means obtained from the Mislevy-Wu model (MM2: M = 497.9; IO2: M = 498.0) were also much larger than the country mean obtained by treating missing items as incorrect (UW: M = 491.8).

5. Discussion

In this reanalysis of the PISA 2018 mathematics data, different scaling models with different treatments of missing item responses were specified. It has been shown that differences in country means across models can be substantial. The present study sheds some light on the ongoing debate about how to properly handling missing item responses in educational large-scale assessment studies. Ignoring missing item responses and treating them as wrong can be seen as opposing strategies. Other scaling models can be interpreted to provide results somewhere between these two extreme poles of handling missingness. We argued that the Mislevy-Wu model contains the strategy of incorrect scoring and the latent ignorable model as submodels. Hence, these missing data treatments can be tested. In our analysis, it turned out that the Mislevy-Wu model fitted the PISA data best. More importantly, the treatment of missing item responses as incorrect provided a better model fit than ignoring them or modeling them by the latent ignorable model that has been strongly advocated in the past [7,8]. It also turned out that the missingness mechanism strongly differed between CR and MC items.

We believe that the call for controlling for test-taking behavior in the reporting in large-scale assessment studies such as response propensity [3] using models that also include response times [87,88] poses a threat to validity because results can be simply manipulated by instructing students to omit items they do not know [20]. Notably, missing item responses are mostly omissions for CR items. Response times might be useful for detecting rapid guessing [56,89–92]. However, it seems likely that students who do not know the solution to CR items do not respond to these items. In this case, the latent ignorability assumption is unlikely to hold, and scaling models that rely on it (see [3,9]) will result in conflated and unfair country comparisons.

In this article, we only investigated the impact of missing item responses on country means. In LSA studies, missing data is also a prevalent issue for student covariates (e.g., sociodemographic status; see [69,93–97]). As covariates also enter the plausible value imputation of latent abilities through the latent background model [34,98] or relationships of abilities and covariates are often of interest in reporting, missing data on covariates is also a crucial issue that needs to be adequately addressed [69].

It could be argued that there is not a unique scientifically sound or publicly accepted scaling model in PISA. The uncertainty in choosing a psychometric model can be reflected by explicitly acknowledging the variability of country means obtained by different model assumptions. This additional source of variance associated with model uncertainty [99–104] can be added to the standard error due to the sampling of students and linking error due to the selection of items [105]. The assessment of specification uncertainty has been discussed in sensitivity analysis [106] and has recently become popular as multiverse analysis [107,108] or specification curve analysis [109,110]. As educational LSA studies are policy-relevant, we think that model uncertainty should be included in statistical inference.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2021 doi:10.20944/preprints202110.0107.v1

Table 5. Country means for PISA 2018 mathematics from 11 different scaling models for missing item responses

Country	%NA	%NR	rk_{UW}	rk_{Int}	Aver	SD	rg	UW	UP	UN1	UN2	UO1	MO2	IO2	MM2	IM2	IF1	IF2
SGP	**2.7**	0.8	1	1–1	568.1	1.5	**5.3**	568.0	567.8	567.6	567.4	567.7	567.0	567.7	567.3	567.8	568.7	572.4
HKG	**3.9**	0.8	2	2–2	548.9	1.3	4.1	550.1	550.0	548.3	548.3	548.2	548.0	547.9	548.3	548.4	548.8	552.0
NLD	**4.4**	1.1	3	3–4	531.4	0.6	2.1	531.6	531.5	531.7	531.6	530.9	530.7	531.1	531.2	531.5	530.9	532.9
JPN	8.4	1.9	4	3–4	532.1	1.8	4.6	530.8	530.6	530.0	530.3	533.8	533.9	533.9	531.1	531.0	533.5	534.6
EST	6.6	2.0	5	5–5	526.7	1.0	3.4	527.9	529.2	526.5	526.8	525.7	526.2	526.4	526.8	525.9	526.1	526.1
KOR	6.4	1.7	6	6–7	522.5	1.3	4.4	523.7	523.6	523.1	521.7	522.1	520.4	520.9	522.2	522.8	522.6	524.8
POL	7.1	1.9	7	6–8	521.5	0.7	2.5	521.4	521.2	520.9	521.0	521.2	521.1	520.9	521.7	521.8	521.5	523.4
CAN	5.8	2.2	8	7–8	520.6	0.8	2.8	519.5	519.9	521.4	521.4	520.1	519.9	519.9	521.0	521.1	520.5	522.2
DNK	5.9	2.0	9	9–10	518.4	0.8	2.3	518.1	518.2	519.4	519.4	517.6	518.1	517.5	519.4	518.9	517.1	518.6
SVN	7.1	1.5	10	10–12	515.2	0.8	2.3	516.4	516.0	514.3	514.9	514.7	515.3	515.7	514.3	514.1	515.2	515.9
BEL	8.3	2.6	11	9–11	517.2	0.7	2.3	516.1	516.7	516.9	517.2	517.4	518.1	517.0	516.7	516.7	517.6	518.4
CHE	8.2	2.5	12	11–12	514.5	0.5	1.5	514.2	514.8	514.0	514.4	514.9	515.2	514.8	513.9	513.7	515.2	514.0
DEU	9.6	2.5	13	13–13	509.8	1.0	3.1	509.1	509.1	509.2	509.2	511.4	511.5	510.4	509.8	509.4	509.8	508.4
FIN	8.9	3.0	14	14–16	506.7	1.0	3.7	506.9	506.5	506.7	507.3	506.5	506.9	506.8	508.0	507.9	506.0	504.3
IRL	5.8	1.3	15	**15–23**	502.2	2.6	**8.2**	505.2	504.8	501.6	502.1	499.9	500.7	500.0	505.1	504.9	497.1	502.5
CZE	9.2	2.0	16	**14–17**	505.1	1.0	3.7	504.9	504.3	503.4	504.2	505.3	505.8	505.8	505.1	505.0	505.5	507.1
GBR	8.2	2.5	17	**14–17**	505.6	1.0	3.3	503.9	504.9	506.6	504.6	507.2	505.7	505.0	505.8	505.8	506.6	505.4
NZL	8.1	2.2	18	**18–22**	502.4	1.2	4.0	503.3	504.3	502.5	502.0	501.8	501.6	501.6	502.4	504.8	501.9	500.8
FRA	**10.1**	3.0	19	**17–20**	502.8	0.9	2.4	502.1	502.2	502.5	503.0	503.9	503.9	503.9	501.8	501.5	503.1	503.3
AUT	8.4	1.8	20	**20–23**	500.8	0.9	2.7	500.9	501.7	500.1	499.4	501.9	500.7	501.4	499.6	500.4	501.0	502.1
PRT	**10.6**	2.8	21	**17–21**	501.9	1.2	3.8	500.1	500.1	500.4	501.2	502.4	502.7	502.5	502.4	502.0	502.6	503.9
LVA	6.4	1.7	22	**22–27**	496.8	1.7	**5.1**	499.7	498.6	497.2	497.7	494.6	495.3	495.1	497.9	497.9	494.7	496.5
NOR	**10.7**	**3.7**	23	**18–23**	501.8	1.5	4.2	499.4	499.8	502.0	502.0	503.7	503.4	503.4	500.6	500.3	502.7	502.0
AUS	7.3	2.4	24	24–26	495.7	0.9	3.7	495.3	496.3	497.8	495.6	496.0	495.5	495.5	495.6	495.7	495.6	494.0
SWE	**12.7**	**5.9**	25	**21–25**	498.4	3.3	**10.1**	491.8	493.4	499.1	499.7	501.3	501.1	501.3	497.9	498.0	501.9	496.8
ITA	**12.4**	2.8	26	25–27	492.0	2.3	**5.4**	490.4	490.1	489.4	490.1	494.7	494.4	494.0	490.1	489.9	494.6	494.1
ISL	9.7	**4.4**	27	**24–27**	494.2	2.8	**8.9**	489.1	491.3	496.5	498.0	495.1	495.6	495.8	495.9	495.2	493.3	490.5
LUX	**10.4**	2.8	28	28–28	486.5	0.9	2.6	486.8	486.5	485.5	486.3	487.2	487.4	486.6	485.3	485.2	487.7	487.5
LTU	7.4	1.5	29	**29–34**	482.0	1.7	**5.5**	485.5	484.4	482.1	483.0	480.1	480.9	480.6	482.0	481.9	480.6	480.9
RUS	7.9	2.2	30	29–31	483.7	0.7	2.1	484.6	484.2	483.6	484.6	482.9	483.7	483.9	484.0	483.7	482.5	482.5
SVK	8.0	1.8	31	**29–32**	483.2	0.6	2.4	484.5	483.9	482.8	483.4	482.8	483.3	483.0	483.2	483.1	482.1	482.7
HUN	8.6	2.0	32	**29–32**	483.2	0.7	2.4	484.1	483.8	483.7	483.4	482.9	482.7	482.9	484.0	483.7	482.6	481.6
ESP	**10.6**	2.9	33	**32–35**	481.5	1.7	**5.8**	482.4	482.4	481.7	482.5	481.6	482.1	482.3	482.4	481.9	480.4	476.7
USA	**4.0**	2.0	34	**29–36**	482.2	2.4	**6.6**	481.6	483.1	484.9	485.7	479.5	480.4	480.5	484.5	484.5	479.1	480.7
BLR	7.8	2.4	35	**32–36**	480.3	1.7	**5.4**	477.7	477.2	481.4	482.6	479.5	480.3	480.5	481.6	481.6	480.1	481.1
MLT	9.8	**3.9**	36	**33–37**	476.6	2.8	**9.2**	474.2	476.0	479.8	471.2	480.3	475.0	476.8	475.6	476.6	480.5	476.5
HRV	**11.8**	2.7	37	36–37	470.8	1.8	5.0	471.8	469.0	468.3	471.7	472.9	471.1	473.3	468.8	468.5	471.1	472.3
TUR	6.7	1.6	38	38–39	460.3	2.0	**6.2**	464.0	462.9	460.8	462.2	458.0	459.3	459.4	460.0	460.0	457.8	458.7
ISR	**12.1**	**4.5**	39	38–39	461.7	1.6	4.3	459.9	461.4	462.4	461.4	463.5	462.6	462.3	459.3	459.2	463.3	463.0
GRC	**10.7**	2.7	40	40–41	439.1	2.5	**9.2**	440.9	440.1	440.0	441.3	438.2	439.8	438.8	439.3	439.2	440.0	432.2
MYS	**1.2**	0.6	41	**41–44**	429.1	4.8	**12.5**	435.8	433.4	432.2	433.5	423.3	424.6	425.4	431.5	432.4	424.4	423.6
ALB	8.0	1.9	42	**41–44**	429.6	2.2	**6.3**	432.0	431.1	430.5	426.5	427.8	427.8	427.8	432.2	432.8	428.5	428.2
BRN	6.1	1.7	43	42–44	427.7	3.2	**8.1**	430.9	430.0	428.8	430.0	423.2	424.4	424.2	428.8	429.4	423.7	431.3
MNE	**17.3**	**3.8**	44	**40–44**	433.1	3.5	**10.7**	430.5	431.3	429.8	428.1	436.8	436.3	436.1	431.1	430.4	438.8	434.8
BIH	**18.8**	**3.9**	45	45–45	413.9	3.8	**10.5**	410.7	410.4	410.3	409.8	417.3	417.3	417.1	412.2	411.2	420.3	416.6

Note. %NA = proportion of item responses with missing data; %NR = proportion of item responses that are not reached; rk_{UW} = country rank from model UW; rk_{Int} = interval of country ranks obtained from 11 different scaling models; Aver = average of country means across 11 models; SD = standard deviation of country means across 11 models; rg = ange of country means across 11 models; UW = scoring as wrong (Sect. 3.1.1) ; UP = MC items scored as partially correct (Sect. 3.1.2); UN1 = ignoring not reached items (Sect. 3.1.3); UN2 = including proportion of not reached items in background model (Sect. 3.1.3); UO1 = ignoring missing items (Sect. 3.1.4); MO2 = model-based latent ignorability (Sect. 3.1.5, Eq. (3)); IO2 = imputed under latent ignorability (Sect. 3.1.5, Eq. (3)); MM2 = Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); IM2 = imputed under Mislevy-Wu model with item format specific δ parameters (Sect. 3.1.6, Eq. (4)); IF1 = FCS imputation based on item responses (Sect. 3.1.7); IF2 = FCS imputation based on item responses and response indicators (Sect. 3.1.7); The following entries in the table are printed in bold: Missing proportions (%NA) larger than 10.0% and smaller than 5.0%, not reached proportions larger than 3.0%, country rank differences larger than 2, ranges in country means larger than 5.0. See Appendix A for country labels.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic model
2PL	two-parameter logistic model
BIC	Bayesian information criterion
CR	constructed-response
FCS	fully conditional specification
GHP	Gilula-Haberman penalty
IRT	item response theory
LSA	large-scale assessment
MC	multiple-choice
PISA	programme for international student assessment
PMM	predictive mean matching

Appendix A. Country Labels

The country labels used in Tables 1, 3, 4 and 5 are as follows: ALB = Albania; AUS = Australia; AUT = Austria; BEL = Belgium; BIH = Bosnia and Herzegovina; BLR = Belarus; BRN = Brunei Darussalam; CAN = Canada; CHE = Switzerland; CZE = Czech Republic; DEU = Germany; DNK = Denmark; ESP = Spain; EST = Estonia; FIN = Finland; FRA = France; GBR = United Kingdom; GRC = Greece; HKG = Hong Kong; HRV = Croatia; HUN = Hungary; IRL = Ireland; ISL = Iceland; ISR = Israel; ITA = Italy; JPN = Japan; KOR = Korea; LTU = Lithuania; LUX = Luxembourg; LVA = Latvia; MLT = Malta; MNE = Montenegro; MYS = Malaysia; NLD = Netherlands; NOR = Norway; NZL = New Zealand; POL = Poland; PRT = Portugal; RUS = Russian Federation; SGP = Singapore; SVK = Slovak Republic; SVN = Slovenia; SWE = Sweden; TUR = Turkey; USA = United States.

References

1. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J., Eds. *Implementation of large-scale education assessments*; Wiley: New York, 2017. doi:10.1002/9781118762462.
2. Rutkowski, L.; von Davier, M.; Rutkowski, D., Eds. *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*; Chapman Hall/CRC Press: London, 2013. doi:10.1201/b16061.
3. Pohl, S.; Ulitzsch, E.; von Davier, M. Reframing rankings in educational assessments. *Science* **2021**, *372*, 338–340. doi:10.1126/science.abd3300.
4. Graham, J.W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **2009**, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530.
5. Little, R.J.A.; Rubin, D.B. *Statistical analysis with missing data*; John Wiley & Sons: New York, 2002. doi:10.1002/9781119013563.
6. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational measurement*; Brennan, R.L., Ed.; Praeger Publishers, 2006; pp. 111–154. https://bit.ly/3nldpF5.
7. Rose, N.; von Davier, M.; Xu, X. *Modeling nonignorable missing data with item response theory (IRT).* (Research Report No. RR-10-11). Educational Testing Service, 2010. doi:10.1002/j.2333-8504.2010.tb02218.x.
8. Pohl, S.; Gräfe, L.; Rose, N. Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educ. Psychol. Meas.* **2014**, *74*, 423–452. doi:10.1177/0013164413504926.
9. Rose, N.; von Davier, M.; Nagengast, B. Modeling omitted and not-reached items in IRT models. *Psychometrika* **2017**, *82*, 795–819. doi:10.1007/s11336-016-9544-7.
10. OECD. *PISA 2015. Technical report*; OECD: Paris, 2017. https://bit.ly/32buWnZ.
11. Pohl, S.; Carstensen, C. *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14), 2012. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. https://bit.ly/2XThQww.
12. Pohl, S.; Carstensen, C.H. Scaling of competence tests in the national educational panel study – Many questions, some answers, and further challenges. *J. Educ. Res. Online* **2013**, *5*, 189–216. https://bit.ly/39AETyE.
13. Kuha, J.; Katsikatsou, M.; Moustaki, I. Latent variable modelling with non-ignorable item nonresponse: Multigroup response propensity models for cross-national analysis. *J. R. Stat. Soc. Series A Stat.* **2018**, *181*, 1169–1192. doi:10.1111/rssa.12350.
14. Holman, R.; Glas, C.A.W. Modelling non-ignorable missing-data mechanisms with item response theory models. *Brit. J. Math. Stat. Psychol.* **2005**, *58*, 1–17. doi:10.1111/j.2044-8317.2005.tb00312.x.
15. Knott, M.; Tzamourani, P. Fitting a latent trait model for missing observations to racial prejudice data. In *Applications of latent trait and latent class models in the social sciences*; Rost, J.; Langeheine, R., Eds.; Waxmann, 1997; pp. 244–252. https://bit.ly/3CMEJ3K.

16. Finch, H. Estimation of item response theory parameters in the presence of missing data. *J. Educ. Meas.* **2008**, *45*, 225–245. doi:10.1111/j.1745-3984.2008.00062.x.
17. Sinharay, S. Score reporting for examinees with incomplete data on large-scale educational assessments. *Educ. Meas.* **2021**, *40*, 79–91. doi:10.1111/emip.12396.
18. Robitzsch, A. Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment [On nonignorable consequences of (partial) ignoring of missing item responses in large-scale assessments]. In *PIRLS & TIMSS 2011. Die Kompetenzen in Lesen, Mathematik und Naturwissenschaften am Ende der Volksschule. Österreichischer Expertenbericht*; Suchan, B.; Wallner-Paschon, C.; Schreiner, C., Eds.; Leykam: Graz, 2016; pp. 55–64. https://bit.ly/2ZnEYDP.
19. Robitzsch, A. About still nonignorable consequences of (partially) ignoring missing item responses in large-scale assessment. *OSF Preprints* **2020**, 20 October 2020. doi:10.31219/osf.io/hmy45.
20. Robitzsch, A.; Lüdtke, O. Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. *PsyArXiv* **2021**, 31 August 2021. doi:10.31234/osf.io/pkjth.
21. Rohwer, G. *Making sense of missing answers in competence tests* (NEPS Working Paper No. 30), 2013. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. https://bit.ly/3AGfsr5.
22. Mislevy, R.J.; Wu, P.K. *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30). Educational Testing Service, 1996. doi:10.1002/j.2333-8504.1996.tb01708.x.
23. Robitzsch, A.; Lüdtke, O. *An item response model for omitted responses in performance tests.* Talk held at IMPS 2017, Zurich, July 2017., 2017. https://bit.ly/3u8rgjy.
24. OECD. *PISA 2018. Technical report*; OECD: Paris, 2020. https://bit.ly/3zWbidA.
25. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*; Lord, F.M.; Novick, M.R., Eds.; MIT Press: Reading, MA, 1968; pp. 397–479.
26. Rasch, G. *Probabilistic models for some intelligence and attainment tests*; Danish Institute for Educational Research: Copenhagen, 1960. https://bit.ly/3sVMlgy.
27. Rose, N.; von Davier, M.; Nagengast, B. Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psych. Test Assess. Model.* **2015**, *57*, 472–498. https://bit.ly/3kD3t89.
28. Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika* **1974**, *39*, 247–264. doi:10.1007/BF02291471.
29. R Core Team. *R: A language and environment for statistical computing*, 2020. Accessed on 20 August 2020. Vienna, Austria. https://www.R-project.org/.
30. Robitzsch, A. *sirt: Supplementary item response theory models*, 2021. R package version 3.10-118. Accessed on 23 September 2021. https://CRAN.R-project.org/package=sirt.
31. Beesley, L.J.; Taylor, J.M.G. A stacked approach for chained equations multiple imputation incorporating the substantive model. *Biometrics* **2020**. [Epub ahead of print], doi:10.1111/biom.13372.
32. van Buuren, S. *Flexible imputation of missing data*; CRC Press: Boca Raton, 2018. doi:10.1201/9780429492259.
33. Chan, K.W.; Meng, X.L. Multiple improvements of multiple imputation likelihood ratio tests. *arXiv Preprint* **2017**, arXiv:1711.08822. https://arxiv.org/abs/1711.08822.
34. von Davier, M.; Sinharay, S. Analytics in international large-scale assessments: Item response theory and population models. In *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*; Rutkowski, L.; von Davier, M.; Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, 2013; pp. 155–174. doi:10.1201/b16061-12.
35. Harel, O.; Schafer, J.L. Partial and latent ignorability in missing-data problems. *Biometrika* **2009**, *96*, 37–50. doi:10.1093/biomet/asn069.
36. Beesley, L.J.; Taylor, J.M.G.; Little, R.J.A. Sequential imputation for models with latent variables assuming latent ignorability. *Aust. N. Z. J. Stat.* **2019**, *61*, 213–233. doi:10.1111/anzs.12264.
37. Debeer, D.; Janssen, R.; De Boeck, P. Modeling skipped and not-reached items using IRTrees. *J. Educ. Meas.* **2017**, *54*, 333–363. doi:10.1111/jedm.12147.
38. Glas, C.A.W.; Pimentel, J.L.; Lamers, S.M.A. Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psych. Test Assess. Model.* **2015**, *57*, 523–541. https://bit.ly/3EOcX8M.
39. Jung, H.; Schafer, J.L.; Seo, B. A latent class selection model for nonignorably missing data. *Comp. Stat. Data An.* **2011**, *55*, 802–812. doi:10.1016/j.csda.2010.07.002.
40. Bacci, S.; Bartolucci, F. A multidimensional finite mixture structural equation model for nonignorable missing responses to test items. *Struct. Equ. Modeling* **2015**, *22*, 352–365. doi:10.1080/10705511.2014.937376.
41. Bartolucci, F.; Montanari, G.E.; Pandolfi, S. Latent ignorability and item selection for nursing home case-mix evaluation. *J. Classif.* **2018**, *35*, 172–193. doi:10.1007/s00357-017-9227-9.
42. Fu, Z.H.; Tao, J.; Shi, N.Z. Bayesian estimation of the multidimensional graded response model with nonignorable missing data. *J. Stat. Comput. Simul.* **2010**, *80*, 1237–1252. doi:10.1080/00949650903029276.
43. Okumura, T. Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educ. Psychol. Meas.* **2014**, *74*, 611–626. doi:10.1177/0013164413516976.
44. Bertoli-Barsotti, L.; Punzo, A. Rasch analysis for binary data with nonignorable nonresponses. *Psicologica* **2013**, *34*, 97–123.
45. Pohl, S.; Becker, B. Performance of missing data approaches under nonignorable missing data conditions. *Methodology* **2020**, *16*, 147–165. doi:10.5964/meth.2805.

46. Rosas, G.; Shomer, Y. Models of nonresponse in legislative politics. *Legis. Stud. Q.* **2008**, *33*, 573–601. doi:10.3162/036298008786403088.
47. Köhler, C.; Pohl, S.; Carstensen, C.H. Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educ. Psychol. Meas.* **2015**, *75*, 850–874. doi:10.1177/0013164414561785.
48. Hughes, R.A.; White, I.R.; Seaman, S.R.; Carpenter, J.R.; Tilling, K.; Sterne, J.A.C. Joint modelling rationale for chained equations. *BMC Med. Res. Methodol.* **2014**, *14*, 28. doi:10.1186/1471-2288-14-28.
49. Yuan, K.H. Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *J. Multivar. Anal.* **2009**, *100*, 1900–1918. doi:10.1016/j.jmva.2009.05.001.
50. Finch, H.W. A comparison of the Heckman selection model, Ibrahim, and Lipsitz methods for dealing with nonignorable missing data. *J. Psychiatry Behav. Sci.* **2021**, *4*, 1045. https://bit.ly/3ERVhJd.
51. Galimard, J.E.; Chevret, S.; Protopopescu, C.; Resche-Rigon, M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat. Med.* **2016**, *35*, 2907–2920. doi:10.1002/sim.6902.
52. Galimard, J.E.; Chevret, S.; Curis, E.; Resche-Rigon, M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Med. Res. Methodol.* **2018**, *18*, 90. doi:10.1186/s12874-018-0547-1.
53. Heckman, J. Sample selection bias as a specification error. *Econometrica* **1979**, *47*, 153–161. doi:10.2307/1912352.
54. Sportisse, A.; Boyer, C.; Josse, J. Imputation and low-rank estimation with missing not at random data. *Stat. Comput.* **2020**, *30*, 1629–1643. doi:10.1007/s11222-020-09963-5.
55. Mislevy, R.J. Missing responses in item response modeling. In *Handbook of item response theory, Vol. 2: Statistical tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, 2016; pp. 171–194. doi:10.1201/b19166-10.
56. Deribo, T.; Kroehne, U.; Goldhammer, F. Model-based treatment of rapid guessing. *J. Educ. Meas.* **2021**, *58*, 281–303. doi:10.1111/jedm.12290.
57. Guo, J.; Xu, X. An IRT-based model for omitted and not-reached items. *arXiv Preprint* **2019**, arXiv:1904.03767. https://arxiv.org/abs/1904.03
58. Rosas, G.; Shomer, Y.; Haptonstahl, S.R. No news is news: Nonignorable nonresponse in roll-call data analysis. *Am. J. Pol. Sc.* **2015**, *59*, 511–528. doi:10.1111/ajps.12148.
59. Gomes, H.; Matsushita, R.; Da Silva, S. Item tesponse theory modeling of high school students' behavior in a high-stakes exam. *Open Access Libr. J.* **2019**, *6*, e5242. doi:10.4236/oalib.1105242.
60. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. doi:10.18637/jss.v045.i03.
61. Bulut, O.; Quo, Q.; Gierl, M.J. A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assess. Educ.* **2017**, *5*, 8. doi:10.1186/s40536-017-0042-x.
62. Debeer, D.; Janssen, R. Modeling item-position effects within an IRT framework. *J. Educ. Meas.* **2013**, *50*, 164–185. doi:10.1111/jedm.12009.
63. Hartig, J.; Buchholz, J. A multilevel item response model for item position effects and individual persistence. *Psych. Test Assess. Model.* **2012**, *54*, 418–431. https://bit.ly/39Y4WQx.
64. Nagy, G.; Nagengast, B.; Becker, M.; Rose, N.; Frey, A. Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psych. Test Assess. Model.* **2018**, *60*, 165–187. https://bit.ly/3Biw74g.
65. Robitzsch, A. Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in calibrating performance tests]. In *Bildungsstandards Deutsch und Mathematik*; Bremerich-Vos, A.; Granzer, D.; Köller, O., Eds.; Beltz Pädagogik: Weinheim, 2009; pp. 42–106.
66. Rose, N.; Nagy, G.; Nagengast, B.; Frey, A.; Becker, M. Modeling multiple item context effects with generalized linear mixed models. *Front. Psychol.* **2019**, *10*, 248. doi:10.3389/fpsyg.2019.00248.
67. Trendtel, M.; Robitzsch, A. Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psych. Test Assess. Model.* **2018**, *60*, 241–263. https://bit.ly/3l4Zi5u.
68. Weirich, S.; Hecht, M.; Böhme, K. Modeling item position effects using generalized linear mixed models. *Appl. Psychol. Meas.* **2014**, *38*, 535–548. doi:10.1177/0146621614534955.
69. Grund, S.; Lüdtke, O.; Robitzsch, A. On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *J. Educ. Behav. Stat.* **2021**, *46*, 430–465. doi:10.3102/1076998620959058.
70. Robitzsch, A.; Pham, G.; Yanagida, T. Fehlende Daten und Plausible Values [Missing data and plausible values]. In *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung*; Breit, S.; Schreiner, C., Eds.; facultas: Vienna, 2016; pp. 259–293. https://bit.ly/2YaZQ0G.
71. Beesley, L.J.; Bondarenko, I.; Elliott, M.R.; Kurian, A.W.; Katz, S.J.; Taylor, J.M.G. Multiple imputation with missing data indicators. *arXiv Preprint* **2021**, arXiv:2103.02033. https://arxiv.org/abs/2103.02033.
72. Kolen, M.J.; Brennan, R.L. *Test equating, scaling, and linking*; Springer: New York, 2014. doi:10.1007/978-1-4939-0317-7.
73. Lee, W.C.; Lee, G. IRT linking and equating. In *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test*; Irwing, P.; Booth, T.; Hughes, D.J., Eds.; Wiley: New York, 2018; pp. 639–673. doi:10.1002/9781118489772.ch21.
74. Haberman, S.J. *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report No. RR-09-40). Educational Testing Service, 2009. doi:10.1002/j.2333-8504.2009.tb02197.x.
75. Robitzsch, A. L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats* **2020**, *3*, 246–283. doi:10.3390/stats3030019.

76. Robitzsch, A.; Lüdtke, O. Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *J. Educ. Behav. Stat.* **2021**. [Epub ahead of print], doi:10.3102/10769986211017479.
77. Joo, S.H.; Khorramdel, L.; Yamamoto, K.; Shin, H.J.; Robin, F. Evaluating item fit statistic thresholds in PISA: Analysis of cross-country comparability of cognitive items. *Educ. Meas.* **2021**, *40*, 37–48. doi:10.1111/emip.12404.
78. Oliveri, M.E.; von Davier, M. Investigation of model fit and score scale comparability in international assessments. *Psych. Test Assess. Model.* **2011**, *53*, 315–333. https://bit.ly/3k4K9kt.
79. von Davier, M.; Yamamoto, K.; Shin, H.J.; Chen, H.; Khorramdel, L.; Weeks, J.; Davis, S.; Kong, N.; Kandathil, M. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* **2019**, *26*, 466–488. doi:10.1080/0969594X.2019.1586642.
80. Lumley, T.; Scott, A. Tests for regression models fitted to survey data. *Aust. N. Z. J. Stat.* **2014**, *56*, 1–14. doi:10.1111/anzs.12065.
81. Lumley, T.; Scott, A. AIC and BIC for modeling with complex survey data. *J. Surv. Stat. Methodol.* **2015**, *3*, 1–18. doi:10.1093/jssam/smu021.
82. Trendtel, M.; Robitzsch, A. A Bayesian item response model for examining item position effects in complex survey data. *J. Educ. Behav. Stat.* **2021**, *46*, 34–57. doi:10.3102/1076998620931016.
83. Haberman, S.J. *The information a test provides on an ability parameter* (Research Report No. RR-07-18). Educational Testing Service, 2007. doi:10.1002/j.2333-8504.2007.tb02060.x.
84. van Rijn, P.W.; Sinharay, S.; Haberman, S.J.; Johnson, M.S. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assess. Educ.* **2016**, *4*, 10. doi:10.1186/s40536-016-0025-3.
85. George, A.C.; Robitzsch, A. Validating theoretical assumptions about reading with cognitive diagnosis models. *Int. J. Test.* **2021**, *21*, 105–129. doi:10.1080/15305058.2021.1931238.
86. Sachse, K.A.; Mahler, N.; Pohl, S. When nonresponse mechanisms change: Effects on trends and group comparisons in international large-scale assessments. *Educ. Psychol. Meas.* **2019**, *79*, 699–726. doi:10.1177/0013164419829196.
87. Lu, J.; Wang, C. A response time process model for not-reached and omitted items. *J. Educ. Meas.* **2020**, *57*, 584–620. doi:10.1111/jedm.12270.
88. Ulitzsch, E.; von Davier, M.; Pohl, S. Using response times for joint modeling of response and omission behavior. *Multivar. Behav. Res.* **2020**, *55*, 425–453. doi:10.1080/00273171.2019.1643699.
89. Goldhammer, F.; Martens, T.; Lüdtke, O. Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assess. Educ.* **2017**, *5*, 18. doi:10.1186/s40536-017-0051-9.
90. Pokropek, A. Grade of membership response time model for detecting guessing behaviors. *J. Educ. Behav. Stat.* **2016**, *41*, 300–325. doi:10.3102/1076998616636618.
91. Ulitzsch, E.; von Davier, M.; Pohl, S. A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *Brit. J. Math. Stat. Psychol.* **2020**, *73*, 83–112. doi:10.1111/bmsp.12188.
92. Weeks, J.P.; von Davier, M.; Yamamoto, K. Using response time data to inform the coding of omitted responses. *Psych. Test Assess. Model.* **2016**, *58*, 671–701. https://bit.ly/3AG33U7.
93. Adams, R.J.; Lietz, P.; Berezner, A. On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assess. Educ.* **2013**, *1*, 5. doi:10.1186/2196-0739-1-5.
94. Aßmann, C.; Gaasch, C.; Pohl, S.; Carstensen, C.H. Bayesian estimation in IRT models with missing values in background variables. *Psych. Test Assess. Model.* **2015**, *57*, 595–618. https://bit.ly/2ZNfzno.
95. Kaplan, D.; Su, D. On imputation for planned missing data in context questionnaires using plausible values: a comparison of three designs. *Large-scale Assess. Educ.* **2018**, *6*, 6. doi:10.1186/s40536-018-0059-9.
96. Rutkowski, L. The impact of missing background data on subpopulation estimation. *J. Educ. Meas.* **2011**, *48*, 293–312. doi:10.1111/j.1745-3984.2011.00144.x.
97. von Davier, M. Imputing proficiency data under planned missingness in population models. In *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*; Rutkowski, L.; von Davier, M.; Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, 2013; pp. 175–201. doi:10.1201/b16061-13.
98. Mislevy, R.J. Randomization-based inference about latent variables from complex samples. *Psychometrika* **1991**, *56*, 177–196. doi:10.1007/BF02294457.
99. Athey, S.; Imbens, G. A measure of robustness to misspecification. *Am. Econ. Rev.* **2015**, *105*, 476–80. doi:10.1257/aer.p20151020.
100. Buckland, S.T.; Burnham, K.P.; Augustin, N.H. Model selection: An integral part of inference. *Biometrics* **1997**, *53*, 603–618. doi:10.2307/2533961.
101. Longford, N.T. 'Which model?' is the wrong question. *Stat. Neerl.* **2012**, *66*, 237–252. doi:10.1111/j.1467-9574.2011.00517.x.
102. Siddique, J.; Harel, O.; Crespi, C.M. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *Ann. Appl. Stat.* **2012**, *6*, 1814–1837. doi:10.1214/12-AOAS555.
103. Young, C. Model uncertainty in sociological research: An application to religion and economic growth. *Am. Sociol. Rev.* **2009**, *74*, 380–397. doi:10.1177/000312240907400303.
104. Young, C.; Holsteen, K. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociol. Methods Res.* **2017**, *46*, 3–40. doi:10.1177/0049124115610347.
105. Robitzsch, A.; Dörfler, T.; Pfost, M.; Artelt, C. Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen [Relevance of item selection and model selection for assessing the development of competencies:

The development in reading competence in primary school students]. *Z. Entwicklungspsychol. Pädagog. Psychol.* **2011**, *43*, 213–227. doi:10.1026/0049-8637/a000052.
106. Saltelli, A.; Ratto, M.; Andres, T.; Campolongo, F.; Cariboni, J.; Gatelli, D.; Saisana, M.; Tarantola, S. *Global sensitivity analysis: the primer*; John Wiley & Sons: New York, 2008. doi:10.1002/9780470725184.
107. Harder, J.A. The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspect. Psychol. Sci.* **2020**, *15*, 1158–1177. doi:10.1177/1745691620917678.
108. Steegen, S.; Tuerlinckx, F.; Gelman, A.; Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **2016**, *11*, 702–712. doi:10.1177/1745691616658637.
109. Simonsohn, U.; Simmons, J.P.; Nelson, L.D. Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN* **2015**, 25 November 2015. doi:10.2139/ssrn.2694998.
110. Simonsohn, U.; Simmons, J.P.; Nelson, L.D. Specification curve analysis. *Nat. Hum. Behav.* **2020**, *4*, 1208–1214. doi:10.1038/s41562-020-0912-z.