METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Performance of Missing Data Approaches Under Nonignorable Missing Data Conditions

Steffi Pohl[a], Benjamin Becker[b]

[a] *Freie Universität Berlin, Berlin, Germany.* **[b]** *Institute for Educational Quality Improvement, Berlin, Germany.*

Methodology, 2020, Vol. 16(2), 147–165, https://doi.org/10.5964/meth.2805

Received: 2018-04-04 • **Accepted:** 2019-12-02 • **Published (VoR):** 2020-06-18

Corresponding Author: Steffi Pohl, Freie Universität Berlin, AB Methoden und Evaluation/Qualitätssicherung, Habelschwerdter Allee 45, 14195 Berlin, Germany. Phone: +49-(0)30-838-62926, E-mail: steffi.pohl@fu-berlin.de

Abstract

Approaches for dealing with item omission include incorrect scoring, ignoring missing values, and approaches for nonignorable missing values and have only been evaluated for certain forms of nonignorability. In this paper we investigate the performance of these approaches for various conditions of nonignorability, that is, when the missing response depends on i) the item response, ii) a latent missing propensity, or iii) both. No approach results in unbiased parameter estimates of the Rasch model under all missing data mechanisms. Incorrect scoring only results in unbiased estimates under very specific data constellations of missing mechanisms i) and iii). The approach for nonignorable missing values only results in unbiased estimates under condition ii). Ignoring results in slightly more biased estimates than the approach for nonignorable missing values, while the latter also indicates the presence of nonignorablity under all simulated conditions. We illustrate the results in an empirical example on PISA data.

Keywords

missing data, nonignorability, item response theory, item nonresponse, large-scale assessment

Test data of low stakes large-scale assessments usually contain a substantial proportion of missing responses on test items. In this paper, we focus on missing values due to item omission. Omitted items are usually nonignorable (see, e.g., Lord, 1974; Mislevy & Wu, 1996). Ignorability for likelihood inference is given when responses are missing at random and the parameters of the data generating model and the missing-data process are distinct (Rubin, 1976). When missing values are nonignorable and not appropriately accounted for parameter estimates can be biased. This can, for example, lead to different country rankings in competence test scores (Rose, von Davier, & Xu, 2010) or biased regression coefficients when predicting competence test data from explaining variables (Köhler, Pohl, & Carstensen, 2017). The goal of this study is to evaluate existing

approaches for dealing with missing values on a wider set of plausible missing data mechanisms than has been done before and to test the robustness of the approaches as well as their limitations.

There are different approaches to dealing with missing values (for an overview see, e.g., De Ayala, Plake, & Impara, 2001; Finch, 2008; or Rose et al., 2010)[1]:

1. Incorrect scoring: Missing responses may be scored as incorrect responses, assuming that the subject did not know the answer. There are different views on the properties of this approach. Assuming that there is a true unobserved response (e.g., Holman & Glas, 2005; Rose et al., 2010), the approach of incorrect scoring results in unbiased estimates only if missing values solely occur on incorrect responses. Researchers furthermore delineated that incorrect scoring does violate assumptions of IRT models. Lord (1974) argued that incorrect scoring is a deterministic scoring approach ignoring the fact that the subject has a positive probability of solving the item, given the individual's trait level. Rose (2013) delineated that incorrect scoring results in local stochastic dependence and measurement non-invariance, which is a violation of the model assumptions.
2. Ignoring: Another approach is ignoring missing values and, thus, treating them as if they were not administered. This approach assumes that missing responses are MAR, given the observed responses on the items in the test (and other covariates in the background model).
3. Nonignoring: Based on work of Moustaki and Knott (2000) and O'Muircheartaigh and Moustaki (1999), Holman and Glas (2005) proposed an approach for dealing with nonignorable missing data in IRT models. In this approach, the tendency to omit items is included as a latent variable in the model and accounted for in the estimation of the item and person parameters. The authors assume a unidimensional IRT measurement model for the responses Y_i of the participants on item *i*. Missing responses due to omitted items are treated as missing values. The latent ability to be estimated is denoted by θ. The authors then define manifest missing indicators D_i as $D_i = 0$, if Y_i is observed and as $D_i = 1$, if Y_i is omitted. A latent variable ξ, which may be interpreted as the propensity of a person to omit items, is then modeled based on these manifest missing indicators. Thus, there is a separate measurement model for observed item responses and for missing responses. An implicit assumption of the model is that the missing indicators fit a unidimensional measurement model.

Studies investigating the performance of the different missing data approaches show that, when the assumptions of the models are met, the respective approaches recover the true parameters. The approach of incorrect scoring results in unbiased parameter

1) While multiple imputation is a sophisticated and often used approach for missing values in single indicator variables (e.g., Vidotto, Vermunt, & Van Deun, 2018; Kleinke, 2018), this is seldom used for dealing with item nonresponse in psychometric models.

estimates when only incorrect responses are missing (Rohwer, 2013). When the missing mechanism is MCAR, MAR, or nonignorable as modeled by the approach for nonignorable missing responses, incorrect scoring results in an underestimation of ability for persons with missing values (e.g., Culbertson, 2011; De Ayala, Plake, & Impara, 2001; Finch, 2008; Lord, 1974; Rose et al., 2010). Ignoring missing values performs well in cases where the missing mechanism is MCAR or MAR (Culbertson, 2011; Finch, 2008; Köhler et al., 2017; Rose et al., 2010)[2]. Ignoring results in biased parameter estimates when the missing mechanism is nonignorable (Culbertson, 2011; De Ayala, Plake, & Impara, 2001; Pohl, Gräfe, & Rose, 2014; Rose et al., 2010). If the nonignorable missing mechanism follows the model for nonignorable missing values, the bias of ignoring missing values becomes negligible when the relationship between ability and missing propensity is rather small (Holman & Glas, 2005; Köhler et al., 2017; Pohl et al., 2014). The approach for nonignorable missing values performs well when the missing mechanism is MCAR, MAR, or when the missing mechanism is generated according to the model for nonignorable missing values (Holman & Glas, 2005; Rose, 2013; Rose et al., 2010).

While the different approaches have been evaluated extensively under MCAR or MAR assumptions, they have only been evaluated for very specific forms of nonignorability (Robitzsch, 2016). Interpretations of study results using approaches for nonignorable missing values, however, seem to neglect that these approaches cannot account for all kinds of nonignorable missing values. Thus, users may overestimate the performance of these approaches. As nonignorability is very plausible for omitted items, information on the performance of the different approaches on various types of nonignorable missing data mechanisms is essential for choosing an approach. In this paper, we distinguish different nonignorable missing data mechanisms:

$$P(D_i) = f(Y_i) \tag{1}$$

$$P(D_i) = f(\xi) \tag{2}$$

$$P(D_i) = f(Y_i, \xi) \tag{3}$$

with D_i denoting the missing indicator of item i, Y_i denoting the item response, and ξ denoting the missing propensity as defined in the model of Holman and Glas (2005). Of course, it is also possible to consider even more complex forms of these missing data mechanisms. The main idea of the nonignorable mechanisms can, however, be characterized by these three basic mechanisms.

Robitzsch (2016) argues that it is necessary to investigate the performance of the approaches under mechanisms in which the missing response is generated as a function

2) Note that some of the studies generated missing values based on (categorized) sum scores of observed items (e.g., De Ayala, et al., 2001; Finch, 2008). Ignoring missing values can then only result in unbiased estimates if the variables that determine the missing values are included in the model.

of the true response itself (Equation 1). This has been done to some extent in previous studies for approaches that assume MAR (e.g., ignoring) or incorrect scoring. Finch (2008), for example, generated the probability for a missing response to an item as a function of the response on that item in the complete data. Culbertson (2011) applied a similar approach, however, only using the special case where only incorrect responses were missing. In this simulation, he only evaluated the approaches of ignoring missing values and scoring missing values as fractional correct. De Ayala et al. (2001) simulated the probability of a missing response on an item as a function of the response to that item in the complete data and on a categorized sum score on other, fully observed item responses. As such, they generated missing responses as a function of the item itself (according to Equation 1) and additionally of a rough proxy of ability. None of the studies that used a missing mechanism as defined in Equation 1 systematically evaluated the performance of all three approaches for this kind of missing mechanism. Ignoring and incorrect scoring have been evaluated on only a limited set of conditions referring to this mechanism. The approach for nonignorable missing values has not at all been evaluated on this missing mechanism. Those studies evaluating the performance of the approach for nonignorable missing data exclusively used the mechanism described in Equation 2 (e.g., Glas et al., 2015; Holman & Glas, 2005; Köhler et al., 2017; Pohl et al., 2014; Rose et al., 2010).

The mechanism in Equation 3 is a general case which includes the mechanisms of Equation 1 and 2 as special cases. Here, the probability of a missing value depends on both, the missing propensity and the response itself. Thus, the general omission tendency of a person as well as the item response determine the probability to omit an item. This missing mechanism has been discussed and modeled by Mislevy and Wu (1996) as well as Robitzsch (2016). So far, it has not been evaluated how the different approaches perform under this more complex missing mechanism.

In this paper, we want to challenge the three missing data approaches (incorrect scoring, ignoring, nonignoring) to evaluate under which kind of ignorable and nonignorable missing data mechanisms they perform well and where their limitations are. We performed three simulation studies, each using one of the three nonignorable missing data mechanisms. Note that the performance of the three approaches for the missing mechanism in Equation 2 has already been extensively investigated in previous studies. We include it in our study in order to investigate whether it yields differential results for data generated according to the missing mechanism in Equations 1 and 2, which would allow us to distinguish between the two mechanisms based on analyses results. In the following the design and the results of each simulation are described.

Simulation Study 1: Missing Values Depending on Item Responses

Data Generation and Data Analyses

In line with typical LSAs, we generated data for $N = 3000$ persons on $k = 32$ binary test items using a Rasch model (Rasch, 1960). The person parameters θ_p were assumed to follow a normal distribution with $\theta \sim N(0,1)$. Item parameters β_i were generated to be approximately normally distributed and symmetric with a mean of zero and a variance of 0.9. The item parameters were fixed across all replications. Response probabilities were then generated based on the Rasch model and item responses were derived by comparing the response probabilities to a randomly drawn set of values from a uniform distribution with a range from zero to one.

We varied the proportion of omitted items as well as the dependency of missing values on the item responses. For omission rates, we chose a proportion of 10% of all item responses to be missing, as this depicts typical applications, as well as a proportion of 30% in order to investigate the effect of proportion of missing values on parameter estimates[3]. For varying the dependency of missing responses on item responses, and as such, the amount of nonignorability, we imposed different conditional probabilities of a missing value on an item given the item response (Table 1)[4]. The larger the differences in the conditional probabilities between incorrect and correct responses, the more severe the mechanism is missing not at random (MNAR) and, thus, the larger the amount of nonignorability. Note that in the strong MNAR condition only incorrectly solved items are missing. The MCAR approach conforms to the assumption of the approach for nonignorable missing values as well as the approach of ignoring missing values. The medium MNAR approach does not correspond to the assumptions of either approach. We ran 500 replications in each of the six conditions.

The data were analyzed using the R-package TAM (Robitzsch, Kiefer, & Wu, 2017). Item omissions were either a) scored as incorrect, b) ignored, or c) accounted for using the approach for nonignorable missing values. Rasch models were specified for the measurement model of ability as well as for missing propensity, and Marginal Maximum Likelihood Estimation (MML) was used for all model estimations. For the measurement model of item responses, we evaluated bias in item parameter estimation and person parameter estimation. When using the approach for nonignorable missing values, we also investigated the correlation of the item parameters of both measurement models, the

3) For example, for a randomly selected replication in the 30% omission proportion and strong MNAR condition this procedure resulted in omission rates on item level ranging from 10% to 49% of responses missing per item.

4) Note that the mean of the item parameters and the mean of the person parameters are zero in the simulation study, resulting on average in an equal amount of correct and incorrect responses in the data.

Table 1

Data Generating Design Used in Simulation 1

	$P(D = 1) = 0.1$			$P(D = 1) = 0.3$		
	Missing mechanism			**Missing mechanism**		
Conditional probability	**MCAR**	**Medium MNAR**	**Strong MNAR**	**MCAR**	**Medium MNAR**	**Strong MNAR**
$P(D = 1\|Y = 1)$	0.1	0.05	0	0.3	0.15	0
$P(D = 1\|Y = 0)$	0.1	0.15	0.2	0.3	0.45	0.6

Note. Y denotes the response variable with $Y = 1$ indicating a correct response and $Y = 0$ an incorrect response; D denotes the missing indicator as described in Equation 1.

variance of the missing propensity, as well as the correlation between ability and missing propensity.

Results

Table 2 depicts the mean and standard deviation of bias in the estimate of the mean ability across all 500 replications for each of the six conditions and three analysis approaches. In the data generating condition of MCAR, ignoring missing values and the approach for nonignorable missing values performed well. There was no average bias and the bias within a single replication was rather small (low standard deviation of bias across replications). For MCAR, incorrect scoring resulted in very large average bias. The bias increased with increased proportion of missing values. For strong MNAR, in which only otherwise incorrect responses are missing, the approach of incorrect scoring resulted in unbiased estimates of the mean of the person parameters. In the condition of strong MNAR, both, ignoring and the approach for nonignorable missing values resulted in highly biased estimates of average ability. The bias increased with an increase in the proportion of missing values. The approach for nonignorable missing values performed slightly better than the approach of ignoring missing values. The bias was similar in size across all levels of ability (see Figure S1 in Supplementary materials). None of the three approaches yielded unbiased estimates in the medium MNAR condition.

Given that the average of the item parameters was fixed to zero, that is to the true values, in the estimation, the approach of ignoring and the approach of nonignoring yielded unbiased item parameter estimates in all conditions (*SD* of item bias < 0.06). However, for incorrect scoring, single item parameters became more biased, when the missing mechanism became less ignorable (*SD* of item bias to up to 0.148 logits for 10% missing values and up to 0.316 for 30% missing values).

Table 2

Mean (SD) Bias in the Estimated Average Ability in Simulation 1

	$P(D = 1) = 0.1$			$P(D = 1) = 0.3$		
	Missing mechanism			**Missing mechanism**		
Approach	**MCAR**	**medium MNAR**	**strong MNAR**	**MCAR**	**medium MNAR**	**strong MNAR**
Incorrect	-0.263 (0.017)	-0.136 (0.019)	0.000 (0.019)	-0.737 (0.015)	-0.384 (0.017)	0.000 (0.019)
Ignore	-0.001 (0.020)	0.114 (0.020)	0.230 (0.020)	-0.001 (0.020)	0.453 (0.021)	0.957 (0.021)
Non-ignoring	-0.001 (0.020)	0.114 (0.020)	0.227 (0.020)	-0.001 (0.020)	0.439 (0.022)	0.931 (0.021)

In order to investigate whether the approach for nonignorable missing values reflected the (non-)ignorability of the data generating mechanism, we evaluated the estimated correlation of the item difficulties and item parameters for the missing indicators $cor(\beta, \delta)$, the estimated variance of the missing propensity $var(\xi)$, as well as the estimated correlation of the missing propensity with the latent ability $cor(\theta, \xi)$ (Table 3). As expected, when the missing mechanism was MCAR, these parameters were all close to zero. There was no correlation between the missing indicators and, thus, there was no common latent missing propensity. The correlation $cor(\beta, \delta)$ was close to zero as the probability of missing an item was the same for all responses and all items. In contrast, for all MNAR conditions, the correlation $cor(\beta, \delta)$ was very high and negative. This reflects the dependency of the probability of missing responses on the actual response with more difficult items showing higher omission rates. The variance of the missing propensity was large, when the probability of missing an item largely differed between correct and incorrect responses, that is, the more severe the missing mechanism was. The variance of the missing propensity reflected the relationships between the missing indicators. Furthermore, the estimated correlation between the missing propensity and ability increased with the amount of nonignorability. The parameters became larger with an increase in the proportion of missing values. Thus, although the approach for nonignorable missing values could not appropriately deal with this kind of nonignorability, it did reflect when the missing mechanism was ignorable and when it deviated from ignorability.

Table 3

Estimated Parameters Regarding the Measurement Model for the Missing Indicators

	$P(D = 1) = 0.1$			$P(D = 1) = 0.3$		
	Missing mechanism			**Missing mechanism**		
Estimated Parameter	**MCAR**	**medium MNAR**	**strong MNAR**	**MCAR**	**medium MNAR**	**strong MNAR**
$\widehat{cor(\beta, \delta)}$	-0.004	-0.951	-0.966	0.003	-0.984	-0.985
$\widehat{var(\xi)}$	0.016	0.018	0.238	0.015	0.113	0.396
$\widehat{cor(\theta, \xi)}$	0.000	-0.092	-0.868	0.000	-0.731	-0.915

Simulation 2: Missing Values Depending on Missing Propensity

Data Generation and Data Analyses

We simulated data according to the approach for nonignorable missing values (Equation 2). Parameters were set at the estimated values found in the data analyses of simulation 1 via the approach for nonignorable missing values. We restricted our second simulation to the condition of strong MNAR and 30% missing values, as this is the most severe case. If there are any differences, these should be most pronounced in this condition. From previous research we know that the model for nonignorable missing values results in unbiased parameter estimates under such data generation (e.g. Holman & Glas, 2005). Thus, we expected the same parameter estimates when using the model for nonignorable missing values for data analysis in simulation 1 and simulation 2.

While in simulation 1, the probability of a missing value was fixed for a given item response (being 0 for correct responses and 0.6 for incorrect ones in the considered condition), in the data generating mechanism of simulation 2, it varied across persons and items. The average missing probability for incorrect responses across all persons and items was .406 with a *SD* of 0.155. The average missing probability for correct answers was .246 with a *SD* of 0.131 across persons and items. Omission rates on item level for a randomly selected replication ranged from 9% to 49% of responses missing per item. The data were analyzed corresponding to simulation 1 and the results were compared between the two simulation studies.

Results

The approach for nonignorable missing values was able to retrieve the true parameters. There was no bias in any of the parameters estimated in this model (see Table 4). Ignoring missing values resulted in only slightly biased parameters, thus, showing robustness to certain kinds of violations of the ignorability assumption. The bias was similar in size across all levels of ability (see Figure S2 in the Supplementary Materials).

Table 4

Bias in Parameter Estimates in Simulation 2

Approach	$\widehat{E(\theta)}$	$\widehat{var(\theta)}$	$\widehat{cor(\beta, \delta)}$	$\widehat{var(\xi)}$	$\widehat{cor(\theta, \xi)}$
Incorrect	-0.9251	-0.1326	NA	NA	NA
Ignore	0.0389	-0.0243	NA	NA	NA
Nonignoring	0.0011	0.0024	0.0043	0.0044	0.007

Note that the differences in the estimated average ability between the three approaches are similar in simulation 1 and 2. Thus, we cannot infer the underlying missing mechanism from the difference in estimates between the different approaches, as this does not differ between the different mechanisms.

In order to look for possible indicators in the data that may help to distinguish the missing process in simulation 1 from that in simulation 2, we investigated model fit indices when using the approach for nonignorable missing values (Table 5). Overall, there was no serious indication of model misfit in either of the two simulation, which would hint at a model misspecification. Thus, one can hardly infer from model fit to the underlying data generating mechanism. We also evaluated the correlation of the missing indicators and found no noticeable differences between the two data generating processes. Both data generations result in substantial correlations between missing indicators (simulation 1: $M = 0.0636$, $SD = 0.002$; simulation 2: $M = 0.0686$, $SD = 0.003$). Summarizing the results, we could not find any indicators from data analyses that could help us to distinguish between the two missing data mechanisms.

Table 5

Model Fit Indices Using the Approach for Nonignorable Missing Values in Simulation 1 and 2

Simulation	BIC	*M* (infit)	*SD* (infit)	*M* (outfit)	*SD* (outfit)
Sim 1	176169.3	1.0004	0.01530	1.00375	0.03459
Sim 2	176316.9	1.0001	0.01507	1.00008	0.03344

Simulation 3: Missing Values Depending on Item Responses and Missing Propensity

Data Generation and Data Analyses

We generated missing data according to the following formula for the probability of a missing response:

$$P(D_i = 1|Y_i, \xi) = \frac{\exp(\xi - \beta_i + c_0 \cdot I_{0i} + c_1 \cdot I_{1i})}{1 + \exp(\xi - \beta_i + c_0 \cdot I_{0i} + c_1 \cdot I_{1i})} \quad (4)$$

with ξ denoting the latent missing propensity as defined by Holman and Glas (2005) and β_i denoting the respective item parameters. I_{0i} is an indicator variable being 1, if $Y_{ij} = 0$ and zero otherwise and I_{1i} being 1, if $Y_{ij} = 1$ and zero otherwise. c_0 is defined as the logit of the missing probabilities for incorrect responses and c_1 as the logit of the missing probabilities for correct responses used in simulation 1 (Table 1)[5]. Thus, item responses have a similar impact in simulation 3 as they have in simulation 1: in the case of low (high) proportion of missing values, for $c_0 = logit(0.1)$ and $c_1 = logit(0.1)$ [$c_0 = logit(0.3)$ and $c_1 = logit(0.3)$] item responses have no impact, for $c_0 = logit(0.15)$ and $c_1 = logit(0.05)$ [$c_0 = logit(0.45)$ and $c_1 = logit(0.15)$] item responses have a medium impact, and for c_0 = 0.2 and $c_1 = 0.0001$ [(c_0 = 0.6 and $c_1 = 0.0001$)] item responses have a strong impact on the missing probability. There are six conditions, two percentages of missing data (about 10% [low] and 30% [high]) and three sets of c parameters describing the impact of the item response on the missing indicator. We used the same item and person parameters for the item responses and missing indicators as in simulation 2 with one exception: To ensure comparable amounts of missing values across simulations, the mean ability parameter and the mean missing propensity parameter were set to 0. The resulting average missing probabilities for incorrect responses and correct responses are depicted in Table 6.

Table 6

Average Probabilities (SD) for Missing Values on Correct and Incorrect Responses Across Persons and Items in the Third Simulation

Conditional probabilities	$P(D = 1) \sim 0.1$			$P(D = 1) \sim 0.3$		
	Impact of item responses			Impact of item responses		
	no	medium	strong	no	medium	strong
$P(D = 1 \mid Y = 0)$	0.166 (0.096)	0.234 (0.121)	0.295 (0.138)	0.405 (0.157)	0.548 (0.162)	0.677 (0.147)
$P(D = 1 \mid Y = 1)$	0.085 (0.059)	0.043 (0.033)	0.000 (0.000)	0.246 (0.130)	0.126 (0.081)	0.000 (0.000)

5) Note that for a missing probability of zero, we set $c_1 = logit(0.0001)$.

Note that in simulation 3 the proportion of missing values is slightly larger than 10 and 30 percent. This resulted in missing rates on item level ranging from 6 to 66% for a randomly selected replication in the condition with about 30% omitted responses and strong MNAR.

Data analyses were in accordance with simulation 1 and 2.

Results

The bias was similar in size across all levels of ability (see Figure S3 in the Supplementary Materials). Table 7 shows the bias in average person parameter estimates using each of the three analysis approaches. As can be seen, the bias was very similar to the bias found in simulation 1. In conditions in which item responses have an impact on the missing propensity, nonignoring as well as ignoring missing data failed to recover the true person parameters. The approach of incorrect scoring only resulted in unbiased estimates, when missing values only occur on otherwise incorrect responses (strong impact of item responses).

Table 7

Mean (SD) Bias in Item Parameter Estimation for Different Missing Data Mechanism in Simulation 3

	$P(D = 1)$~0.1			$P(D = 1)$~0.3		
	Impact of item responses			**Impact of item responses**		
Approach	**no**	**medium**	**strong**	**no**	**medium**	**strong**
Incorrect	-0.235 (0.008)	-0.119 (0.008)	0.000 (0.008)	-0.692 (0.008)	-0.349 (0.008)	0.000 (0.008)
Ignore	0.012 (0.008)	0.148 (0.008)	0.285 (0.008)	0.041 (0.010)	0.525 (0.009)	1.060 (0.010)
Nonignoring	0.000 (0.008)	0.132 (0.008)	0.265 (0.008)	0.000 (0.010)	0.462 (0.009)	0.981 (0.010)

This is because in these data generating conditions, the impact of the item response on the probability of a missing response was very strong, while it was much lower for the missing propensity. This can be seen in Table 6: The difference between the average conditional probability for a missing response between correct and incorrect responses (which is represented by the effect of the item response) is larger than the variation of these conditional abilities across persons (which is represented by the effect of the missing propensity). In the case of strong impact, the variation of conditional probabilities of missing responses for correct responses is even zero across persons. That means that persons do not differ in that probability and, since its average is zero, missing values only occurred on incorrect responses. Thus, in conditions in which the missing probability is a function of item response and missing propensity, for values that may

be assumed in practice, the impact of item response is stronger than the impact of the missing propensity.

The approach for nonignorable missing values reflected not only the nonignorability of the data due to the missing propensity, but also due to the impact of the item response (Table 8). In the condition of no impact of item response on the missing probability, the parameters correspond to the simulated ones. With greater impact of item responses on the missing probability, the variance of the missing propensity and the correlation of the missing propensity with ability increased.

Table 8

Parameter Estimates of the Approach for Nonignorable Missing Values Regarding the Measurement Model for the Missing Indicators in Simulation 3

	Amount of missing responses					
	low			high		
	Impact of item responses			Impact of item responses		
Estimated parameter	no	medium	strong	no	medium	strong
$\widehat{cor}(\beta, \delta)$	-0.979	-0.983	-0.982	-0.981	-0.985	-0.983
$\widehat{var}(\xi)$	0.403	0.688	0.892	0.401	0.740	1.053
$\widehat{cor}(\theta, \xi)$	-0.906	-0.938	-0.947	-0.908	-0.939	-0.924

Empirical Example

In the following empirical example, we illustrate what can be inferred from analyses results in practice. We reanalyzed the Italian PISA 2012 data on assessing math competence, which is publicly available (see Supplementary materials). For illustrative purposes, we only used the responses to the first booklet. We excluded 43 persons with not reached items, as we wanted to focus on dealing with omitted items. The resulting data subsample consisted of 2616 persons and 22 items. In this subsample 11.7% of all responses were missing due to item omission. The missing rates on item level ranged from 0.5% to 47% (i.e., percentages of responses per item being missing).

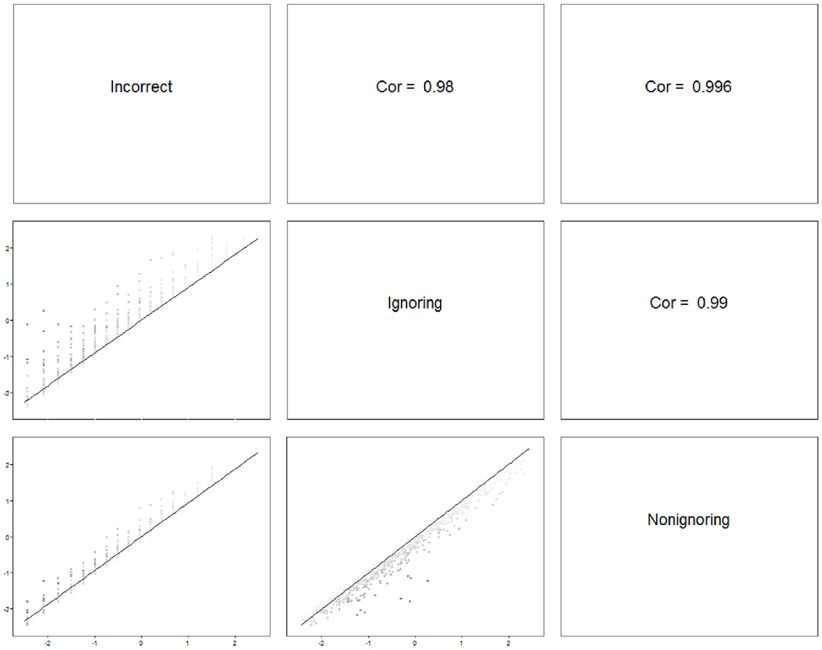
Applying the approach for nonignorable missing values resulted in an estimated correlation between the item difficulties and item parameters for the missing indicators of -0.664, indicating that more difficult items are more often omitted. The estimated variance of the omission propensity was large $\widehat{var}(\xi)$ = 2.894 and omission propensity highly correlated with ability $\widehat{cor}(\theta, \xi)$ = -0.605, indicating that there is a common variable underlying the missing indicators across items and that the persons with lower ability tend to omit more items. These correlations are lower than those found in the simula-

tions in which the missing indicator depended on the item response (simulation 1 and 3). They are nevertheless considerable and indicate nonignorability of item omissions.

We compared the ability estimates using each of the three approaches. Fixing the average item difficulty in all three analyses to zero, we found considerable differences in estimated mean ability: Incorrect scoring results in much lower average ability estimates $\widehat{E(\theta)}$ = -0.052 than ignoring missing values $\widehat{E(\theta)}$ = 0.167 or nonignoring $\widehat{E(\theta)}$ = 0.145. The comparison of EAP estimates across the different approaches (Figure 1) show that the differences in ability estimates occur specifically for persons with many missing values, which in this application are the ones with lower ability. Similar to all three simulation studies, incorrect scoring leads to lower ability scores for persons with missing values compared to the other two approaches, and ignoring results in even higher scores than nonignoring.

Figure 1

EAP-Estimates of Person Ability for the Domain Mathematics in a Single Booklet of the Italian PISA 2012 Data Set



Note. Darker points represent more missing values; For this graph the mean of the person parameters was set to 0.

The non-zero correlation between ability and missing propensity indicates that there is some form of nonignorability in the data. From the results we can, however, not infer

to the specific kind of nonignorable mechanism. Any of the nonignorable mechanisms described in Equation 1 to 3, and also more complex forms including further covariates, may be responsible for the results.

Discussion

Approaches for nonignorable missing values have often been over-interpreted by users of being able to deal with all kinds of nonignorable missing data. So far, the performance of the different approaches for missing values has only been investigated on a limited set of nonignorable missing data mechanisms. We investigated the performance of incorrect scoring, ignoring missing values, and using the approach for nonignorable missing values on three different kinds of nonignorable missing data mechanisms. The three mechanisms differ in whether the probability of a missing value is a) a function of the item response, b) a function of a unidimensional missing propensity (and the respective model parameters), or c) a function of both, item response and missing propensity. The results found in this study are in line with previous studies. Similar to Holman and Glas (2005) as well as Rose and colleagues (2010), we found good performance of the approach for nonignorable missing values under all missing data conditions in simulation 2. Ignoring missing values resulted in only slightly biased parameters and incorrect scoring resulted in highly biased parameter estimates.

Additional to previous studies, we also investigated the performance of the approaches under other nonignorable missing data approaches. There is no approach that can deal with all mechanisms of nonignorable missing values. Regarding nonignorable missing data mechanisms, the approach for nonignorable missing values is only appropriate under a missing data mechanism, in which the probability of a missing value is solely a function of the missing propensity. For any of the other mechanisms it results in biased parameter estimates. Ignoring missing values results in similar estimates as the approach for nonignorable missing values with only slightly larger bias. Incorrect scoring only results in unbiased parameter estimates when missing values solely occur on incorrect responses. Note that none of the approaches can deal with missing data mechanisms in which the missing value depends on the item response and missing values also occur on correct responses.

The simulation studies in this paper have some limitations. First, we did only consider Rasch models as measurement model for item responses as well as missing indicators. In practice often other models, such as the 2PL or 3PL model or (generalized) partial credit model are used. While the current study did not investigate this, the results will most likely also hold for these types of measurement models as previous research (Glas & Pimentel, 2008; Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) has found no substantial different effect. Second, in this study we assume that ability (and missing propensity) are unidimensional and (multivariate-)

normal distributed. Previous research acknowledged that the missing process may be even more complex and tried to more closely depict that mechanism by including further covariates (Glas, Pimentel, & Lamers, 2015; Köhler et al., 2015b; Moustaki & Knott, 2000) or relaxing model assumptions (Köhler, Pohl, & Carstensen, 2015a; Rose, 2013). In practice, these assumptions may be tested and appropriate model extensions included. Third, we only simulated a selected variety of conditions, which do not cover the whole range of possible values that may occur in practice. The simulated condition suffices for evaluating for which missing data mechanisms the different approaches are suited. For more information on the impact of different amounts of missing values or a different correlation between ability and missing propensity, we refer the reader to previous work (Holman & Glas, 2005; Rose, von Davier, & Xu, 2010; Sachse, Mahler, & Pohl, 2019). Fourth, we did not make use of all information available in current testing. Recent research (Lu, Wang, & Tao, 2018; Pohl, Ulitzsch, & von Davier, 2019; Ulitzsch, von Davier, & Pohl, 2019) has shown that response time data may provide useful information for disentangling and modeling different missing data mechanisms.

Even though the nonignoring approach cannot account for all types of missing mechanisms, it can be useful for investigating the mechanism of missing values as it does reflect the existence of some form of nonignorability of the missing process in the data. The results show that one needs to be cautious when interpreting the results of analyses using any of the proposed approaches. When choosing an approach for dealing with missing values, we must evaluate the plausibility of each underlying mechanism by making use of pilot studies, empirical analyses, and theory. We must discuss our assumptions on the underlying mechanism to make our arguments available to readers who may then judge whether these are plausible. If a nonignorable mechanism as described in Equation 2 seems to be most plausible, then the approach for nonignorable missing values should be used. Ignoring missing values is a good alternative if the correlation is not that high (< 0.3, see Holman & Glas, 2005). If one assumes that missing values only occur on otherwise incorrect responses (special case of Equation 1), scoring missing values as incorrect would be the appropriate analysis approach. If any form other form of dependency of missing values on the response itself (aside from missing responses only occurring on otherwise incorrect responses), no approach will be able to result in unbiased estimates. A researcher should be aware of this and consider this in the interpretation of the results.

Based on work of Mislevy and Wu (1996), Robitzsch (2016) proposed a unified framework, in which the three approaches discussed in this paper can all be subsumed. It is available in the R-package sirt (Robitzsch, 2017). The model can incorporate all three missing data mechanisms discussed in our paper, even those in which the three approaches fail to recover the true parameters. However, the model requires a specification of the impact of item responses on the missing indicator. As such, it assumes that this impact is known. In practice we usually do not know this impact, instead that is what we

usually want to investigate. In such situations the framework cannot help in identifying the missing mechanism, it can, however, be used for sensitivity analyses.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO1655/2-1).

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: We thank two anonymous reviewers for helpul comments on our manuscript.

Data Availability: Data for the empirical example is freely available from the OECD (see Supplementary Materials section).

Supplementary Materials

1. Scatterplots of EAP-estimates of person ability for Simulations 1 to 3 for a single replication are available via the PsychArchives repository (for access see Index of Supplementary Materials below).
2. Data for the empirical example (Italian PISA 2012 data) is freely available from the OECD (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Pohl, S., & Becker, B. (2020). *Supplementary materials to: Performance of missing data approaches under nonignorable missing data conditions*. PsychOpen. https://doi.org/10.23668/psycharchives.2905

The Organisation for Economic Co-operation and Development (OECD), Programme for International Student Assessment (PISA). (2012). *PISA 2012 dataset*. PISA Database. https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

References

Culbertson, M. (2011). *Is it wrong? Handling missing responses in IRT* [Paper presentation]. The annual meeting of the National Council of Measurement in Education, New Orleans, LA, USA.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213-234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*(6), 907-922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling, 57*(4), 523-541.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical & Statistical Psychology, 58*(1), 1-17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology, 14*, 3-15. https://doi.org/10.1027/1614-2241/a000141

Köhler, C., Pohl, S., & Carstensen, C. H. (2015a). Taking the missing propensity into account when estimating competence scores – Evaluation of IRT models for non-ignorable omissions. *Educational and Psychological Measurement, 75*(5), 850-874. https://doi.org/10.1177/0013164414561785

Köhler, C., Pohl, S., & Carstensen, C. H. (2015b). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57*(4), 499-522.

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

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*(2), 247-264. https://doi.org/10.1007/BF02291471

Lu, J., Wang, C., & Tao, J. (2018). *Modeling nonignorable missing for not-reached items incorporating item response times.* Talk given at the annual meeting of the Psychometric Society, NY, USA.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*. Princeton, NJ, USA: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163*(3), 445-459. https://doi.org/10.1111/1467-985X.00177

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A, 162*(2), 177-194. https://doi.org/10.1111/1467-985X.00129

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423-452. https://doi.org/10.1177/0013164413504926

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika, 84*(3), 892-920. https://doi.org/10.1007/s11336-019-09669-2

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Robitzsch, A. (2016). Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment. In B. Suchań, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS & TIMSS 2011: Die Kompetenzen in Lesen, Mathematik und Naturwissenschaften am Ende der Volksschule* (pp. 55-64). Graz, Austria: Leykam.

Robitzsch, A. (2017). sirt: Supplementary Item Response Theory Models (R package version 2.3-57). Retrieved from https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules (R package version 2.4-9). Retrieved from https://CRAN.R-project.org/package=TAM

Rohwer, G. (2013). *Making sense of missing answers in competence tests* (NEPS working paper No. 30). National Educational Panel Study, University of Bamberg, Bamberg, Germany.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* [Doctoral thesis, Friedrich-Schiller-University, Jena, Germany]. Retrieved from https://www.db-thueringen.de/receive/dbt_mods_00022476

Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with Item Response Theory (IRT). *ETS Research Reports Series, 2010*(1), i-53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

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

Sachse, K., Mahler, N., & Pohl, S. (2019). Effects of changing nonresponse mechanisms on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement, 79*(4), 699-726. https://doi.org/10.1177/0013164419829196

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2019.1643699

Vidotto, D., Vermunt, J. K., & Van Deun, K. (2018). Bayesian latent class models for the multiple imputation of categorical data. *Methodology, 14*, 56-68. https://doi.org/10.1027/1614-2241/a000146

Methodology
2020, Vol.16(2), 147–165
https://doi.org/10.5964/meth.2805