



# On the Treatment of Missing Data in Background Questionnaires in Educational Large-Scale Assessments: An Evaluation of Different Procedures

Simon Grund 

Oliver Lüdtke

Alexander Robitzsch

*Leibniz Institute for Science and Mathematics Education, Kiel, Germany  
Centre for International Student Assessment, Germany*

*Large-scale assessments (LSAs) use Mislevy's "plausible value" (PV) approach to relate student proficiency to noncognitive variables administered in a background questionnaire. This method requires background variables to be completely observed, a requirement that is seldom fulfilled. In this article, we evaluate and compare the properties of methods used in current practice for dealing with missing data in background variables in educational LSAs, which rely on the missing indicator method (MIM), with other methods based on multiple imputation. In this context, we present a fully conditional specification (FCS) approach that allows for a joint treatment of PVs and missing data. Using theoretical arguments and two simulation studies, we illustrate under what conditions the MIM provides biased or unbiased estimates of population parameters and provide evidence that methods such as FCS can provide an effective alternative to the MIM. We discuss the strengths and weaknesses of the approaches and outline potential consequences for operational practice in educational LSAs. An illustration is provided using data from the PISA 2015 study.*

**Keywords:** *missing data; multiple imputation; plausible values; measurement error; large-scale assessment*

One of the main goals of educational large-scale assessments (LSAs) such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) is to provide information about student proficiency and the relations between student proficiency and other cognitive and noncognitive variables such as students' socioeconomic status, self-concept, attitudes, or interests (Singer & Braun, 2018). To this end, LSAs employ scaling procedures by which proficiency scores are estimated for each student on the basis of (a) their performance on an achievement test and (b) the information they provide in a background questionnaire. Combining these two

sources of information, the scaling procedure is often used to generate a set of “plausible values” (PVs) for the proficiency of each student (Mislevy, 1991). This method follows Rubin’s (1987) multiple imputation (MI) approach by regarding the latent proficiency scores as missing data and allows for unbiased estimates of population parameters (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; for other approaches to estimating population parameters, see also Culpepper & Park, 2017; Rijmen et al., 2014; von Davier & Sinharay, 2007, 2010).

An important requirement of these scaling procedures is that the background variables are completely observed. However, missing data are common in LSAs. They occur, for example, because students fail to answer some of the items in the background questionnaire or because some studies use a rotated questionnaire (or planned missing data) design (Graham et al., 2006; Rhemtulla & Hancock, 2016). For example, socioeconomic status is often missing for a substantial number of participants, thereby making it difficult to derive statements about the relation between students’ socioeconomic background and educational attainment. This difficulty raises the question of how missing data in background variables should be treated in educational LSAs.

The treatment of missing data in LSAs is further complicated by the fact that the data are prepared and used by two different parties sometimes called the “imputer,” whose task is to generate the PVs, and the secondary “analyst” who uses the data to answer a substantive research question (Meng, 1994; see also Shin, 2013). For the imputer, an ideal statistical procedure would allow for the generation of PVs while simultaneously treating the missing data in background variables (e.g., using MI; see also Blackwell et al., 2017a, 2017b). No such procedure is currently used in educational LSAs. Instead, LSAs rely on the missing indicator method (MIM; Cohen & Cohen, 1975) whose application has been discouraged in the statistical literature (e.g., Jones, 1996; Schafer & Graham, 2002). For the analyst, this method creates additional challenges because, although it provides PVs, it does not provide imputations for missing data and requires that these be treated by other means.

Several articles have investigated the treatment of missing data in background variables. von Davier (2013) provided a discussion of the challenges associated with rotated questionnaires in LSAs. Adams et al. (2013) investigated the effects of rotated questionnaires on the estimates of population parameters and concluded that these designs allow for accurate estimates of the marginal properties of student proficiency (e.g., population means). However, Rutkowski (2011) and Rutkowski and Zhou (2015) found that the current treatment of missing data in LSAs can lead to biased parameter estimates in subpopulations (e.g., means and mean differences) when missing data occur in a systematic manner. Kaplan and Su (2016, 2018) showed that different rotated questionnaire designs can differ greatly in how well they recover variables’ marginal properties and the relations between them. For the treatment of missing data, Assmann et al. (2015) proposed a Bayesian estimation procedure for estimating the scaling model

with missing background data, and Weirich et al. (2014) evaluated procedures based on two-stage (or nested) MI (see also Harel, 2007; Rubin, 2003). Finally, Wetzal et al. (2015) considered alternatives to the current practice based on latent class models.

In writing this article, we had three goals. First, we aimed to evaluate the procedures currently used for handling missing data in background variables in educational LSAs (i.e., those based on the MIM) in an attempt to clarify the conditions under which they can cause problems. Second, we aimed to implement and evaluate a strategy that allows for a joint modeling of PVs for the proficiency scores and imputations of missing data in background variables. Finally, we attempted to contrast the strengths and weaknesses of these approaches and provide recommendations for the operational practices applied in educational LSAs.

The article is organized as follows. In the first section, we review the statistical procedures used in the scaling of proficiency data and the generation of PVs in educational LSAs. In the second and third sections, we extend our discussion to the case with missing data in background variables, reviewing (a) the current practices applied in educational LSAs (based on the MIM) and (b) the joint modeling of measurement error and missing data. We then present the results of two simulation studies in which we evaluated the performance of these methods in two scenarios with different measurement models and statistical complexity. Finally, we illustrate their application with data from the PISA 2015 study (Organization for Economic Cooperation and Development [OECD], 2017). We close with a discussion of our findings and recommendations for practice.

### Statistical Models in LSAs

The statistical procedures used in educational LSAs are applied to represent the relations between students' responses on the achievement test, their latent (i.e., unobserved) proficiency on these tests, and their responses on the background questionnaire. Let  $\mathbf{y}$  denote item responses on the achievement test,  $\mathbf{x}$  the responses on the background questionnaire, and  $\boldsymbol{\theta}$  the (multidimensional) latent proficiency. Then, under the conditional independence and invariance assumptions usually employed in LSAs (e.g., absence of differential item functioning), the joint distribution of  $\mathbf{y}$ ,  $\mathbf{x}$ , and  $\boldsymbol{\theta}$  can be written as

$$P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\psi}) = P(\mathbf{y}|\boldsymbol{\theta}; \boldsymbol{\xi})P(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\psi}), \quad (1)$$

where  $P(\mathbf{y}|\boldsymbol{\theta}; \boldsymbol{\xi})$  is a measurement (or scaling) model describing the relations between the item responses and the latent proficiency, and  $P(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\psi})$  is a structural (or population) model describing the relations between the latent proficiency and the background variables. In the following, we briefly review the scaling procedures and the generation of PVs in the hypothetical case in which no data are missing in the background variables.<sup>1</sup>

### Scaling Model

In educational LSAs, the students' responses to the items on the achievement test are modeled with item response theory (IRT), which describes item responses as a function of both person ability and item characteristics. In general, IRT models express the probability of observing a response pattern  $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$  for a person  $i$  ( $i = 1, \dots, N$ ) to a set of items  $j$  ( $j = 1, \dots, J$ ) in a statistical model for each item

$$P(\mathbf{y}_i | \boldsymbol{\theta}_i; \boldsymbol{\xi}) = \prod_{j=1}^J P(y_{ij} | \boldsymbol{\theta}_i; \boldsymbol{\xi}_j), \quad (2)$$

where  $\boldsymbol{\theta}_i$  denotes the latent proficiency of person  $i$  and  $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_J)$  denotes a set of item parameters. One particular IRT model that is frequently used in educational LSAs is the generalized partial credit model (Muraki, 1992; for an overview, see von Davier & Sinharay, 2013; von Davier et al., 2006). Suppose that a multidimensional latent proficiency  $\boldsymbol{\theta}_i$  is measured with a set of items, each with  $K_j + 1$  ordered categories, that is, with possible scores  $k = 0, \dots, K_j$ . Then the probability of person  $i$  achieving score  $k$  on item  $j$  can be modeled as

$$P(y_{ij} = k | \boldsymbol{\theta}_i; \boldsymbol{\xi}_j) = \frac{\exp[k \mathbf{a}_j^T \boldsymbol{\theta}_i - b_{jk}]}{1 + \sum_{v=1}^{K_j} \exp[v \mathbf{a}_j^T \boldsymbol{\theta}_i - b_{jv}]}, \quad (3)$$

where the item parameters  $\boldsymbol{\xi}_j = (\mathbf{a}_j, \mathbf{b}_j)$  denote item slopes and intercepts, respectively (with  $b_{j0} = 0$ ). The response probabilities in the scaling model give rise to an individual likelihood function about  $\boldsymbol{\theta}_i$ , thus allowing proficiency scores to be estimated for each person. Although point estimates such as the maximum likelihood estimate (MLE; Lord, 1983) and weighted likelihood estimate (WLE; Warm, 1989) are often used to measure an individual's proficiency, their application in LSAs has been shown to be inappropriate because they do not fully account for measurement error, thus leading to biased estimates of population parameters (e.g., inflated variance of  $\boldsymbol{\theta}$ , distorted estimates of subgroup characteristics, correlations, and regression coefficients; Braun & von Davier, 2017; von Davier et al., 2009; Wu, 2005). Instead, proficiency scores in educational LSAs are generated in the form of PVs, which allow for an accurate estimation of population parameters (e.g., Mislevy, 1991; Mislevy, Beaton, et al., 1992).

### Population Model and PVs

The generation of PVs follows Rubin (1987) by regarding the latent proficiency scores as missing data, for which imputations can be generated by drawing repeatedly from the posterior predictive distribution of the proficiency  $\boldsymbol{\theta}$ , given the responses on the achievement test, the parameters of the scaling model,

and the responses on the background questionnaire (Mislevy, 1991). Let  $\mathbf{x}_i$  denote the values on the background variables for person  $i$ . Then the posterior distribution of  $\boldsymbol{\theta}_i$  can be expressed as

$$P(\boldsymbol{\theta}_i | \mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\xi}, \boldsymbol{\psi}) \propto P(\mathbf{y}_i | \boldsymbol{\theta}_i; \boldsymbol{\xi}) P(\boldsymbol{\theta}_i | \mathbf{x}_i; \boldsymbol{\psi}), \quad (4)$$

where the first term denotes the scaling model with item parameters  $\boldsymbol{\xi}$  and the second term denotes a population model for the latent proficiency with parameters  $\boldsymbol{\psi}$ . This model often takes the form of a latent regression model with parameters  $\boldsymbol{\psi} = (\boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ , in which the proficiency scores are regressed on the responses from the background questionnaire (see Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992):

$$\boldsymbol{\theta}_i = \mathbf{x}_i \boldsymbol{\Gamma} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad (5)$$

where  $\boldsymbol{\Gamma}$  is a matrix of regression coefficients, and  $\boldsymbol{\Sigma}$  is the residual variance-covariance matrix. Notice that the expression for the posterior distribution in Equation 4 is proportional to the joint distribution in Equation 1. In other words, sampling PVs from the posterior distribution given the observed data for each respondent is equivalent to obtaining samples from the joint distribution of  $\mathbf{y}$ ,  $\mathbf{x}$ , and  $\boldsymbol{\theta}$ . In practice, the parameters of the population model can be estimated with an expectation-maximization (EM) algorithm (Dempster et al., 1977; for alternative methods, see also Culpepper & Park, 2017; Johnson & Jenkins, 2004), and samples from the posterior distribution of  $\boldsymbol{\theta}$  can be obtained by substituting  $\boldsymbol{\xi}$ ,  $\boldsymbol{\Gamma}$ , and  $\boldsymbol{\Sigma}$  with their MLEs or with approximate draws from their posterior distributions (Bock & Aitkin, 1981; Mislevy, Johnson, & Muraki, 1992; Thomas & Gan, 1997; for an overview, see von Davier & Sinharay, 2013; von Davier et al., 2006).

This procedure for generating PVs can be used directly when the background variables are completely observed. In practice, however, the background variables often contain missing data and require treatment before the PVs can be generated (Adams et al., 2013; Wetzal et al., 2015). In the following, we describe the method currently used to deal with missing data in the background questionnaire in LSAs and consider its statistical properties.

### Missing Data in the Background Questionnaire

In educational LSAs, the background variables are seldom complete. Suppose that some elements of  $\mathbf{x}$  are missing, with their observed and missing parts denoted as  $\mathbf{x}^{\text{obs}}$  and  $\mathbf{x}^{\text{mis}}$ , and let  $\mathbf{r}$  be a set of indicator variables that describe which values in  $\mathbf{x}$  are missing and observed, respectively. Rubin (1976) referred to data as missing completely at random (MCAR) if the propensity for missing data is independent of both the observed and the missing data, that is,  $P(\mathbf{r} | \mathbf{x}, \mathbf{y}) = P(\mathbf{r})$ . This can occur, for example, when missing data result from the use of a rotated questionnaire (or planned missing data) design (e.g., Graham

TABLE 1.  
*Schematic Illustration of the MIM*

Continuous			Categorical										
			With a Missing Indicator						With an Extra Category				
$x$	$x^*$	$r$	$x$	$x^*$	$r$	$x_1^*$	$x_2^*$	$r$	$x$	$x^*$	$x_1^*$	$x_2^*$	$x_3^*$
-1	-1	0	0	0	0	0	0	0	0	0	0	0	0
1	→ 1	0	1	→ 1	0	→ 1	0	0	1	→ 1	→ 1	0	0
3	3	0	2	2	0	0	1	0	2	2	0	1	0
—	0	1	—	0	1	0	0	1	—	3	0	0	1

et al., 2006). By contrast, if the data are missing at random (MAR), then the propensity for missing data depends on the observed data but not on the missing data once the observed data are taken into account, that is,  $P(\mathbf{r}|\mathbf{x}, \mathbf{y}) = P(\mathbf{r}|\mathbf{x}^{\text{obs}}, \mathbf{y})$ . In the following, we assume that the background data are either MCAR or MAR.<sup>2</sup>

The current method for dealing with missing data in the background variables in LSAs relies on including the indicator variables  $\mathbf{r}$  in the latent regression model and recoding the data to temporarily render them complete (Adams et al., 2013). This strategy is also known as the missing indicator method (MIM) in the missing data literature (Allison, 2001; Cohen & Cohen, 1975). In the following, we describe this method in more detail.

### Missing Indicator Method

The general idea behind the MIM is to recode the incomplete background questionnaire data by replacing missing data with predefined values and by including additional indicator variables that represent differences between cases with missing and observed data, respectively (Adams et al., 2013). The MIM can be applied to both continuous and categorical data, which is illustrated in Table 1. For any background variable with missing data  $x_q$ , an indicator variable  $r_q$  is created such that

$$r_{iq} = \begin{cases} 1 & \text{if } x_{iq} \text{ is missing} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In addition, the background variables are recoded to render them temporarily complete. A continuous background variable  $x_q$ , possibly centered at its mean or median, is recoded into a new variable  $x_q^*$  such that

$$x_{iq}^* = \begin{cases} x_{iq} & \text{if } x_{iq} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

This is illustrated in Table 1. For categorical data, the same principle is used but combined with dummy or effect coding. For example, when using dummy codes to represent a categorical variable  $x_q$  with  $L_q$  categories and values  $0, \dots, L_q - 1$ , the variable is recoded into  $L_q - 1$  dummy variables  $x_{ql}^*$  ( $l = 1, \dots, L_q - 1$ ) such that

$$x_{iql}^* = \begin{cases} 1 & \text{if } x_{iq} = l \\ 0 & \text{if } x_{iq} \neq l \text{ or missing,} \end{cases} \quad (8)$$

where the remaining category acts as a reference category. This strategy is sometimes described as treating missing responses as an “extra category” because the missing indicator  $r_q$  acts as an additional dummy variable in the coding scheme. In other words, treating missing responses as an extra category is equivalent with coding missing responses as zero and adding a missing indicator to the coding scheme. This is illustrated in Table 1 for a single variable with three categories.

The recoded variables  $\mathbf{x}^*$  and the indicators  $\mathbf{r}$  are then used as conditioning variables in the latent regression model, often in the form of principal components or a similar lower dimensional representation, and the PVs are generated from the (adjusted) posterior distribution

$$P(\boldsymbol{\theta}_i | \mathbf{y}_i, \mathbf{x}_i^*, \mathbf{r}_i; \boldsymbol{\xi}, \boldsymbol{\psi}) \propto P(\mathbf{y}_i | \boldsymbol{\theta}_i; \boldsymbol{\xi}) P(\boldsymbol{\theta}_i | \mathbf{x}_i^*, \mathbf{r}_i; \boldsymbol{\psi}). \quad (9)$$

Because the variables  $\mathbf{x}^*$  take on a constant value whenever the corresponding values in  $\mathbf{x}$  are missing, the recoded background variables contribute to the posterior distribution only to the extent to which they have data (for further discussion, see von Davier, 2013).

Despite its convenience, the MIM has been criticized because it can distort parameter estimates (Jones, 1996). The same may be true in LSAs when PVs are generated under the MIM (see also Rutkowski, 2011). Furthermore, because the MIM provides only a temporary solution to missing data in background variables, secondary analysts must take additional steps to treat them, for example, by removing cases with incomplete data (e.g., listwise deletion) or by using MI (see also Schafer & Graham, 2002).<sup>3</sup> It is largely an open question whether using the MIM, by itself or in combination with these methods for handling missing data, can provide unbiased estimates of population parameters in LSAs. In the following, we consider this question in more detail.

### *Estimators Under the MIM*

Depending on the method used to treat missing data after generating the PVs via the MIM, it is possible to obtain different estimators for population parameters. In the following, we consider three such methods: pairwise estimation of variances and covariances (MIM-PE), listwise deletion (MIM-LD), and nested MI of the incomplete data (MIM-MI). The three methods provide different

estimators for the population parameters that differ in how they make use of the data. Specifically, MIM-PE estimates variances and covariances on the basis of the (pairwise) available data, and other parameters such as correlation or regression coefficients are then estimated on the basis of the variance-covariance matrix. MIM-LD estimates variances and covariances in the same way, but other parameters such as correlation and regression coefficients are obtained on the basis of only the subset of the data in which all of the required variables are complete. Finally, MIM-MI generates replacements for missing data in background variables on the basis of the PVs generated under the MIM and the observed data, and the population parameters of interest are then estimated on the basis of the imputed data.

To illustrate the difference between the estimators, consider a hypothetical scenario with two variables  $u$  and  $v$ , both of which contain missing data. In this case, there are two missing data indicators  $r_u$  and  $r_v$ . Suppose the parameter of interest is the correlation between  $u$  and  $v$ . Using MIM-PE, the correlation is estimated by first estimating the variances and covariances of the variables using the (pairwise) available data, that is,

$$\hat{\rho}_{uv}^{\text{PE}} = \frac{\hat{\sigma}_{uv, r_u=0, r_v=0}}{\hat{\sigma}_{u, r_u=0, r_v \in \{0,1\}} \hat{\sigma}_{v, r_u \in \{0,1\}, r_v=0}}, \quad (10)$$

where the subscripts refer to the subsets of the data in which  $u$  and  $v$  are observed ( $r = 0$ ) or missing ( $r = 1$ ), respectively. By contrast, using MIM-LD, the correlation is estimated using only the subset of the data in which both  $u$  and  $v$  are observed, that is,

$$\hat{\rho}_{uv}^{\text{LD}} = \frac{\hat{\sigma}_{uv, r_u=0, r_v=0}}{\hat{\sigma}_{u, r_u=0, r_v=0} \hat{\sigma}_{v, r_u=0, r_v=0}}. \quad (11)$$

Although the two estimators use the same information about the covariance between the variables, they make different use of the information that is available about the variances. The same principle applies more generally to all parameters whose sufficient statistics include the variance-covariance matrix of the variables. For example, suppose that  $\mathbf{v}$  now includes two or more variables, and the parameters of interest are the regression coefficients in the multiple regression model

$$u = \mathbf{v}\boldsymbol{\beta} + e. \quad (12)$$

The sufficient statistics for the regression coefficients (ignoring the intercept) are the variances and covariances among the variables, and the coefficients can be estimated as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{uv} \hat{\boldsymbol{\Sigma}}_v^{-1}. \quad (13)$$

In MIM-PE, the variances and covariances are estimated using the (pairwise) available data, that is, every element in the variance-covariance matrix is



estimated separately using the data available for the respective (pair of) variables. In MIM-LD, all values in the variance-covariance matrix are estimated at once using the subset of the data in which all variables are observed. Finally, in MIM-MI, the variances and covariances are estimated using the imputed data. In the following, we discuss the statistical properties of the different estimators under the MIM in more detail.

### *Statistical Properties of the MIM*

To better understand the statistical properties of the MIM, we derived the asymptotic bias in the parameter estimates under the MIM in a “minimal” setting with one latent variable  $\theta$ , measured by a single indicator  $y$ , and two background variables  $x_1$  and  $x_2$ . We assumed that the sample was large ( $n \rightarrow \infty$ ) and that  $x_1$  was MCAR. The parameters of interest were the variances and covariances among the variables as well as correlation and regression coefficients. In the following, we present the main results of this investigation. The full derivation is presented in the Supplement A in the online version of this article. The main results for the variances and covariances are also summarized in Table 2, and an additional simulation study confirming and illustrating the results is presented in Online Supplement B.

*Pairwise estimation.* Our investigation showed that, under MIM-PE, the variances and covariances can be estimated without bias under MCAR. Consequently, this would also allow for an unbiased estimation of other parameters such as correlation and regression coefficients by employing a two-step procedure in which the variances and covariances of the variables are estimated in a pairwise fashion, and regression coefficients are then calculated on the basis of the variance-covariance matrix. This is an important finding because it shows that unbiased estimates of many key parameters can be obtained under the MIM at least when the data are MCAR.

*Listwise deletion.* Under MIM-LD, the variances and covariances are estimated without bias, as with MIM-PE. However, the estimates of correlation and regression coefficients can be biased even when the data are MCAR. This is because variances and covariances can be biased in the subsets of the data that pertain to complete and incomplete cases, respectively ( $\text{MIM}_{r=0}$  and  $\text{MIM}_{r=1}$ ; see Table 2). For example, the estimated regression coefficients for the regression of  $\theta$  on  $x_1$  and  $x_2$  under MIM-LD are

$$\hat{\beta}_1 = \beta_1 - (1 - \lambda_c)B\rho_{12} \quad \text{and} \quad \hat{\beta}_2 = \beta_2 + (1 - \lambda_c)B, \quad (14)$$

where  $B = \beta_1 \frac{p\rho_{12}^2}{1-(1-p)\rho_{12}^2}$ ,  $p$  is the proportion of missing data,  $\lambda_c$  is the conditional reliability of  $y$  as a measure of  $\theta$  given  $x_1$  and  $x_2$ , and  $\rho_{12}$  is the correlation between  $x_1$  and  $x_2$  (for details, see Online Supplement A). For other parameters,

TABLE 2.

*Expected Values for the Estimated Variances and Covariances Under the MIM Organized by the Pattern of Missing Data and the Method Used to Treat Missing Data Under MCAR*

Estimator	$\text{Var}(\theta)$	$\text{Var}(x_1)$	$\text{Var}(x_2)$	$\text{Cov}(\theta, x_1)$	$\text{Cov}(\theta, x_2)$	$\text{Cov}(x_1, x_2)$
True value	$\sigma_\theta^2$	$\sigma_{x_1}^2$	$\sigma_{x_2}^2$	$\sigma_{\theta x_1}$	$\sigma_{\theta x_2}$	$\sigma_{x_1 x_2}$
By pattern of missing data						
MIM <sub>r=0</sub>	$\sigma_\theta^2 + pV$	$\sigma_{x_1}^2$	$\sigma_{x_2}^2$	$\sigma_{\theta x_1}$	$\sigma_{\theta x_2} + pC$	$\sigma_{x_1 x_2}$
MIM <sub>r=1</sub>	$\sigma_\theta^2 - pV$		$\sigma_{x_2}^2$		$\sigma_{\theta x_2} - pC$	
By method for the treatment of missing data						
MIM-PE	$\sigma_\theta^2$	$\sigma_{x_1}^2$	$\sigma_{x_2}^2$	$\sigma_{\theta x_1}$	$\sigma_{\theta x_2}$	$\sigma_{x_1 x_2}$
MIM-LD	$\sigma_\theta^2$	$\sigma_{x_1}^2$	$\sigma_{x_2}^2$	$\sigma_{\theta x_1}$	$\sigma_{\theta x_2}$	$\sigma_{x_1 x_2}$
MIM-MI	$\sigma_\theta^2$	$\sigma_{x_1}^2 - p(\gamma_1^2 V + 2\gamma_1 \gamma_2 C)$	$\sigma_{x_2}^2$	$\sigma_{\theta x_1} - p(\gamma_1 V + \gamma_2 C)$	$\sigma_{\theta x_2}$	$\sigma_{x_1 x_2} - p\gamma_1 C$

*Note.* MIM<sub>r=0</sub> = MIM with estimates based on only the complete cases; MIM<sub>r=1</sub> = MIM with estimates based on only the incomplete cases (i.e., with missing data  $x_1$ ); MIM-PE = MIM with pairwise estimation; MIM-LD = MIM with listwise deletion; MIM-MI = MIM with nested MI;  $p$  = proportion of missing data in  $x_1$ ;  $\gamma_1, \gamma_2$  = parameters of the imputation model for  $x_1$ ;  $V, C$  = bias in variances and covariances (see Online Supplement A).

the bias is also a function of the proportion of missing data, the conditional reliability, the true regression coefficients from the regression of  $\theta$  on  $x_1$  and  $x_2$ , and the correlation between  $x_1$  and  $x_2$ . The bias appears to be relatively small in most of the cases that are relevant for practice, but it may increase if the reliability is low or the proportion of missing data is high. This illustrates that naive estimates of correlation and regression coefficients, which often correspond to the default methods in statistical software, can be biased even if the estimates of the variances and covariances are not and even if the data are MCAR.

*Nested MI.* Under MIM-MI, some but not all variances and covariances can be estimated without bias (see Table 2). Specifically, because the estimates of the variances and covariances in the subsets of the data pertaining to complete and incomplete cases (MIM<sub>r=0</sub> and MIM<sub>r=1</sub>) are biased under the MIM, the parameters of the imputation model that is used to impute missing data in  $x_1$  are biased as well. Consequently, the estimates of the parameters relating to  $x_1$  can be biased even when the data are MCAR, including some of the variances and covariances as well as the correlation and regression coefficients. The form of the bias tends to be complex even for simple parameters such as correlation and regression coefficients but is again a function of the proportion of missing data,

the conditional reliability of  $y$ , the true regression coefficients of  $x_1$  and  $x_2$ , and the correlation between  $x_1$  and  $x_2$ .

In summary, our investigation suggested that, although unbiased estimates can be obtained with MIM-PE, naive use of MIM-LD or even MIM-MI can lead to distorted parameter estimates even if the data are MCAR. Nonetheless, these results should be interpreted with care due to the restrictive nature of the assumptions made in their derivation. In the following section, we present an alternative to the MIM based on a joint treatment of PVs and missing data in the background questionnaire. Then, we present the results of two simulation studies that we conducted to evaluate these procedures in more general and realistic scenarios.

### Joint Treatment of PVs and Missing Data

If there are missing data in the background variables, the joint distribution of  $y$ ,  $\mathbf{x}$ , and  $\boldsymbol{\theta}$  in Equation 1 can be extended to include both the observed and missing parts of  $\mathbf{x}$ , that is, both  $\mathbf{x}^{\text{obs}}$  and  $\mathbf{x}^{\text{mis}}$ . Under the assumption that conditional independence and invariance hold as before and that the data are MAR, the joint distribution can be written as

$$P(y, \mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}}, \mathbf{r}, \boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\xi}, \boldsymbol{\psi}) = P(\mathbf{r}|y, \mathbf{x}^{\text{obs}}; \boldsymbol{\xi})P(y|\boldsymbol{\theta}; \boldsymbol{\xi})P(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}}, \boldsymbol{\theta}; \boldsymbol{\psi}), \quad (15)$$

where  $P(y|\boldsymbol{\theta}; \boldsymbol{\xi})$  is a measurement model,  $P(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}}, \boldsymbol{\theta}; \boldsymbol{\psi})$  is a population model describing the (hypothetical) relations between  $\boldsymbol{\theta}$  and both  $\mathbf{x}^{\text{obs}}$  and  $\mathbf{x}^{\text{mis}}$ , and  $P(\mathbf{r}|y, \mathbf{x}^{\text{obs}}; \boldsymbol{\xi})$  is the missing data mechanism (i.e., MAR).

In order to draw inferences on the basis of only the observed data, the missing data can be “integrated out” of the function used to fit the model of interest (e.g., the likelihood function or the posterior distribution; see Little & Rubin, 2002). The following posterior distribution for the missing data and the latent proficiency scores can then be used to draw PVs for  $\boldsymbol{\theta}$  and imputations for  $\mathbf{x}^{\text{mis}}$ :

$$P(\boldsymbol{\theta}_i, \mathbf{x}_i^{\text{mis}}|y_i, \mathbf{x}_i^{\text{obs}}; \boldsymbol{\xi}, \boldsymbol{\psi}) \propto P(y_i|\boldsymbol{\theta}_i; \boldsymbol{\xi})P(\boldsymbol{\theta}_i, \mathbf{x}_i^{\text{mis}}|\mathbf{x}_i^{\text{obs}}; \boldsymbol{\psi}). \quad (16)$$

Finally, inferences about population parameters can be drawn by averaging over the PVs for  $\boldsymbol{\theta}$  and the imputations and for  $\mathbf{x}^{\text{mis}}$  generated in this manner (see also Tanner & Wong, 1987). In the following, we describe a general strategy for sampling from the joint distribution of  $\boldsymbol{\theta}$  and  $\mathbf{x}^{\text{mis}}$ .

### Sampling From the Joint Distribution

There are several computational approaches to sampling from the joint distribution of  $\mathbf{x}^{\text{mis}}$  and  $\boldsymbol{\theta}$ . In this article, we outline a fully conditional specification (FCS) algorithm that can be used to draw from the joint distribution using a sequence of conditional models (Raghunathan et al., 2001; van Buuren et al., 2006; for a similar approach, see also Assmann et al., 2015). Other approaches are possible and are considered in the Discussion section.

*FCS algorithm.* The general idea of FCS is to approximate draws from the joint posterior distribution of  $\boldsymbol{\theta}$  and  $\mathbf{x}^{\text{mis}}$  by iterating along a sequence of conditional models and generating imputations for  $\boldsymbol{\theta}$  and  $\mathbf{x}^{\text{mis}}$  in a step-by-step manner, that is, by iterating back and forth between sampling steps pertaining to  $\boldsymbol{\theta}$  and  $\mathbf{x}^{\text{mis}}$ . This procedure is very flexible because it can employ different imputation models for each variable with missing data. The sampling algorithm in the FCS approach can be summarized as follows. Let  $\boldsymbol{\zeta}_p = (\boldsymbol{\gamma}_p, \sigma_p^2)$  denote the parameters of the latent regression model for the  $p$ th latent proficiency  $\theta_p$  ( $p = 1, \dots, P$ ), and suppose that there is a measurement model for each proficiency that induces an individual likelihood function about  $\theta_{ip}$ , given the item responses  $\mathbf{y}_i$  and the item parameters  $\boldsymbol{\xi}_p$ . Further, let  $\boldsymbol{\psi}_q$  denote the parameters for the imputation model of the  $q$ th background variable with missing data ( $q = 1, \dots, Q$ ). Then, at iteration  $t$ ,

1. For every latent proficiency  $\theta_p$  ( $p = 1, \dots, P$ ),
  - (a) Draw  $\boldsymbol{\zeta}_p^{(t+1)} \sim P(\boldsymbol{\zeta}_p | \mathbf{y}_i, \mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis},(t)}, \boldsymbol{\theta}_i^{(t)})$  from appropriate posterior distributions (or approximations thereof).
  - (b) Impute  $\theta_{ip}^{(t+1)} \sim P(\theta_{ip} | \mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis},(t)}, \boldsymbol{\theta}_{i(-p)}^{(t)}, \boldsymbol{\zeta}_p^{(t+1)})$  from the posterior predictive distribution of  $\theta_{ip}$ , given (i) the observed data, (ii) the most recent imputations of the other proficiency variables  $\boldsymbol{\theta}_{i(-p)}$  and the background variables  $\mathbf{x}_i^{\text{mis}}$ , and (iii) the parameters  $\boldsymbol{\zeta}_p$ .
2. For every background variable  $x_q$  ( $q = 1, \dots, Q$ ),
  - (a) Draw  $\boldsymbol{\psi}_q^{(t+1)} \sim P(\boldsymbol{\psi}_q | \mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis},(t)}, \boldsymbol{\theta}_i^{(t+1)})$  from appropriate posterior distributions.
  - (b) Impute  $x_{iq}^{\text{mis},(t+1)} \sim P(x_{iq}^{\text{mis}} | \mathbf{x}_i^{\text{obs}}, \mathbf{x}_{i(-q)}^{\text{mis},(t+1)}, \boldsymbol{\theta}_i^{(t+1)}, \boldsymbol{\psi}_q^{(t+1)})$  from the posterior predictive distribution of  $x_{iq}^{\text{mis}}$ , given (i) the observed data, (ii) the most recent imputations of  $\boldsymbol{\theta}_i$  and the other background variables  $\mathbf{x}_{i(-q)}^{\text{mis}}$ , and (iii) the parameters  $\boldsymbol{\psi}_q$ .

Notice that the FCS algorithm divides the sampling of the multidimensional proficiency  $\boldsymbol{\theta}$  across separate univariate models. Notice also that the sampling step for drawing the imputations for  $\mathbf{x}_i^{\text{mis}}$  conditions only on  $\boldsymbol{\theta}_i$  but not on  $\mathbf{y}_i$ . This step follows the conditional independence assumption in the scaling model (Equation 1) but can be modified to include  $\mathbf{y}_i$  if needed. For computational convenience, it is possible to approximate the posterior distribution of  $\theta_{ip}$  with a normal distribution and to use a measurement model with fixed item parameters (e.g., as estimated in an earlier step; see also Thomas & Gan, 1997; von Davier et al., 2006). Posterior draws for the model parameters can be obtained most easily by employing conjugate prior distributions and combining them with their MLEs (e.g., using the EM algorithm). This approach is implemented in the miceadds package (Robitzsch et al., 2019) for the statistical software R. The

imputation models for missing data in background variables may be chosen from among any univariate (or multivariate) models suited for the data (e.g., normal distribution, predictive mean matching; see Schafer, 1997; van Buuren et al., 2006). In the following, we present the results of two simulation studies that we conducted to evaluate these procedures.

### Study 1

In the first study, we focused on a minimal setting, which included a 2PL IRT measurement model and two continuous background variables, one of which contained missing data.

#### Data Generation

The data were simulated from a multivariate normal distribution with three variables  $\theta$ ,  $x_1$ , and  $x_2$ , where  $x_1$  and  $x_2$  are two background variables, and  $\theta$  represents the latent proficiency that is measured by a number of dichotomous indicators  $y_j$ ). For person  $i$ ,

$$(\theta_i, x_{1i}, x_{2i})^T \sim N(\mathbf{0}, \mathbf{\Sigma}), \quad (17)$$

where  $\mathbf{\Sigma}$  is the population correlation matrix. The measurement model was a 2PL IRT model defined as follows. For person  $i$  and item  $j$ ,

$$P(y_{ij} = 1 | \theta_i; \xi_j) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))}, \quad (18)$$

where the item parameters  $\xi_j = (a_j, b_j)$  were chosen in such a way that the item difficulties were equidistant in the interval  $[-2, 2]$ , the item discrimination parameters were equidistant in the interval  $[0.5, 1.5]$ , and the item difficulty and discrimination were uncorrelated for any given set of items.<sup>4</sup>

*Missing data.* Missing data were induced in  $x_1$  on the basis of a latent response propensity  $r^*$ . Specifically, we used the following linear model to simulate  $r^*$  given the values of some predictor variable  $z$ ,

$$r_i^* = \alpha + \delta z_i + e_i, \quad e_i \sim N(0, 1 - \delta^2), \quad (19)$$

where  $\alpha$  is a quantile of the standard normal distribution according to the probability of missing data (e.g.,  $\alpha = -0.84$  for 20% missing data), and  $\delta$  denotes the effect of  $z$  on missingness in  $x_1$ . A value  $x_{1i}$  was deleted if  $r_i^* > 0$ . We varied the missing data mechanism by (a) changing which variable predicted  $r^*$  (i.e., which variable was represented by  $z$ ) and (b) choosing  $\delta$  in such a way that the predictor explained more or less of the variance in  $r^*$ .

The simulated conditions are summarized in Table 3 and were chosen so that they reflected typical conditions in LSAs or were of theoretical interest.

TABLE 3.  
*Simulated Conditions in Studies 1 and 2*

Design Variables	Study 1	Study 2
Sample size ( $n$ )	100, 500, 2,000	2,000
Correlation of $\theta$ , $x_1$ , and $x_2$	$\begin{bmatrix} 1 & & \\ .50 & 1 & \\ .35 & .30 & 1 \end{bmatrix}$ , $\begin{bmatrix} 1 & & \\ .50 & 1 & \\ .70 & .30 & 1 \end{bmatrix}$	complex
Number of items ( $J$ )	7, 15, 60	16, 32
Missing mechanism	MCAR, MAR <sub><math>x</math></sub> , MAR <sub><math>y</math></sub>	MCAR, MAR <sub><math>x</math></sub> , MAR <sub><math>y</math></sub>
Strength of MAR <sup>a</sup>	50%, 100%	50%, 100%
Probability of missing data	20%, 40%	33%
No. of conditions	180	10

<sup>a</sup> The strength of the missing data mechanism denotes the amount of variance explained in the latent response propensity of  $x_2$  when the data are MAR.

Specifically, we varied the sample size ( $n = 100, 500, 2,000$ ) to reflect applications with small, moderate, or large samples (e.g., subgroups vs. countries); the number of indicators  $y_j$  ( $J = 7, 15, 60$ ) to mimic conditions in which the test had low, moderate, or high reliability (e.g., facets vs. general domains); and the correlation between the variables to reflect conditions with more or less information about  $\theta$  in the background variables. The data in  $x_1$  were either MCAR, MAR as a function of  $x_2$  (MAR <sub>$x$</sub> ), or MAR as a function of the WLE of  $\theta$  (MAR <sub>$y$</sub> ; for a similar approach, see Cham et al., 2017). Both MAR <sub>$x$</sub>  and MAR <sub>$y$</sub>  were simulated in such a way that either 50% or 100% of the variance in the response propensity was explained by  $x_2$  or the WLE of  $\theta$ , respectively. Finally, we varied the probability of missing data (20% or 40%). This resulted in a total of 180 conditions, each replicated 1,000 times.

*Scaling and imputation procedures.* In the scaling model for  $\theta$ , the item parameters were fixed to their true values. The population model was a regression of  $\theta$  on  $x_1$  and  $x_2$  as suggested by the data generating model. The procedures used to treat missing data are summarized in Table 4 and included MIM-PE, MIM-LD, and MIM-MI as outlined above as well as the FCS approach. For the purpose of comparison, we also considered PVs generated from the complete data (CD), that is, without missing data in  $x_1$ . For CD, FCS, and the MIM, we generated 20 PVs. For FCS, this also provided 20 imputations of the missing data. For MIM-MI, we generated 10 imputations of the missing data for every PV generated under the MIM, resulting in a total of 200 imputed data sets. All procedures were implemented in the statistical software R, using the packages mice (van Buuren &

TABLE 4.  
*Scaling and Imputation Procedures in Study 1*

Procedure	Description	Imputation model	$M$
CD	PV imputation of $\theta$ (with complete data in $x_1$ )	$\theta \sim y, x_1, x_2$	20
FCS	Fully conditional specification	$\theta \sim y, x_1, x_2$ $x_1 \sim x_2, \theta$	20
MIM-PE	PV imputation of $\theta$ (MIM) with pairwise estimation for missing data in $x_1$	$\theta \sim y, x_1^*, x_2, r$	20
MIM-LD	PV imputation of $\theta$ (MIM) with listwise deletion for missing data in $x_1$	$\theta \sim y, x_1^*, x_2, r$	20
MIM-MI	PV imputation of $\theta$ (MIM) with nested MI for missing data in $x_1$	$\theta \sim y, x_1^*, x_2, r$ (Stage 1) $x_1 \sim x_2, \theta$ (Stage 2)	$20 \times 10$

*Note.*  $M$  = number of imputations (nested imputations are denoted by “ $\times$ ”).

Groothuis-Oudshoorn, 2011), miceadds (Robitzsch et al., 2019), and TAM (Robitzsch et al., 2018).

*Parameters of interest and pooling.* The parameters of interest included the means, variances, and covariances for all variables as well as their correlations and the regression coefficients for the regression of  $\theta$  on  $x_1$  and  $x_2$ . For CD, FCS, and MIM-MI, the parameter estimates were obtained on the basis of the PVs for  $\theta$  and the complete or imputed background data. For MIM-PE and MIM-LD, they were obtained on the basis of the PVs for  $\theta$  and the incomplete background data by using pairwise estimation and listwise deletion, respectively. The accuracy of the parameter estimates was evaluated in terms of bias and the root mean squared error (RMSE). In addition, we calculated the coverage rate of the 95% confidence interval to assess the accuracy of the statistical inferences. The standard errors for MIM-PE were obtained through a grouped jackknife procedure with 50 jackknife samples. The standard errors for the correlation coefficients were obtained by applying the Fisher transformation (Fisher, 1915). Results obtained from the CD, FCS, MIM-PE, and MIM-LD were pooled with Rubin’s (1987) rules; those obtained from MIM-MI were pooled by applying pooling rules for nested MI (Rubin, 2003; Shen, 2000).

## Results

Because the simulation yielded many results, we focus on only the main findings here.

*Bias.* Figure 1 shows the estimated bias for the mean and variance of  $\theta$  for conditions with different sample sizes and missing data mechanisms. The estimates of the overall mean and variance of  $\theta$  were approximately unbiased for all

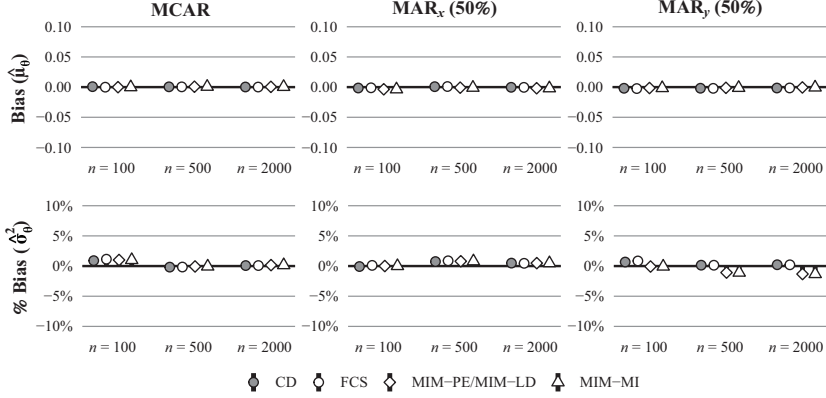


FIGURE 1. Bias (raw and in %) for the estimated mean and variance of  $\theta$  in conditions with a moderate number of items ( $J = 15$ ), a strong correlation between  $\theta$  and  $x_2$  ( $\rho_{\theta x_2} = .70$ ), and 40% missing data. CD = complete data (only PVs); FCS = fully conditional specification; MIM-PE/MIM-LD = missing indicator method with pairwise estimation or listwise deletion (point estimates are identical); MIM-MI = MIM with nested MI.

procedures regardless of the sample size and the missing data mechanism. Similarly, Figure 2 shows the bias for the mean and variance of  $x_1$  as well as the covariances between the variables. The estimates of these parameters were approximately unbiased for CD and FCS. By contrast, the results for MIM-PE, MIM-LD, and MIM-MI were more mixed. Specifically, MIM-PE and MIM-LD provided unbiased estimates only under MCAR but not under  $MAR_x$  and  $MAR_y$ . MIM-MI provided less biased estimates under  $MAR_x$ , but some bias remained for the covariance of  $x_1$  with  $\theta$  and (to a lesser extent) with  $x_2$ . In addition, MIM-MI yielded slightly biased results under MCAR, which was not the case for MIM-PE and MIM-LD. Under  $MAR_y$ , MIM-PE, MIM-LD, and MIM-MI showed noticeable bias in these parameter estimates, where the bias was smaller with MIM-MI and pointed in the opposite direction.<sup>5</sup> The extent of the bias with MIM-PE, MIM-LD, and MIM-MI also depended on the number of items ( $J$ ) and the proportion of missing data. This effect is illustrated in Figure 3 for the bias in the covariance of  $\theta$  with  $x_1$  under  $MAR_x$ . Bias with MIM-PE and MIM-LD was larger when more data were missing (40%). MIM-MI reduced the bias under  $MAR_x$ , but the reduction was less effective in conditions with a small or moderate number of items ( $J = 7$  and  $15$ ) and more missing data (40%).

In addition to the means, variances, and covariances, we also investigated the bias in the estimates of the correlations between variables and the regression coefficients in the multiple regression of  $\theta$  on  $x_1$  and  $x_2$ . The results are summarized in Figure 4. Consistent with our expectations, CD, FCS, and MIM-PE



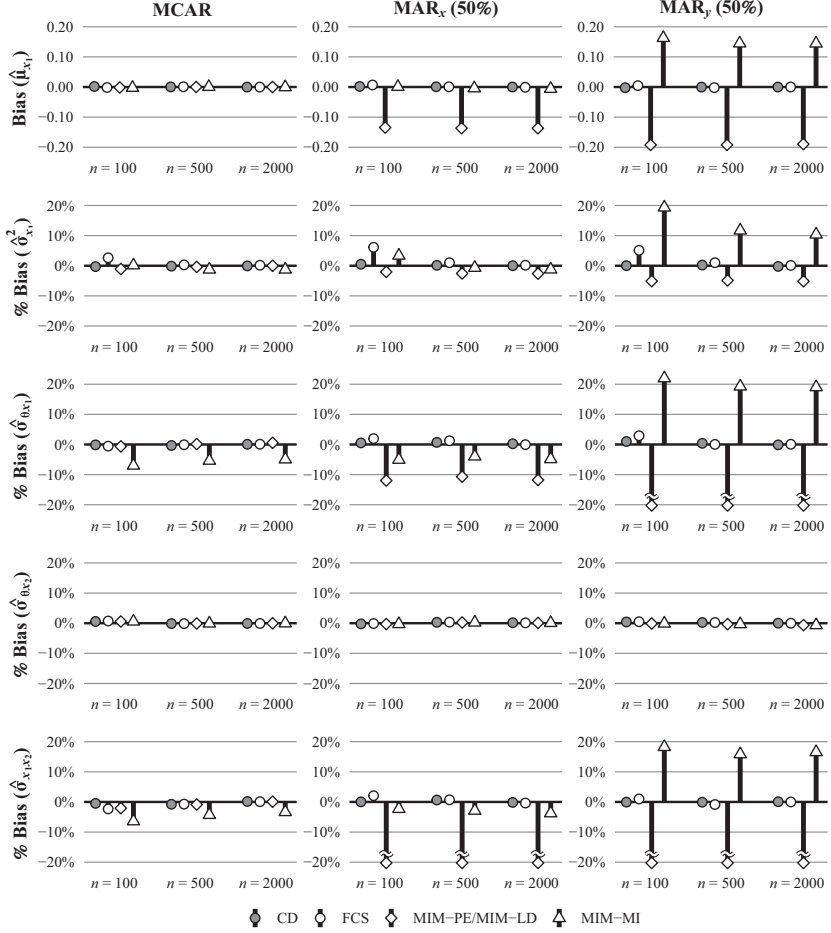


FIGURE 2. Bias (raw and in %) for the estimated the mean and variance of  $x_1$  and the covariances between  $\theta$ ,  $x_1$ , and  $x_2$  in conditions with a moderate number of items ( $J = 15$ ), a strong correlation between  $\theta$  and  $x_2$  ( $\rho_{\theta x_2} = .70$ ), and 40% missing data. The double tilde symbol ( $\approx$ ) denotes values outside the plotted range. CD = complete data (only PVs); FCS = fully conditional specification; MIM-PE/MIM-LD = missing indicator method with pairwise estimation or listwise deletion (point estimates are identical); MIM-MI = MIM with nested MI.

provided approximately unbiased estimates of the parameters. By contrast, the estimates were slightly biased with both MIM-LD and MIM-MI. This was true even when the data were MCAR (see Figure 4). The bias for the two regression coefficients pointed in opposite directions and was largest in conditions with a

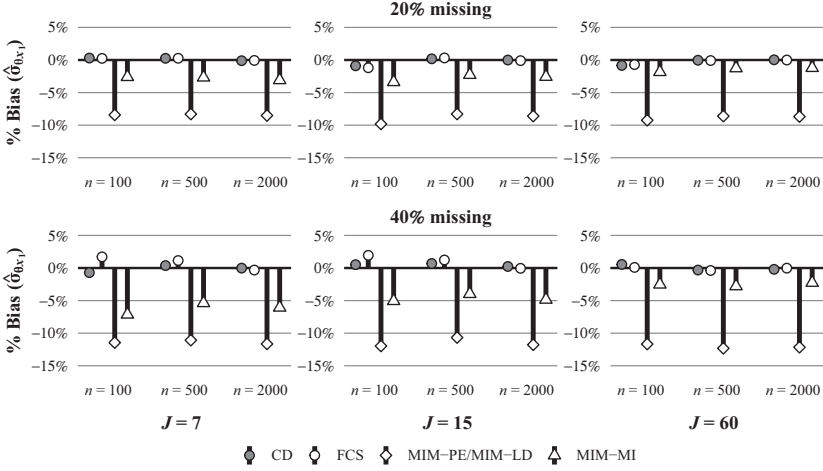


FIGURE 3. Bias (in %) for the estimated covariance between  $\theta$  and  $x_1$  in conditions with a strong correlation between  $\theta$  and  $x_2$  ( $\rho_{\theta x_1} = .70$ ) and data  $MAR_x$  (50%). CD = complete data (only PVs); FCS = fully conditional specification; MIM-PE/MIM-LD = missing indicator method with pairwise estimation or listwise deletion (point estimates are identical); MIM-MI = MIM with nested MI.

small number of items ( $J = 7$ ) and a moderate correlation between  $\theta$  and  $x_2$  ( $\rho_{\theta x_2} = .35$ ). In most of the other conditions, this bias remained relatively small (usually below 10%). Under MIM-LD, the bias was larger for the regression coefficient of  $x_2$  ( $\beta_2$ ) than for  $x_1$  ( $\beta_1$ ), that is, for the variable *not* affected by missing data. Under MIM-MI, the opposite was true.

**RMSE.** The RMSE followed a pattern similar to the bias and was usually lower under FCS than under MIM-LD, MIM-PE, and MIM-MI. However, despite the fact that the parameter estimates were sometimes biased under MIM-LD and MIM-MI, they were sometimes more accurate (i.e., had a lower RMSE) than under FCS. This was the case primarily in small samples ( $n = 100$ ) and some conditions with moderate samples ( $n = 500$ , only for MIM-MI) when the data were MCAR or  $MAR_x$  (50%). This finding may be explained by the fact that the bias for MIM-LD and MIM-MI was often negative in the respective conditions (i.e., toward zero), which may have reduced the variability in the parameter estimates. By contrast, the RMSE of the estimates under MIM-PE were sometimes significantly larger than for the other methods when the data were  $MAR_x$  and  $MAR_y$ .

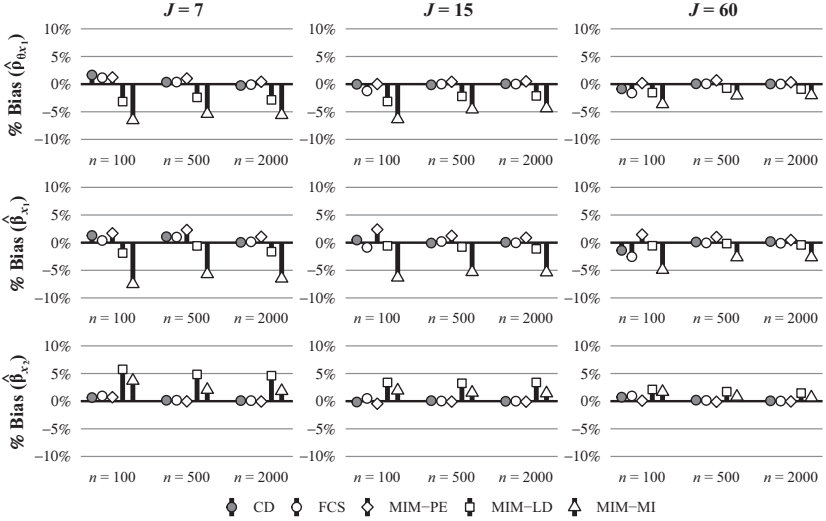


FIGURE 4. Bias (in %) for the estimated regression coefficients in the regression of  $\theta$  on  $x_1$  and  $x_2$  in conditions with a strong correlation between  $\theta$  and  $x_2$  ( $\rho_{\theta x_2} = .70$ ) and 40% missing data (MCAR). CD = complete data (only PVs); FCS = fully conditional specification; MIM-PE = missing indicator method with pairwise estimation; MIM-LD = MIM with listwise deletion; MIM-MI = MIM with nested MI.

**Coverage.** The results for the coverage rates of the 95% confidence intervals are shown in Table 5 for the estimated correlation between  $\theta$  and  $x_1$  in selected conditions with data that were MCAR. The coverage rates under FCS were usually close to those obtained with the complete data and to the nominal value of 95%. Under MIM-LD and MIM-MI, the coverage rates sometimes dropped well below the nominal value of 95%, especially in conditions with moderate or large samples ( $n = 500$  and  $2,000$ ) and a small number of items ( $J = 7$  and  $15$ ). This can be regarded as a direct consequence of bias in the parameter estimates. Under MIM-PE, the coverage rates were often too low in conditions with smaller samples ( $n = 100$  and  $500$ ) and a lot of missing data (40%). This indicates that the standard errors under MIM-PE obtained from the grouped jackknife procedure were sometimes too small.

### Summary

The results of Study 1 illustrate several important points. First, the application of the MIM allows for unbiased estimates of the mean and variance of  $\theta$  regardless of the missing data mechanism. Second, the MIM allows for unbiased estimates of means, variances, and covariances when the data are MCAR. For

more complex parameters, such as correlation and regression coefficients, estimates based on the MIM are unbiased when combined with a pairwise estimation approach (MIM-PE) but not when combined with listwise deletion (MIM-LD) or MI (MIM-MI) even when the data are MCAR. Third, parameter estimates based on the MIM can be biased when the data are MAR (i.e.,  $MAR_x$  or  $MAR_y$ ), in which case using MI (MIM-MI) may reduce bias. Finally, the FCS approach led to unbiased and efficient parameter estimates in most conditions with good inferential properties overall. These findings were in line with our expectations, which were based on the theoretical properties of the MIM. However, it is important to take into account that, in Study 1, we considered only the minimal case with two background variables. For this reason, we conducted a second study with a more complex data structure that bears a closer resemblance to educational LSAs. The main purpose of this study was to investigate whether and to what extent our findings can be expected to carry over to real data, where there is often a large number of background variables with a complex correlation structure and diverse patterns of missing data, which may show compensatory effects that are not observable in cases with fewer variables.

## **Study 2**

In the second study, we extended the simulation setting to be more similar to conditions commonly found in LSAs. Because the simulation design is much more complex than in the first study, we only provide a brief overview here. For a complete description, we refer to Online Supplement D.

### *Data Generation, Procedures, and Parameters of Interest*

The second simulation featured an extended design with a latent proficiency variable  $\theta$  as well as a total of 54 categorical background variables, which we divided into three blocks,  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ , with 18 variables each. The data were generated as follows. First, we treated the variables as if they were continuous and generated them from a multivariate normal distribution. Then, we discretized the background variables using thresholds, so that we could create different types of variables with three to eight categories and varying distributions (symmetric, skewed, or bimodal), which we distributed equally across the three blocks. The measurement model for  $\theta$  was a 2PL IRT model with items assigned to persons according to a booklet design. Finally, we induced missing data in all variables in the first and the second blocks of background variables, where the variables in the first block were MCAR,  $MAR_x$ , or  $MAR_y$  as in the first study, and the variables in the second block were induced according to a three-form rotated questionnaire design (i.e., MCAR; see Graham et al., 2006). The simulated conditions were chosen in such a way that they reflected conditions found in educational LSAs (e.g., PISA) and are summarized in Table 3. For a complete

TABLE 5.  
*Coverage of the 95% Confidence Interval for the Correlation of  $\theta$  and  $x_1$  With a Strong Correlation of  $\theta$  and  $x_2$  ( $\rho_{\theta x_2} = .70$ ) and Different Proportions of Missing Data (MCAR)*

		Small Samples ( $n = 100$ )						Moderate Samples ( $n = 500$ )						Large Samples ( $n = 2,000$ )					
$J$	MD	CD	FCS	MIM- PE	MIM- LD	MIM- MI		CD	FCS	MIM- PE	MIM- LD	MIM- MI		CD	FCS	MIM- PE	MIM- LD	MIM- MI	
7	20%	<b>91.4</b>	92.7	<b>90.8</b>	93.4	<b>92.5</b>		<b>90.4</b>	<b>90.7</b>	<b>89.4</b>	<b>92.2</b>	<b>90.8</b>		<b>91.0</b>	<b>92.5</b>	<b>90.9</b>	<b>92.4</b>	<b>89.2</b>	
	40%	<b>89.9</b>	<b>91.7</b>	<b>92.2</b>	96.9	94.6		<b>90.7</b>	92.8	<b>90.8</b>	95.4	93.1		<b>89.1</b>	92.7	<b>90.0</b>	92.9	<b>83.0</b>	
15	20%	94.2	94.6	93.9	96.9	95.6		93.3	93.5	<b>91.7</b>	94.2	93.8		93.9	94.4	92.6	94.1	<b>91.7</b>	
	40%	92.8	94.7	<b>92.1</b>	96.8	95.7		93.2	95.1	93.3	96.0	94.8		96.1	94.7	92.6	96.3	<b>89.1</b>	
60	20%	94.3	95.4	94.7	96.8	95.8		94.1	94.0	92.9	94.7	94.6		94.5	94.7	93.1	95.5	94.9	
	40%	93.6	94.1	<b>90.2</b>	95.9	94.3		94.7	96.3	93.6	97.4	96.3		94.9	94.8	93.6	97.0	93.9	

*Note.* Coverage rates below 92.5% are printed in bold.  $n$  = sample size;  $J$  = number of items; MD = proportion of missing data; CD = complete data (only PVs); FCS = fully conditional specification; MIM = missing indicator method; MIM-MI = MIM with nested MI.

description of the simulation design and the simulated conditions, we refer to Online Supplement D.

The scaling and imputation procedures were the same as in the previous study (i.e., CD, FCS, MIM-PE, MIM-LD, and MIM-MI). In accordance with the operational practices applied in LSAs, we converted the background variables into contrast codes and used principal components analysis (PCA) to extract 100 principal components, which were then used as conditioning variables in the latent regression model. Under the MIM, the effect coding also included additional codes to indicate missing responses for each variable (i.e., treating missing responses as an “extra category”). For FCS and MIM-MI, we imputed missing data in the background variables using predictive mean matching (PMM).<sup>6</sup> For CD, FCS, and the MIM, we generated 10 PVs, and for MIM-MI, we generated five imputations for every PV, resulting in a total of 50 imputations.

The parameters of interest were the means, variances, and covariances among the variables as well as the correlations between the variables and the regression coefficients for the multiple regression of  $\theta$  on pairs of any two background variables. Because the total number of parameters was too large to consider them all, we only estimated a subset of the parameters that pertained to five of the 18 background variables in each block (i.e., one per type and block). The parameter estimates were evaluated in terms of bias, RMSE, and the coverage of the 95% confidence interval. However, in contrast to the previous study, we used the average estimates obtained with CD as a point of reference in the calculation of these criteria because the original population values could not be applied to the discretized background variables. In the interest of space, we focus on the bias in the presentation of the results.

## *Results*

The estimated bias in the means, variances, and covariances of the variables is summarized in Table 6 for each type of parameter that pertains to background variables in different blocks of the background questionnaire ( $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ ). In contrast to the first study, all procedures provided estimates of the means, variances, and covariances with little or no bias as long as the data were MCAR. Under  $\text{MAR}_{\mathbf{x}}$ , MIM-PE and MIM-LD led to noticeable bias in parameter estimates pertaining to variables in the first block of the background questionnaire (i.e.,  $\mathbf{x}_1$ ). By contrast, MIM-MI yielded parameter estimates with only a little bias, that is, with the median bias near zero for all types of parameters and only a little bias remaining in the parameters. Likewise, FCS provided parameter estimates with little or no bias as in the first study. Under  $\text{MAR}_{\mathbf{y}}$ , MIM-PE and MIM-LD again led to noticeable bias in some parameter estimates. In contrast to the first study, this included the estimates for the variance of  $\theta$ . MIM-MI reduced the bias in most parameters with the exception of the bias in the variance of  $\theta$ . By contrast, FCS provided estimates with little or no bias for all parameters.

TABLE 6.

Summary of the Estimated Bias in Estimated Means, Variances, and Covariances by Block ( $\theta$ ,  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ ,  $\mathbf{x}_3$ ) in Conditions With High Reliability ( $J = 32$ )

Par.	FCS				MIM-PE/MIM-LD				MIM-MI							
	Md.	Min.	25%	75%	Max.	Md.	Min.	25%	75%	Max.	Md.	Min.	25%	75%	Max.	
MCAR																
$\mu_0$	0.1	—	—	—	—	0.1	—	—	—	—	0.1	—	—	—	—	—
$\mu_{x_1}$	-0.0	-0.0	-0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.1	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
$\mu_{x_2}$	-0.0	-0.1	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.1	-0.0	-0.0	-0.0	0.0
$\sigma_0^2$	-0.1	—	—	—	—	-0.1	—	—	—	—	-0.1	—	—	—	—	—
$\sigma_{x_1}^2$	0.3	0.1	0.3	0.4	0.5	0.0	-0.0	0.0	0.0	0.0	0.3	0.1	0.3	0.4	0.5	0.5
$\sigma_{x_2}^2$	0.4	0.0	0.3	0.4	0.4	-0.1	-0.1	-0.1	-0.0	0.1	0.4	0.0	0.3	0.4	0.4	0.4
$\sigma_{0x_1}$	-0.6	-0.9	-0.8	-0.1	-0.1	-0.2	-0.8	-0.4	-0.1	-0.1	-1.1	-1.3	-1.2	-0.9	-0.6	-0.6
$\sigma_{0x_2}$	-1.5	-2.0	-1.8	-1.0	-0.8	-0.8	-1.3	-1.1	0.3	0.3	-1.3	-1.8	-1.8	-0.8	-0.4	-0.4
$\sigma_{0x_3}$	-0.3	-0.5	-0.3	-0.2	0.0	-0.3	-0.6	-0.3	-0.2	0.1	-0.3	-0.6	-0.3	-0.2	0.1	0.1
$\sigma_{x_1x_1}$	-0.3	-0.5	-0.4	-0.0	0.8	0.1	-0.2	-0.1	0.2	0.2	-0.3	-0.5	-0.4	-0.0	0.8	0.8
$\sigma_{x_1x_2}$	-0.7	-1.8	-1.1	-0.3	0.1	-0.6	-1.5	-0.8	-0.1	0.6	-0.6	-1.7	-1.2	-0.4	0.2	0.2
$\sigma_{x_1x_3}$	-0.2	-1.2	-0.4	0.1	0.5	-0.1	-1.2	-0.3	0.1	0.4	-0.2	-1.1	-0.4	0.1	0.5	0.5
$\sigma_{x_2x_2}$	-0.3	-0.6	-0.4	-0.3	0.3	-0.1	-0.2	-0.2	-0.0	0.1	-0.4	-0.6	-0.4	-0.3	0.2	0.2
$\sigma_{x_2x_3}$	-0.3	-1.1	-0.6	-0.1	-0.0	-0.7	-1.3	-0.8	-0.3	0.4	-0.4	-1.1	-0.5	-0.1	0.1	0.1
MAR <sub>x</sub> (50%)																
$\mu_0$	0.1	—	—	—	—	0.0	—	—	—	—	0.0	—	—	—	—	—
$\mu_{x_1}$	0.1	-0.4	-0.4	0.1	0.1	-4.2	-6.1	-4.9	-3.9	-3.6	0.1	-0.4	-0.4	0.1	0.1	0.1
$\mu_{x_2}$	-0.0	-0.1	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.1	-0.0	-0.0	0.0	0.0
$\sigma_0^2$	-0.2	—	—	—	—	-0.3	—	—	—	—	-0.3	—	—	—	—	—
$\sigma_{x_1}^2$	0.3	0.2	0.2	0.6	0.7	-0.6	-0.8	-0.6	1.7	2.1	0.3	0.2	0.2	0.6	0.7	0.7
$\sigma_{x_2}^2$	0.3	0.1	0.2	0.3	0.4	-0.0	-0.1	-0.1	0.0	0.1	0.3	0.1	0.2	0.3	0.4	0.4
$\sigma_{0x_1}$	-0.0	-4.1	-3.7	0.7	0.9	-12.4	-13.2	-13.0	-12.0	-11.6	-2.3	-6.1	-5.8	-1.6	-1.4	-1.4
$\sigma_{0x_2}$	-1.1	-1.6	-1.5	-1.1	-0.2	-0.2	-0.7	-0.6	0.3	1.0	-1.1	-1.6	-1.5	-1.0	0.0	0.0

(continued)

TABLE 6. (continued)

Par.	FCS				MIM-PE/MIM-LD				MIM-MI						
	Md.	Min.	25%	75%	Max.	Md.	Min.	25%	75%	Max.	Md.	Min.	25%	75%	Max.
$\sigma_{\theta_3}$	-0.3	-0.5	-0.4	-0.2	-0.0	-0.7	-0.8	-0.8	-0.6	-0.4	-0.7	-0.8	-0.8	-0.6	-0.4
$\sigma_{\lambda_1 x_1}$	0.4	-0.1	0.1	0.5	1.0	-0.5	-2.5	-1.8	-0.2	1.8	0.3	-0.1	0.0	0.4	1.0
$\sigma_{\lambda_1 x_2}$	-0.9	-3.1	-1.8	-0.1	0.6	-5.1	-6.4	-5.8	-4.2	-2.8	-0.8	-2.9	-1.6	-0.1	0.7
$\sigma_{\lambda_1 x_3}$	0.5	-8.1	-6.4	1.3	1.6	-23.6	-28.4	-26.9	-21.5	-5.3	0.5	-7.9	-6.2	1.4	1.9
$\sigma_{\lambda_2 x_2}$	-0.4	-0.6	-0.5	-0.3	-0.1	-0.0	-0.2	-0.1	-0.0	0.1	-0.4	-0.6	-0.4	-0.3	-0.0
$\sigma_{\lambda_2 x_3}$	-0.1	-1.1	-0.3	0.1	0.4	0.1	-0.6	-0.0	0.2	0.5	-0.3	-1.0	-0.4	-0.1	0.3
MAR <sub>y</sub> (50%)															
$\mu_0$	0.1	—	—	—	—	2.2	—	—	—	—	2.2	—	—	—	—
$\mu_{x_1}$	-0.0	-0.2	-0.2	-0.0	0.0	-5.2	-7.2	-6.2	-4.8	-4.2	0.0	-0.2	-0.1	0.2	0.3
$\mu_{x_2}^2$	-0.0	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.1	-0.0	-0.0	-0.0
$\sigma_0^2$	-0.2	—	—	—	—	19.0	—	—	—	—	19.0	—	—	—	—
$\sigma_1^2$	0.2	0.1	0.2	0.4	0.6	-1.0	-1.2	-1.1	1.7	1.9	0.2	0.1	0.1	0.4	0.8
$\sigma_{x_2}$	0.3	0.1	0.2	0.4	0.5	-0.1	-0.1	-0.1	-0.0	-0.0	0.3	0.1	0.2	0.4	0.5
$\sigma_{\theta_1}$	-0.9	-3.7	-3.3	-0.5	-0.2	-26.9	-28.2	-27.0	-26.6	-26.4	0.9	-2.0	-1.1	2.6	4.0
$\sigma_{\theta_2}$	-1.3	-2.0	-1.7	-1.2	-0.5	-0.1	-0.5	-0.5	-0.0	0.3	-1.1	-1.3	-1.2	-0.5	-0.5
$\sigma_{\theta_3}$	-0.3	-0.6	-0.4	-0.2	-0.0	0.2	-0.3	-0.1	0.4	0.5	0.2	-0.3	-0.1	0.4	0.5
$\sigma_{\lambda_1 x_1}$	-0.2	-0.6	-0.3	-0.1	0.3	-1.5	-4.0	-3.4	-1.2	1.4	-0.0	-0.6	-0.2	0.0	0.3
$\sigma_{\lambda_1 x_2}$	-1.0	-2.6	-1.7	-0.8	-0.0	-9.3	-11.2	-10.6	-9.0	-7.4	-0.5	-2.3	-1.1	-0.0	0.6
$\sigma_{\lambda_1 x_3}$	-0.1	-4.8	-3.7	0.5	0.9	-19.5	-22.2	-20.7	-17.2	-5.3	0.9	-3.6	-2.8	1.8	2.4
$\sigma_{\lambda_2 x_2}$	-0.3	-0.5	-0.4	-0.3	0.2	-0.1	-0.2	-0.1	-0.1	-0.0	-0.3	-0.5	-0.4	-0.2	0.2
$\sigma_{\lambda_2 x_3}$	-0.2	-1.1	-0.4	0.0	0.4	-0.0	-0.7	-0.2	0.1	0.7	-0.2	-1.2	-0.4	-0.0	0.6

Note. Biases larger than  $\pm 5\%$  are printed in bold. The bias given for  $\mu_0$  was calculated as the absolute bias  $\times 100$ . FCS = fully conditional specification; MIM = missing indicator method with pairwise estimation or listwise deletion (point estimates are identical); MIM-MI = MIM with nested MI.



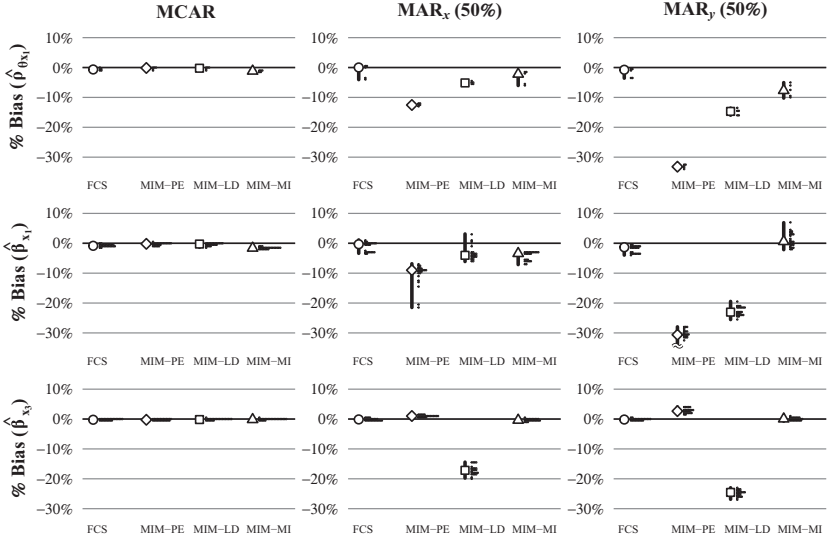


FIGURE 5. Bias (in %) for the estimated correlation coefficients of  $\theta$  with  $\mathbf{x}_1$  and the regression coefficients in the regression of  $\theta$  on any pair of variables in  $\mathbf{x}_1$  and  $\mathbf{x}_3$  in conditions with high reliability ( $J = 32$ ). The point range and symbols indicate the range of and median bias. The double tilde symbol ( $\approx$ ) denotes values outside the plotted range. The individual points represent the bias for each individual parameter. FCS = fully conditional specification; MIM-PE = missing indicator method with pairwise estimation; MIM-LD = MIM with listwise deletion; MIM-MI = MIM with nested MI.

Similar to the first study, we also investigated bias in the correlation and regression coefficients. This is illustrated in Figure 5 for the correlations of  $\theta$  with the background variables in the first block (i.e.,  $\mathbf{x}_1$ ) and the regression coefficients for the regressions of  $\theta$  on the pairs of variables in the first and third blocks of the background questionnaire (i.e.,  $\mathbf{x}_1$  and  $\mathbf{x}_3$ ). Under MCAR, all methods led to approximately unbiased estimates of these parameters. This is in contrast to the first study, in which only MIM-PE provided unbiased estimates of these parameters under MCAR. Under both  $MAR_x$  and  $MAR_y$ , both MIM-PE and MIM-LD led to noticeable bias in the parameter estimates. MIM-MI provided a substantial reduction of the bias in most cases although some bias remained in individual parameters. By contrast, FCS provided estimates with little or no bias for these parameters in all conditions.

### Summary

The second simulation study provided some important insights about the statistical properties of the methods we considered. First, although the theoretical

properties of the methods based on the MIM (MIM-PE, MIM-LD, and MIM-MI) suggested that a naive use of MIM-LD and MIM-MI may lead to biased parameter estimates even when the data are MCAR, this may not necessarily be a reason for concern in the context of educational LSAs, where the large number of background variables may compensate for the effects of missing data under the MIM. This is an encouraging finding because it illustrates that the MIM allows for approximately unbiased parameter estimates at least under MCAR. However, under  $MAR_x$  and  $MAR_y$ , MIM-PE and MIM-LD still had the potential to provide biased parameter estimates. Second, FCS appeared to provide estimates with good statistical properties even with a larger number of background variables. The same was true for MIM-MI, which led to a substantial reduction in the bias in most cases, although some bias remained in a few individual parameters under  $MAR_x$  and  $MAR_y$ .

### **Example Analysis: PISA 2015**

To illustrate the procedures considered in this article, we used data from the German subsample ( $N = 6,504$ ) of PISA 2015 (OECD, 2017). The data included the 184 cognitive items from the *science* domain and 214 variables from the students' background questionnaire. Missing data occurred in all cognitive items (range: 77.8%–92.6%) and almost all of the background variables (range: 10.9%–54.0%). The procedures were implemented in a manner that was similar to the operational practices used in PISA 2015. In the interest of space, we only provide a brief description of the procedures here and provide a full description along with the computer code in Online Supplement E. We generated the PVs using the item parameters and contrast coding scheme used in PISA 2015 (OECD, 2017) and the R package TAM (Robitzsch et al., 2018). This entailed the use of dummy codes to indicate missing responses for all background variables, which were then subjected to PCA in order to obtain components that we used as conditioning variables in the latent regression model. To impute the missing data, we used a FCS approach based on the R package mice (van Buuren & Groothuis-Oudshoorn, 2011), where categorical and ordinal variables were imputed with polytomous regression models and PMM, respectively (see also Kaplan & Su, 2016, 2018). The imputation methods for the background variables also used the contrast-coded background data but relied on partial least squares (PLS) to reduce the dimensionality of the data. All steps in the procedure included the final student weights.

For the analysis of the data, we obtained WLEs for each of the 25 noncognitive scales from the background questionnaire using the item parameters used in PISA 2015. The parameters of interest were the regression coefficients in the regressions of science proficiency on any two of the noncognitive scales from the background questionnaire. This resulted in 600 regression coefficients (not counting the intercept) obtained from 300 regression analyses. For MIM-PE and

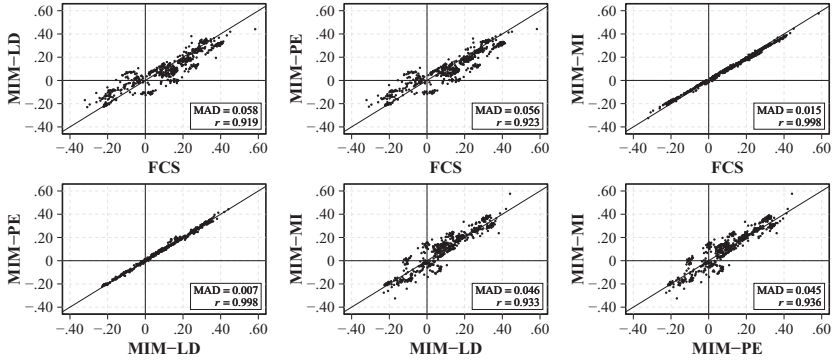


FIGURE 6. Comparison of the estimated standardized regression coefficients from the regressions of the PVs on any two background variables in the example analysis obtained from different procedures. MAD = mean absolute difference;  $r$  = correlation; FCS = fully conditional specification; MIM-PE = missing indicator method with pairwise estimation; MIM-LD = MIM with listwise deletion; MIM-MI = MIM with nested MI.

MIM-LD, the WLEs were calculated only for students who responded to at least three items of that scale and set to missing otherwise (see also OECD, 2017). The results are summarized in Figure 6, which provides a direct comparison of the parameter estimates obtained from different procedures. Overall, the regression coefficients obtained from the three procedures were similar with differences primarily occurring between MIM-PE and MIM-LD on the one hand and FCS and MIM-MI on the other hand, such that the estimates under MIM-PE and MIM-LD tend to be somewhat closer to zero. This difference is visible both in the correlation and the mean absolute difference between the procedures as well as in the slight “tilt” that is present in some of the panels of the figure. However, in summary, these results are in line with the results of the simulation studies and illustrate that, whereas differences between the procedures can sometimes be observed, the differences were relatively small in most cases, especially when there were many background variables available and the (conditional) reliability of the test was high.

## Discussion

In this study, we investigated different procedures for the treatment of missing data in background variables in educational LSAs. This included the joint treatment of proficiency scores and missing data (i.e., FCS) as well as the procedures currently used in operational practice (i.e., the MIM), combined with different methods for the treatment of missing data (MIM-PE, MIM-LD, MIM-MI). From a theoretical perspective, our investigation revealed that the MIM can provide

unbiased results when the data are MCAR. This included estimates of means, variances, and covariances, but it required the use of a pairwise estimation strategy (MIM-PE) for more complex parameters such as correlation and regression coefficients, whereas its use with listwise deletion (MIM-LD) or MI (MIM-MI) sometimes led to biased parameter estimates even under MCAR. In two simulation studies, we investigated these properties across a wider range of conditions. In all conditions, we found that the FCS procedure provided unbiased estimates with good inferential properties. By contrast, the MIM provided unbiased estimates only when the data were MCAR but not when they were MAR, in which case the bias tended to be lowest when the MIM was combined with MI (MIM-MI). However, in conditions with a larger number of background variables, which are the most similar to conditions encountered in LSAs, the differences between the procedures were much less pronounced, indicating that some of the negative properties of the MIM can be compensated for with the large amount of information available in LSAs. In particular, we found that the MIM when combined with MI often provided results similar to FCS in these conditions for all except a few of the parameters of interest.

Overall, these results are encouraging because they (a) illustrate a number of positive theoretical properties of the MIM that hold when the data are MCAR (i.e., with rotated questionnaires; see also Adams et al., 2013) and (b) provide additional support for the recommendations already in place regarding the analyses of incomplete background data with rotated questionnaires in educational LSAs (e.g., for MIM-PE and MIM-MI; see also OECD, 2014). However, the results also highlight some possible weaknesses of the MIM, namely, that (a) parameter estimates might not be completely unbiased when the data are MAR even when MI is used and (b) some parameter estimates may be biased even when the data are MCAR when obtained with listwise deletion, which is the default setting for many analyses in statistical software.

Based on our findings, it may be worth considering the joint treatment of proficiency scores and missing data as an alternative to the operational practices applied in educational LSAs. However, this idea can be met with scrutiny: Not only does a joint treatment complicate the generation of PVs, but it also blurs the line between the imputer and the (secondary) analyst by placing the burden of specifying an imputation model for the missing background data on the imputer. This requires the imputer to anticipate potential analyses, so that the imputation model will “fit” the intended analyses (Meng, 1994). However, this problem is not unique to the treatment of missing data and applies to the generation of the PVs in the same way. In principle, it may be worth considering whether the data for secondary analyses could be released in two versions: one containing PVs and imputations for missing background data and one containing only the PVs but with the imputations for missing data deleted. In this context, two-stage or nested MI (Rubin, 2003)—in which the PVs for the latent proficiency variables and the imputations for the missing background

data are generated in two separate stages—may be considered as an alternative. The little research that has evaluated these approaches seems to imply that they enjoy qualities similar to the joint imputation of PVs and missing data (Weirich et al., 2014; see also Kaplan & Su, 2018).

The implementation of the FCS approach considered here is only one of several ways to adopt a joint imputation of PVs and missing data. First, the FCS approach is not restricted to the use of univariate models to generate PVs, and multivariate models can be used instead to maintain the multidimensional scaling procedures employed in educational LSAs (see von Davier & Sinharay, 2013). To our knowledge, such an approach is currently not implemented in statistical software. Second, instead of using an FCS approach, the joint distribution of the variables can sometimes be modeled directly (Blackwell et al., 2017a, 2017b; King et al., 2001; see also Cole et al., 2006). Third, the FCS approach presented here uses parametric models for the treatment of missing data. By contrast, nonparametric methods (e.g., Assmann et al., 2015; Si & Reiter, 2013) or procedures based on latent class models may allow for an even more flexible representation of the relations (and possible interactions) between the variables (Vermunt et al., 2008; Vidotto et al., 2018; Wetzel et al., 2015; Xu & von Davier, 2019; see also von Davier, 2013). Finally, the imputation of PVs need not be restricted to the achievement test data but can also be applied to the noncognitive constructs in the background questionnaire (e.g., interest and attitude scales). In other words, missing data need not be imputed on the item level but may be combined with intermittent steps to generate PVs for these scales (see also Gottschall et al., 2012).

This study comes with multiple limitations and points to consider. First, educational LSAs often feature hundreds of variables that require the dimensionality of the data set to be reduced before the scaling model can be applied. For this purpose, PCAs or PLS can be used (Oranje et al., 2009; Oranje & Ye, 2013). This was illustrated in our analysis of the PISA 2015 data, in which we used PLS to reduce the dimensionality of the imputation models for the background variables. Second, the data in educational LSAs often have a multilevel structure that results from students being clustered in schools. This structure needs to be taken into account in the scaling of the proficiency data (Adams et al., 1997; Adams & Wu, 2007; Li et al., 2009), the imputation of missing background data (Enders et al., 2016; Lüdtke et al., 2017), and secondary analyses (Monseur & Adams, 2009). By contrast, if the multilevel structure is represented with fixed effects, such as in PISA, then the methods considered in this article can be applied directly by including an additional set of indicator variables to represent school membership in the imputation models for missing data and the PVs (see the example with the PISA 2015 data).

Further research is needed to fully understand the proper treatment of missing data in background variables in educational LSAs. This includes the comparison of the available methods in a wider range of settings, for example, when the data

are MNAR (see also Ibrahim et al., 2005; Moustaki & Knott, 2000; O’Muircheartaigh & Moustaki, 1999). The same is true for applications with rotated questionnaire designs that are gaining popularity in educational LSAs. The procedures studied in this article can be applied to missing data that stem from rotated questionnaires, but their performance under different forms of rotation is an important topic that has yet to receive more attention (see also Kaplan & Su, 2016, 2018). Finally, future research should take into account the relationship between the imputer and the (secondary) analyst (Meng, 1994). This is particularly important because the data from educational LSAs are often used for a variety of purposes, comprising an immense number of potential analyses (e.g., models with nonlinear effects, multilevel models; see also Li et al., 2009; Schofield, 2015; Schofield et al., 2015).


### **Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### **ORCID iD**

Simon Grund  <https://orcid.org/0000-0002-1290-8986>

### **Notes**

1. The procedures presented in this article can also be applied when data are missing unsystematically in the achievement test (e.g., with rotated booklet or multimatrix designs). The case in which data are missing systematically in the achievement test is a different topic and will not be considered here (for further discussion, see Moustaki & Knott, 2000; Pohl et al., 2014; Rose et al., 2017).
2. If this assumption is not satisfied, and the data are missing *not* at random, that is,  $P(\mathbf{r}|\mathbf{x}, \mathbf{y}) = P(\mathbf{r}|\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}}, \mathbf{y})$ , then the treatment of missing data is more complicated and requires the missing data mechanism to be modeled explicitly (for a detailed discussion, see Carpenter & Kenward, 2013; Little & Rubin, 2002).
3. In practice, full information maximum likelihood (FIML) estimation is also a popular alternative for dealing with missing data. We did not consider FIML further in this study because the statistical properties of parameter estimates under FIML and multiple imputation (MI) appeared to be virtually the same. For interested readers, we provide simulation results comparing the parameter estimates under the missing indicator method (MIM) when combined with FIML or MI in Online Supplement B.

4. The optimal permutations of the item difficulty and discrimination parameters that best fulfilled these criteria were determined a priori by using a large simulated data set ( $N = 20,000$ ) and comparing all possible permutations for any given number of items. The same permutations of the item parameters were used throughout the study and included in Online Supplement C.
5. The relatively large bias with MIM-PE, MIM-LD, and MIM-MI under MAR<sub>y</sub> may be explained by the fact that this mechanism implies a dependency between the item responses ( $y_1, \dots, y_J$ ) and the missing data indicator  $r$ . This violates the assumptions of the measurement model used with the MIM.
6. In principle, the imputation of missing data in background variables can also be combined with dimension reduction techniques such as principal components analysis (PCA) or partial least squares (PLS). In the present case, no dimension reduction was used because none was needed. However, in an additional simulation study, we found that PLS yielded approximately the same results as predictive mean matching without dimension reduction, whereas the results were more mixed with PCA.

## References

- Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education*, 1(5), 1–27. <http://doi.org/10.1186/2196-0739-1-5>
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47–76. <http://doi.org/10.3102/10769986022001047>
- Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 57–75). Springer.
- Allison, P. D. (2001). *Missing data*. Sage.
- Assmann, C., Gaasch, C., Pohl, S., & Carstensen, C. H. (2015). Bayesian estimation in IRT models with missing values in background variables. *Psychological Test and Assessment Modeling*, 57, 595–618. <http://www.psychologie-aktuell.com/>
- Blackwell, M., Honaker, J., & King, G. (2017a). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research*, 46, 342–369. <http://doi.org/10.1177/0049124115585360>
- Blackwell, M., Honaker, J., & King, G. (2017b). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, 46, 303–341. <http://doi.org/10.1177/0049124115585360>
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459. <http://doi.org/10.1007/BF02293801>
- Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education*, 5(17), 1–16. <http://doi.org/10.1186/s40536-017-0050-x>
- Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Wiley.

- Cham, H., Reshetnyak, E., Rosenfeld, B., & Breitbart, W. (2017). Full information maximum likelihood estimation for latent variable interactions with incomplete indicators. *Multivariate Behavioral Research*, 52, 12–30. <http://doi.org/10.1080/00273171.2016.1245600>
- Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences* (1st ed.). Erlbaum.
- Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35, 1074–1081. <http://doi.org/10.1093/ije/dyl097>
- Culpepper, S. A., & Park, T. (2017). Bayesian estimation of multivariate latent regression models: Gauss versus Laplace. *Journal of Educational and Behavioral Statistics*, 42(5), 591–616. <http://doi.org/10.3102/1076998617700598>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38. <http://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21, 222–240. <http://doi.org/10.1037/met0000063>
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507–521. <http://doi.org/10.2307/2331838>
- Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, 47, 1–25. <http://doi.org/10.1080/00273171.2012.640589>
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11, 323–343. <http://doi.org/10.1037/1082-989X.11.4.323>
- Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4, 75–89. <http://doi.org/10.1016/j.stamet.2006.03.002>
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100, 332–346. <http://doi.org/10.1198/016214504000001844>
- Johnson, M. S., & Jenkins, F. (2004). A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress. *ETS Research Report Series*, 2004(2), i–28. <http://doi.org/10.1002/j.2333-8504.2004.tb01965.x>
- Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91, 222–230. <http://doi.org/10.2307/2291399>
- Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, 41, 57–80. <http://doi.org/10.3102/1076998615622221>



- Kaplan, D., & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: A comparison of three designs. *Large-Scale Assessments in Education*, 6(6), 1–31. <http://doi.org/10.1186/s40536-018-0059-9>
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Association*, 95, 49–69. <http://doi.org/10.1017/S0003055401000235>
- Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics*, 34, 433–463. <http://doi.org/10.3102/1076998609332757>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233–245. <http://doi.org/10.1007/BF02294018>
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22, 141–165. <http://doi.org/10.1037/met0000096>
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538–558. <http://doi.org/10.1214/ss/1177010269>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177–196. <http://doi.org/10.1007/BF02294457>
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133–161.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, 17, 131–154.
- Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, 10(3), 320–334. <https://orbi.uliege.be/>
- 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 (Statistics in Society)*, 163(3), 445–459. <http://doi.org/10.1111/1467-985X.00177>
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176. <http://doi.org/10.1177/014662169201600206>
- 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 (Statistics in Society)*, 162, 177–194. <http://doi.org/10.1111/1467-985X.00129>
- Oranje, A., Staniewska, D., & Ye, L. (2009). *An exploration of model reduction approaches for educational survey population models*. Meeting of the National Council on Measurement in Education, San Diego, CA.
- Oranje, A., & Ye, L. (2013). Population model size, bias, and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 203–228). CRC Press.
- Organization for Economic Cooperation and Development. (2014). *PISA 2012 technical report*. <https://www.oecd.org/pisa/>

- Organization for Economic Cooperation and Development. (2017). *PISA 2015 technical report*. <https://www.oecd.org/pisa/>
- 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, 423–452. <http://doi.org/10.1177/0013164413504926>
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–96. <http://www.statcan.gc.ca/>
- Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, 51, 305–316. <http://doi.org/10.1080/00461520.2016.1208094>
- Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, 39, 235–256. <http://doi.org/10.3102/1076998614531045>
- Robitzsch, A., Grund, S., & Henke, T. (2019). *miceadds: Some additional multiple imputation functions, especially for mice (Version 3.6-1)*. <http://CRAN.R-project.org/package=miceadds>
- Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules (Version 2.10-11)*. <https://CRAN.R-project.org/package=TAM>
- Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, 82, 795–819. <http://doi.org/10.1007/s11336-016-9544-7>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592. <http://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3–18. <http://doi.org/10.1111/1467-9574.00217>
- Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, 48, 293–312.
- Rutkowski, L., & Zhou, Y. (2015). The impact of missing and error-prone auxiliary information on sparse-matrix sub-population parameter estimates. *Methodology*, 11, 89–99. <http://doi.org/10.1027/1614-2241/a000095>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. <http://doi.org/10.1037/1082-989X.7.2.147>
- Schofield, L. S. (2015). Correcting for measurement error in latent variables used as predictors. *The Annals of Applied Statistics*, 9, 2133–2152. <http://doi.org/10.1214/15-AOAS877>
- Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, 80, 727–747. <http://doi.org/10.1007/s11336-014-9415-z>
- Shen, Z. (2000). *Nested multiple imputation* (Unpublished doctoral dissertation), Harvard University, Cambridge, MA.
- Shin, Y. (2013). Efficient handling of predictors and outcomes having missing values. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international*

- large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 451–479). CRC Press.
- Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38, 499–521. <http://doi.org/10.3102/1076998613480394>
- Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science*, 360(6384), 38–40. <http://doi.org/10.1126/science.aar4952>
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540. <http://doi.org/10.2307/2289457>
- Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, 22, 425–445.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 1049–1064. <http://doi.org/10.1080/10629360600810434>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <http://doi.org/10.18637/jss.v045.i03>
- Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38, 369–397. <http://doi.org/10.1111/j.1467-9531.2008.00202.x>
- Vidotto, D., Vermunt, J. K., & van Deun, K. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, 43, 511–539. <http://doi.org/10.3102/1076998618769871>
- von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). CRC Press.
- von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 9–36). IEA-ETS Research Institute.
- von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*, 32, 233–251. <http://doi.org/10.3102/1076998607300422>
- von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, 35(2), 174–193. <http://doi.org/10.3102/1076998609346970>
- von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In Rutkowski, L., von Davier, M., & Rutkowski, M. (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). CRC Press.

- von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1039–1055). Elsevier.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. <http://doi.org/10.1007/BF02294627>
- Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-Scale Assessments in Education*, 2(9), 1–18. <http://doi.org/10.1186/s40536-014-0009-0>
- Wetzel, E., Xu, X., & von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement*, 75, 739–763. <http://doi.org/10.1177/0013164414558843>
- Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31, 114–128. <http://doi.org/10.1016/j.stueduc.2005.05.005>
- Xu, X., & von Davier, M. (2019). Applying the general diagnostic model to proficiency data from a national skills survey. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 489–501). Springer.

### Authors

SIMON GRUND is a researcher at the Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, 24118 Kiel, Germany; email: [grund@ipn.uni-kiel.de](mailto:grund@ipn.uni-kiel.de). His research interests include missing data, educational measurement, research synthesis, and multilevel modeling.

OLIVER LÜDTKE is a professor of educational measurement at the Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, 24118 Kiel, Germany; email: [oluedtke@ipn.uni-kiel.de](mailto:oluedtke@ipn.uni-kiel.de). His research interests include the application of multilevel modeling in psychological and educational research, international student achievement studies, and personality development in adolescence.

ALEXANDER ROBITZSCH is a principal researcher at the Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, 24118 Kiel, Germany; email: [robitzsch@ipn.uni-kiel.de](mailto:robitzsch@ipn.uni-kiel.de). His research interests include computational aspects of psychometric models, item response modeling, missing data, regularization methods, causal effects, and multilevel modeling.

Manuscript received December 21, 2018

First revision received March 27, 2020

Second revision received August 19, 2020

Accepted August 21, 2020