Multiple Imputation of Multilevel Missing Data: An Introduction to the R Package pan

Simon Grund[1,2], Oliver Lüdtke[1,2], and Alexander Robitzsch[1,2]

Abstract

The treatment of missing data can be difficult in multilevel research because state-of-the-art procedures such as multiple imputation (MI) may require advanced statistical knowledge or a high degree of familiarity with certain statistical software. In the missing data literature, pan has been recommended for MI of multilevel data. In this article, we provide an introduction to MI of multilevel missing data using the R package pan, and we discuss its possibilities and limitations in accommodating typical questions in multilevel research. To make pan more accessible to applied researchers, we make use of the mitml package, which provides a user-friendly interface to the pan package and several tools for managing and analyzing multiply imputed data sets. We illustrate the use of pan and mitml with two empirical examples that represent common applications of multilevel models, and we discuss how these procedures may be used in conjunction with other software.

Keywords

multiple imputation, missing data, multilevel, R

In recent years, multilevel models have become one of the standard tools for analyzing clustered empirical data. Such data often occur in organizational and educational psychology and other fields of the social sciences; for example, when employees are nested within work groups, students are nested within school classes, or in longitudinal studies when measurement occasions are nested within persons. In addition, empirical data are often incomplete, for example, when participants drop out of the study or do not answer all of the items on a questionnaire. Several authors have advocated the use of modern missing data techniques such as multiple imputation (MI) rather than traditional approaches such as listwise or pairwise deletion (Allison, 2001; Enders, 2010; Newman, 2014; Schafer & Graham, 2002; van Buuren, 2012). One central requirement of MI is that the imputation model must be at least as general as the model of interest to preserve relationships among variables (Enders, 2010). In the case of incomplete multilevel data, it is important that the imputation model takes the multilevel structure into account to ensure valid statistical inferences in subsequent multilevel analyses (Black, Harel, & McCoach, 2011; Graham, 2012; van Buuren, 2011).

Although MI is gaining popularity among applied researchers, multilevel imputation models are rarely used in practice. One of the most commonly recommended software solutions for multilevel imputation is the pan package (Schafer & Yucel, 2002; Schafer & Zhao, 2014), which is freely available in the statistical software R (R Core Team, 2015; see also Culpepper & Aguinis, 2011). However, the application of pan can be challenging, and its documentation is rather technical, especially for users who are not familiar with R. For instance, for multilevel missing data, Graham (2012) recommended "that you obtain a copy of the PAN program (…), and that you find an expert in R who can help you get started" (p. 137).

The present article is intended as a gentle introduction to the pan package for MI of multilevel missing data. We assume that readers have a working knowledge of multilevel models (see Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). To make pan more accessible to applied researchers, we make use of the R package mitml, which provides a user-friendly interface to the pan package and some additional tools for organizing and analyzing multiply imputed data (Grund, Robitzsch, & Lüdtke, 2016). The first section of this article introduces an empirical example that is used for illustrating the application of pan to multilevel data. In the following section, we briefly describe the main ideas

[1]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]Centre for International Student Assessment, Germany

Corresponding Author:
Simon Grund, Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, Kiel 24118, Germany.
Email: grund@ipn.uni-kiel.de

Table 1. Pairwise Observed-Data Correlations Among Variables and Amount of Missing Data.

	MA	RA	CA	SES	DPM	DPR	SC
MA		.528	.530	.232	−.234	−.238	−.217
RA			.493	.299	−.291	−.294	−.327
CA				.240	−.265	−.251	−.221
SES					−.154	−.155	−.123
DPM						.782	.399
DPR							.419
Missing Data	19.4%	0%	0%	35.0%	61.4%	21.5%	21.7%

Note. MA = math achievement; RA = reading achievement; CA = cognitive ability; SES = socioeconomic status; DPM = disciplinary problems in math class; DPR = disciplinary problems in reading class; SC = school climate.

behind pan and MI, and we discuss which features of multilevel models must be considered when conducting MI. Finally, we use the mitml package to carry out MI for the empirical example. In that context, we will discuss possibilities for model diagnostics and tests of statistical hypotheses (e.g., model constraints, model comparisons).

Multilevel Modeling: An Empirical Example

Multilevel models account for dependencies in the data and allow relationships between variables to be estimated at different levels of analysis or effects that may vary across higher level observational units. For the purpose of this article, we assume that the multilevel structure consists of persons (e.g., students, employees) nested within groups (e.g., classes, work groups). If only the regression intercept varies across groups, the model is referred to as a random-intercept model. For example, Chen and Bliese (2002) examined the effects of individual characteristics (e.g., psychological strain) and leadership climate on the self-efficacy of U.S. soldiers. Kunter, Baumert, and Köller (2007) investigated the effects of student- and group-level ratings of classroom management on students' interest in mathematics. If the effects of additional predictor variables vary across groups, the model is referred to as a random-slope or random-coefficients model. For example, Hofmann, Morgeson, and Gerras (2003) investigated varying effects of leader-member exchange on safety behavior across work teams in the U.S. army.

The example data set used in this article is from the field of educational research and was taken from the German sample of primary school students who participated in the Progress in International Reading Literacy Study (PIRLS; Bos et al., 2005; Mullis, Martin, Gonzales, & Kennedy, 2003). The data set includes test scores in both mathematics (MA) and reading achievement (RA), a measure of cognitive ability (CA), a measure of socioeconomic status (SES), students' ratings of the quality of teaching in their math and reading classes (the prevalence of disciplinary problems), and ratings of the general learning environment (school climate). For the purpose of this article, we considered only

students for whom RA and CA scores were available, which was true for approximately 99.3% of the sample (8,767 students in 475 classes). Ratings of disciplinary problems in math classes (DPM) were missing for half of the sample due to a planned missing data design (Graham, Taylor, Olchowski, & Cumsille, 2006). Table 1 provides an overview of the data set, along with the observed correlations and the percentages of missing values among variables. Some variables contain additional, unplanned missing data. In such cases, it is useful to examine the missing data patterns that occur in the data set. This is shown in Table 2. Approximately 50% of the sample adhered to the planned missing data design (Patterns 1 and 2). In another 25% of the sample (Patterns 3 and 4), SES was additionally missing. The remaining patterns were more diverse, and data were missing for MA scores, disciplinary problems in reading classes (DPR), or school climate (SC). Planned missing data designs are becoming increasingly popular in large-scale observational studies because such designs can reduce the burden that is placed on each individual participant (Graham et al., 2006). The missing data mechanism is usually ignorable for variables recorded in this manner, thus enabling us to focus on more specific aspects of MI in multilevel research.

Example 1: Random-Intercept Model

Our first model of interest examined the effect of teaching quality in math classes (disciplinary problems; DPM) on students' MA scores. In addition, we included SES to control for differences in socioeconomic background between students and classes. The student-level variables were centered around the group mean, and the group means were included as predictor variables to separate within-group from between-group effects (see Enders & Tofighi, 2007; Raudenbush & Bryk, 2002). For student i in class j, the model reads

$$\begin{aligned} \mathrm{MA}_{ij} = {} & \beta_0 + \beta_1 \left(\mathrm{DPM}_{ij} - \overline{\mathrm{DPM}}_j \right) + \\ & \beta_2 \overline{\mathrm{DPM}}_j + \beta_3 \left(\mathrm{SES}_{ij} - \overline{\mathrm{SES}}_j \right) + \\ & \beta_4 \overline{\mathrm{SES}}_j + \upsilon_{0j} + \varepsilon_{ij} \, . \end{aligned} \quad (1)$$

Table 2. Frequent Missing Data Patterns.

Pattern	MA	RA	CA	SES	DPM	DPR	SC	Cases #	Relative %	Cumulative %
1	o	o	o	o	x	o	o	2,306	26.3	26.3
2	o	o	o	o	o	o	o	2,134	24.3	50.6
3	o	o	o	x	x	o	o	1,173	13.4	64.0
4	o	o	o	x	o	o	o	1,125	12.8	76.9
5	x	o	o	o	x	x	x	1,027	11.7	88.6
6	x	o	o	x	x	x	x	622	7.1	95.7

Note. The patterns displayed here account for ≥95% of the sample. MA = math achievement; RA = reading achievement; CA = cognitive ability; SES = socioeconomic status; DPM = disciplinary problems in math class; DPR = disciplinary problems in reading class; SC = school climate; o = observed; x = missing.

Here, the β coefficients denote fixed effects, and υ_{0j} and ε_{ij} denote the residuals at the class and student level, respectively. We refer to the effects of the average DPM and SES of a class as *between-group* effects, whereas *within-group* effects account for the students' individual deviations from that average. For example, β_4 denotes the effect of a class' average SES on class-level MA, whereas β_3 denotes the effect of students' individual deviations from the class average on their individual MA scores. The class- and student-level residuals are each assumed to be distributed normally, independently and identically, with zero mean and variances $Var(\upsilon_{0j})$ and $Var(\varepsilon_{ij})$, respectively.

Example 2: Random-Slope Model

Our second model of interest examined the relationship between students' CA and their MA scores. We assumed that the relationship between the two variables would vary across groups (random slope) because some teachers may nurture students' individual strengths and weaknesses, whereas others may strive to "equalize" them. As before, we included SES to control for differences in socioeconomic background. In line with recent recommendations for analyzing random-slope variation, we centered the variables around the group means (Aguinis, Gottfredson, & Culpepper, 2013; Hofmann & Gavin, 1998). The group means were included as additional predictors to "reintroduce" the group-level part into the model. The model reads,

$$
\begin{aligned}
\text{MA}_{ij} = \beta_0 + \beta_1\left(\text{CA}_{ij} - \overline{\text{CA}}_j\right) + \beta_2\overline{\text{CA}}_j + \\
\beta_3\left(\text{SES}_{ij} - \overline{\text{SES}}_j\right) + \beta_4\overline{\text{SES}}_j + \\
\upsilon_{0j} + \upsilon_{1j}\left(\text{CA}_{ij} - \overline{\text{CA}}_j\right) + \varepsilon_{ij},
\end{aligned}
\tag{2}
$$

where υ_{1j} denotes the random effect of CA on MA per class. The two random effects (intercept and slope) are assumed to follow a multivariate normal distribution, independently and identically across classes, and the remaining notation is the same as above.

MI of Incomplete Multilevel Data

Missing data could be addressed by restricting the analyses to completely observed cases (listwise deletion, LD). However, this approach is more likely to suffer from low power and to give biased results (e.g., Little & Rubin, 2002; see also Newman, 2014). MI has become one of the preferred methods for overcoming these problems (Rubin, 1987; Schafer & Graham, 2002). Using MI, a number of replacements for the missing data are drawn from the distribution of the missing values, given the observed data and an imputation model. The completed data sets are then analyzed separately, and the results are combined across data sets to form final parameter estimates and inferences (see Enders, 2010, for details about the general MI procedure).

General Aspects of MI

In most applications of MI, the data are assumed to be missing at random (MAR), a notion that was introduced by Rubin (1976) in his well-known classification of missing data mechanisms. Consider the hypothetical complete data matrix **Y** which is decomposed into observed and unobserved portions, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. An indicator matrix **R** denotes whether values are observed or missing. If the missing data are simply a random sample of the hypothetical complete data, that is, $P(\mathbf{R} \mid \mathbf{Y}) = P(\mathbf{R})$, then the data are missing completely at random (MCAR). One such scenario occurs in planned missing data designs, where missing values are "assigned" randomly to each participant. If the occurrence of missing data depends on the observed data but missing data occur "at random" with these taken into account, that is, $P(\mathbf{R} \mid \mathbf{Y}) = P(\mathbf{R} \mid \mathbf{Y}_{obs})$, then the data are MAR. The two missing data mechanisms MCAR and MAR are often called "ignorable" because the exact missing data mechanism need not be known to perform MI (for a more general discussion of the role of "ignorability," see Enders, 2010). If neither condition holds, that is, $P(\mathbf{R} \mid \mathbf{Y}) = P(\mathbf{R} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, then the data are missing not at random (MNAR). Most software implementations of MI rely on the assumption that the data are MAR. Performing MI under MNAR is possible but requires

making strong assumptions about the missing data mechanism and is most often used for sensitivity analyses (see Carpenter & Kenward, 2013). To enhance the plausibility of the MAR assumption, it has been suggested that auxiliary variables be included in the imputation model. These variables are related to either the propensity of missing data or the variables with missing values themselves, without necessarily being part of the model of interest (Collins, Schafer, & Kam, 2001). In our empirical example, some data are missing by design and are thus MCAR. For the remaining data, we will assume that the data are MAR, given the observed portions of the data that can be included as auxiliary variables.

Furthermore, the imputation model must be at least as complex as the analysis model. If variables or parameters that are relevant for the analysis model are not included in the imputation model, then the procedure could yield biased results (Meng, 1994; Schafer, 2003). For example, assume that a researcher is interested in testing an interaction between two variables in a multiple regression analysis with partially missing data. In this case, it would be important that the interaction effect (i.e., product term) is incorporated in the imputation model (Enders, Baraldi, & Cham, 2014). Similarly, if one is interested in estimating the intraclass correlation (ICC; that is, the variance within and between groups) with incomplete data, it would be crucial to take into account the clustered data structure (Taljaard, Donner, & Klar, 2008). If the model of interest includes random slopes, then the imputation model should allow for different slopes across groups. Choosing an appropriate imputation model can be challenging, and it may be tempting to resort to ad hoc methods for treating multilevel missing data (Graham, 2012). For example, it has been suggested that the multilevel structure be represented by creating a set of dummy-indicator variables (Drechsler, 2015; Graham, 2009; White, Royston, & Wood, 2011). In this approach, the dummy indicators are included in the single-level imputation model, and a separate intercept (or fixed effect) is estimated for each group. However, recent simulation research has indicated that such methods can distort parameter estimates and standard errors in multilevel analyses (Andridge, 2011; Enders, Mistler, & Keller, 2016; Lüdtke, Robitzsch, & Grund, in press).

Two broad approaches to performing MI can be distinguished. In the joint modeling approach, a single statistical model is used for imputing all incomplete variables simultaneously (e.g., Schafer & Yucel, 2002). In contrast, in the fully conditional specification of MI, each variable is imputed in turn using a sequence of imputation models (van Buuren & Groothuis-Oudshoorn, 2011). In the present article, we focus on the pan package, which follows the joint modeling paradigm (for a discussion, see Carpenter & Kenward, 2013).

The Multivariate Linear Mixed-Effects Model

The statistical model underlying the pan package is a multivariate extension of regular (univariate) multilevel models; that is, it represents multiple dependent variables simultaneously. In addition, the model may feature a number of predictor variables with associated fixed and random effects. In the following, we refer to this model as the multivariate linear mixed-effects model (MLMM, see Schafer & Yucel, 2002). The model reads,

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_j + \mathbf{e}_{ij}, \qquad (3)$$

where $\mathbf{y}_{ij}$ is the $(1 \times r)$ vector of responses for person i in group j, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are $(1 \times p)$ and $(1 \times q)$ vectors of covariate values, $\boldsymbol{\beta}$ is a $(p \times r)$ matrix of fixed effects, $\mathbf{b}_j$ is a $(q \times r)$ matrix of random effects, and $\mathbf{e}_{ij}$ is a $(1 \times r)$ vector of residuals. In most cases, the matrix $\mathbf{z}_{ij}$ contains a subset of the values in $\mathbf{x}_{ij}$, and both will contain at least a "one" for the regression intercept. The random-effects matrix $\mathbf{b}_j$, with columns stacked upon another, is assumed to follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$, independently and identically for all groups. The vector of residuals $\mathbf{e}_{ij}$ is assumed to follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, independently and identically for all individuals. The MLMM imputes all variables on the left-hand side of the model equation given the variables on the right-hand side (with fixed and random effects). Only the variables on the left-hand side (i.e., in $\mathbf{y}_{ij}$) may contain missing values, whereas the variables on the right-hand side must be completely observed (i.e., in $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$). In the following, we will distinguish between two broad approaches to MI of incomplete multilevel data using pan's MLMM (see Table 3 for an illustration).

Multivariate empty model. In the first approach, the emphasis is placed on the left-hand side of the model (i.e., the $\mathbf{y}_{ij}$), whereas the right-hand side includes only the intercept ($\mathbf{x}_{ij} = \mathbf{z}_{ij} = 1$). For all variables included on the left-hand side, the MLMM decomposes their variances and covariances into separate between- ($\boldsymbol{\Psi}$) and within-group portions ($\boldsymbol{\Sigma}$). We refer to this approach as the multivariate *empty* model. This model can be understood as a multivariate variant of the regular empty multilevel model—also known as the null model or the intercept-only model—in which the dependent variable is also decomposed into between- and within-group components, but the predictor side of the model remains empty. The upper half of Table 3 contains an example with three variables, each of which may or may not contain missing data. As can be seen in Table 3, the three variables decompose into a fixed term common to all persons and groups, a random intercept unique to each group, and an error term unique to each person. The covariance matrices of random effects and errors, $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$, contain the variances and covariances of the $\mathbf{y}_{ij}$ and allow for different relations between the dependent variables at the group and the person level, respectively. For that reason, the empty model is especially useful if researchers are interested in

Table 3. Two Multivariate Linear Mixed-Effects Models for Missing Data.

General notation

$$\mathbf{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_{ij} + \mathbf{e}_{ij}$$

Multivariate empty model

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}_{ij} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}_j + \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}_{ij}$$

target variables   fixed effects (intercepts)   random effects (intercepts)   residuals

$$\begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}_j^T \sim N(\mathbf{0},\Psi) \quad \text{and} \quad \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}_{ij} \sim N(\mathbf{0},\Sigma)$$

Full mixed-effects model

$$\begin{bmatrix} y_1 & y_2 \end{bmatrix}_{ij} = \begin{bmatrix} 1 & x_1 \end{bmatrix}_{ij}\begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \end{bmatrix} + \begin{bmatrix} 1 & x_1 \end{bmatrix}_{ij}\begin{bmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \end{bmatrix}_j + \begin{bmatrix} e_1 & e_2 \end{bmatrix}_{ij}$$

target variables   fixed effects (intercepts, slopes)   random effects (intercepts, slopes)   residuals

$$\begin{bmatrix} b_{01} & b_{11} & b_{02} & b_{12} \end{bmatrix}_j^T \sim N(\mathbf{0},\Psi) \quad \text{and} \quad \begin{bmatrix} e_1 & e_2 \end{bmatrix}_{ij} \sim N(\mathbf{0},\Sigma)$$

Note. The predictor x_1 is assumed to be completely observed. Vectorization of the random-effects matrix $\mathbf{b}_j$ is achieved by stacking its columns.

estimating relationships at the individual and the group level as in random-intercept models with group-level predictors (see Example 1).

Full mixed-effects model. The second approach utilizes both sides of the model. For all variables included on the right-hand side (i.e., in $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$), the MLMM estimates fixed and/or random effects, respectively. The lower half of Table 3 contains an example with two dependent variables with missing data and one fully observed predictor variable x_1 in $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$. As can be seen, the model includes both fixed and random effects for the intercept and x_1. We refer to this model as the *full* mixed-effects model because it includes both random intercepts and slopes where possible. Note that x_1 is not decomposed in this model, and the fixed and random effects represent the overall effects of that variable on the dependent variables. To include separate within- and between-group effects of x_1, the variable must be decomposed into between- and within-group portions prior to performing MI (e.g., by including the group mean as an additional predictor). The full mixed-effects model is particularly useful if the model of interest includes random slopes because the slope variance is represented in the imputation model (see Example 2).

Software Alternatives

A number of software packages have introduced procedures for MI of multilevel data. The software M*plus* (Muthén & Muthén, 2012) implements a two-level model similar to the empty model in pan (denoted H1) as well as a second procedure (denoted H0) for more complex models (e.g., random-slope models; see Asparouhov & Muthén, 2010b). Joint modeling approaches are also available in SAS (Mistler, 2013), REALCOM (Carpenter, Goldstein, & Kenward, 2011), and the R package jomo (Quartagno & Carpenter, 2016). A fully conditional specification of MI is available in the R package mice (van Buuren & Groothuis-Oudshoorn, 2011). For some of these packages, it is possible to follow similar analysis steps as outlined in this article for the pan package. We return to this possibility in a later section.

Example Applications With Multilevel Missing Data

To demonstrate the application of pan for imputing incomplete multilevel data, we made use of the mitml package. This package provides a more convenient interface for the pan algorithm and some additional tools for handling multiply imputed data sets and combining their results (Grund, Robitzsch, & Lüdtke, 2016). Following the imputation, we used the package lme4 for estimating the two models of interest (Bates, Maechler, Bolker, & Walker, 2014). We repeated the imputation and estimation in both examples using the popular software M*plus*.[1] The results were mostly consistent with those of pan and will not be discussed in detail. Input files for M*plus* are provided in Supplement A in the supplemental online materials.

The example data set is structured as follows. The first variable (ID) denotes the class membership of each student. The remaining variables are as described above and may contain different amounts of missing data, which are denoted as NA.

ID	MathAchiev	ReadAchiev	CognAbility	SES	MathDis	ReadDis	SchClimate
1	517.92	547.65	52	70	1.6	1.8	1.25
1	524.78	633.82	46	40	3.0	2.4	2.50
1	544.50	474.04	59	34	NA	2.4	1.00

Treating and analyzing multilevel missing data usually involves the following steps. First, an appropriate imputation model must be specified. As outlined above, the analysis model must be considered at that point so that the relevant variables, parameters, and auxiliary variables are included in the imputation model. Second, the imputation procedure must be carried out, resulting in a number of imputed data sets. Third, the data sets must be analyzed separately, and the resulting parameter estimates are combined according to the rules described in Rubin (1987; for alternatives, see Carpenter & Kenward, 2013; Reiter & Raghunathan, 2007). These steps can be carried out using the mitml package. To illustrate the impact of different approaches for handling incomplete multilevel data, we also provide the results obtained from single-level MI, which ignores the multilevel structure, and from LD (i.e., complete case analysis). The computer code and output files are provided in Supplement B in the supplemental online materials.

Example 1: Random-Intercept Model

In the first example, the model of interest examined the between- and within-group effects of DPM on MA, while controlling for SES at the individual and class level.

$$\mathrm{MA}_{ij} = \beta_0 + \beta_1\left(\mathrm{DPM}_{ij} - \overline{\mathrm{DPM}}_j\right) +$$
$$\beta_2\left(\mathrm{SES}_{ij} - \overline{\mathrm{SES}}_j\right) + \beta_3\overline{\mathrm{DPM}}_j + \quad (1, \text{revisited})$$
$$\beta_4\overline{\mathrm{SES}}_j + \upsilon_{0j} + \varepsilon_{ij}.$$

Choosing an appropriate imputation model is straightforward in this case because the multivariate empty model is suitable for random-intercept models in general. In addition, the empty model includes between- as well as within-group relations as required by the model of interest (in Ψ and Σ; see Table 3). Recall that the empty model is specified by writing all variables on the left-hand side of the model equation. In R, the imputation model for this example is set up as follows.

```
# SETUP: imputation model (variance
# decomposition model)
fml <- MathAchiev + MathDis + SES +
 ReadAchiev + CognAbility + ReadDis +
 SchClimate ~ 1 + (1|ID)
```

The mitml package uses formula objects to represent the imputation model. The "~" symbol separates the left- and right-hand side of the model. The left-hand side contains the three variables of interest and the auxiliary variables (i.e., reading achievement, CA, ratings of DPR and school climate). On the right-hand side, the intercept is specified both as a fixed (1) and a random effect (1|ID), where the "|" symbol denotes clustering.

For running the pan algorithm, the mitml package offers the function panImpute as its main interface. The pan algorithm uses Markov Chain Monte Carlo (MCMC) techniques to draw replacements for the missing values. At each iteration of the procedure, a new set of parameters and replacements is simulated. The distribution from which the replacements are drawn is called the posterior predictive distribution of the missing data (Gelman, Carlin, Stern, & Rubin, 2013). The full procedure is divided into a burn-in phase and an imputation phase (see Enders, 2010). During burn-in, the algorithm performs a number of iterations without saving any imputations, thus ensuring that the parameters of the imputation model have converged to stationary distributions. In other words, the burn-in phase must be long enough for the algorithm to "stabilize" before any replacements are drawn. Then, during the imputation phase, a number (m) of imputed data sets are drawn, each spread a number of iterations apart. The fact that imputations are not drawn directly from consecutive iterations ensures that the imputed data sets constitute independent random draws from the posterior predictive distribution. Specifically, consecutive iterations in MCMC are often correlated to some degree (autocorrelation), whereas multiply imputed data sets must be drawn independently of one another. Thus, the number of iterations chosen between imputations must be large enough for autocorrelation to vanish.

In the first example, we ran pan for 50,000 burn-in iterations, after which $m = 100$ imputed data sets were drawn, each spread 5,000 iterations apart. While these numbers may seem large, recent studies have advocated generating such large numbers of imputations, particularly when large portions of the data are missing (Bodner, 2008; Graham, Olchowski, & Gilreath, 2007). The number of iterations for burn-in and between imputations was chosen such that convergence could be ensured, as described below. The respective command using mitml was as follows.

```
# IMPUTATION:
imp <- panImpute(dat, formula=fml,
 n.burn=50000, n.iter=5000, m=100,
 seed=1234)
```

The mitml package saves the imputation in a special format that is designed to handle large data sets. To obtain a list containing all the imputed data sets, the function mitmlComplete is used. The necessary command is printed below.

```
# list of imputed data sets
impList <- mitmlComplete(imp, print="all")
```

Convergence diagnostics. For the analysis to yield reliable results, it must be ensured that the pan algorithm has converged and that the imputed data sets are approximately independent draws from the posterior predictive distribution (for a detailed discussion of convergence assessment in MCMC, see Cowles & Carlin, 1996; Gill, 2014; Jackman, 2009). The mitml package offers two ways of doing so. The first option is to examine the potential scale reduction factor (also called $\hat{R}$; Gelman & Rubin, 1992) for the parameters of the imputation model. Originally intended for analyzing multiple MCMC chains, $\hat{R}$ is calculated

here by discarding the burn-in iterations and dividing the single MCMC chain for each parameter into multiple segments (see Asparouhov & Muthén, 2010a). The $\hat{R}$ statistic then compares the variance within and between segments to detect a potential "drifting" of the chain, that is, chains that are more variable overall than one would expect, based on the variability within segments. In the mitml package, $\hat{R}$ is included in the summary of an imputed data object.

```
# DIAGNOSTIC: summary and potential scale
# reduction
summary(imp)
```

In addition to the potential scale reduction, the output of summary includes details about the imputation procedure and the missing data rate per variable. In this example, the output was as follows (truncated for better readability).

```
Call:

panImpute(data = dat, formula = fml, n.burn = 50000, n.iter = 5000,m = 100, seed = 1234)

Cluster variable:        ID
Target variables:        MathAchiev MathDis SES ReadAchiev CognAbility ReadDis SchClimate
Fixed effect predictors: (Intercept)
Random effect predictors: (Intercept)

Performed 50000 burn-in iterations, and generated 100 imputed data sets, each 5000
iterations apart.

Potential scale reduction (Rhat, imputation phase):

          Min     25%    Mean   Median     75%    Max
Beta:    1.000   1.000   1.000   1.000   1.000  1.000
Psi:     1.000   1.000   1.001   1.000   1.001  1.011
Sigma:   1.000   1.000   1.000   1.000   1.000  1.001

Largest potential scale reduction:
Beta: [1,6], Psi: [1,1], Sigma: [1,1]

Missing data per variable:

    ID    MathAchiev  MathDis   SES   ReadAchiev  CognAbility ReadDis SchClimate

MD% 0      19.4        61.4     35.0  0            0           21.5    21.7
```

Ideally, $\hat{R}$ should be close to one for all parameters (Gelman & Rubin, 1992). If larger values occur (say, above 1.050), a longer burn-in period may be required. Due to the potentially large number of statistical parameters, the mitml package displays only summary statistics for these values

while emphasizing the parameters with the largest $\hat{R}$. As shown in the output, the $\hat{R}$ was well below 1.050 for all parameters. The parameter with the largest $\hat{R}$ was the first diagonal entry of the random–effects covariance matrix Ψ, that is, the intercept variance for MA scores ($\hat{R} = 1.011$).
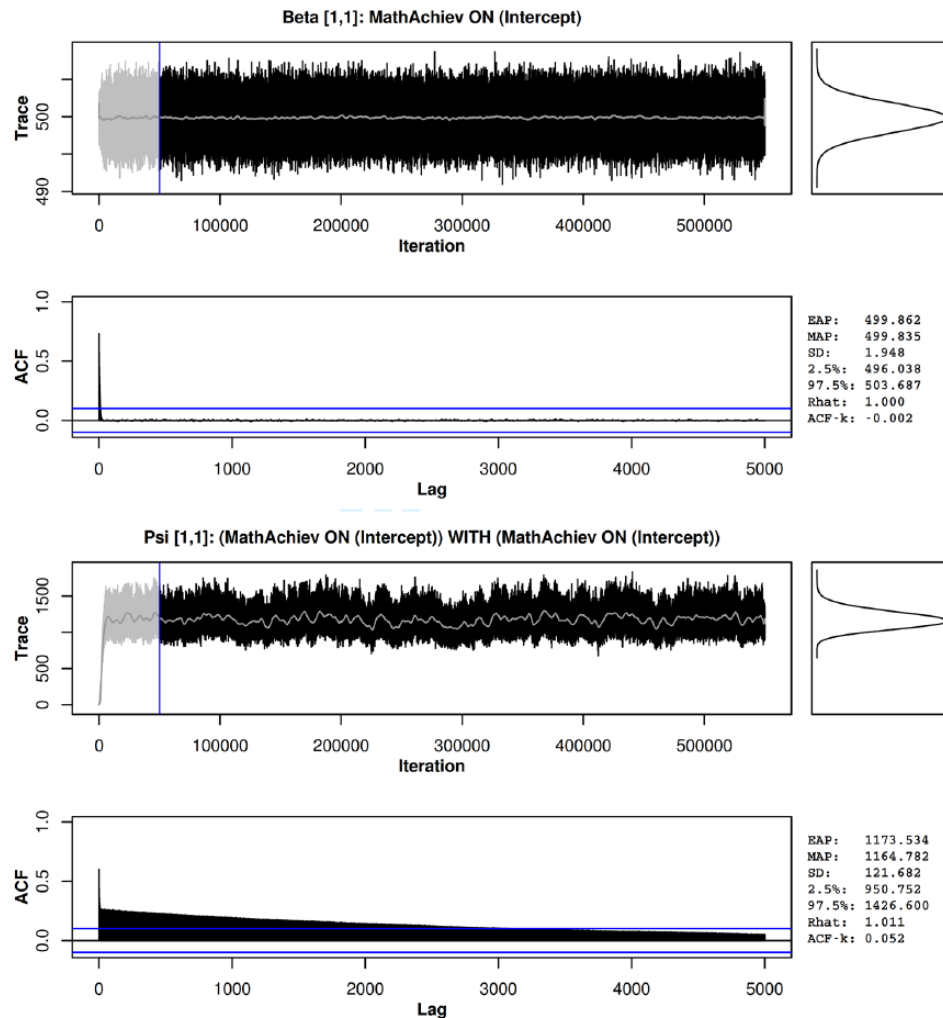
Figure 1. Diagnostic plots for the fixed intercept (top) and the intercept variance (bottom) of math achievement in the imputation model.
Note. The trace plot includes all iterations from the burn-in and the imputation phase. The autocorrelation plot and the posterior summaries are calculated only from the imputation phase.

However, $\hat{R}$ has been criticized, and large values of $\hat{R}$ need not always indicate poor convergence (e.g., Geyer, 1992). Therefore, as a second option, diagnostic plots should be considered. For each parameter in the imputation model, the plot function may produce a trace plot for all iterations during and/or after burn-in, an autocorrelation plot for all iterations after burn-in, and a summary of the parameter's posterior distribution. The trace plot is a graphical representation of the MCMC chain for each parameter, and it shows the values of that parameter at each iteration. The autocorrelation plot shows the degree to which consecutive elements of the MCMC chain are correlated (when spread a number of iterations apart). The posterior summary includes a density plot of the MCMC chain and a number of summary statistics relating to both the MCMC

chain and its autocorrelation. The diagnostic plots can be requested as follows.

```
plot(imp, trace="all")
```

Here, we discuss the diagnostic plots only for the fixed intercept and the intercept variance for MA, which exhibited the worst convergence behavior of all parameters (see Figure 1). The trace plots showed no sign of "drifting" or substantial change after the burn-in phase, indicating that 50,000 iterations were sufficient for the parameters to reach their respective target distributions. Autocorrelation was quite persistent for the intercept variance but had essentially died out by lag 5,000. Therefore, imputations spread 5,000 iterations apart

Table 4. Estimates of the Intraclass Correlation for the Variables of Interest in Example 1.

	Multilevel MI	Single-level MI	Listwise deletion
ICC_{MA}	.121	.111	.115
ICC_{SES}	.122	.072	.134
ICC_{DPM}	.179	.100	.169

Note. MI = multiple imputation; ICC = intraclass correlation; MA = math achievement; SES = socioeconomic status; DPM = disciplinary problems in math classes.

could be considered independent. We concluded that the parameters had converged and that the imputed data sets constituted independent draws from the posterior predictive distribution of the missing data.

ICCs. Usually the first step in analyzing multilevel data is to estimate the ICC of the variables of interest. Therefore, before proceeding with the model of interest, we will illustrate the analysis of multiply imputed data sets by fitting intercept-only models for MA, SES, and DPM to estimate their ICCs. To obtain final parameter estimates from multiply imputed data sets, the analysis model must be fit separately to each data set, and the resulting estimates must be combined. In the mitml package, the list of imputed data sets (here impList) can be analyzed by using the functions with and within. The within function is used to transform the imputed data sets and carry out smaller computations prior to fitting the analysis model. The with function returns the model fit itself. The intercept-only model for MA can be fit as shown below (for DPM and SES, see Supplement B in the supplemental online materials). We used the lmer function from the lme4 package to fit the analysis models.

```
# FIT: null model for math achievement
fit <- with(impList, lmer(MathAchiev ~ 1 +
 (1|ID)))
```

This results in a list of 100 fitted analysis models, one for each imputed data set. The parameter estimates of the fitted models can be combined by using the rules described in Rubin (1987). The mitml package implements Rubin's rules in the testEstimates function, which returns the combined estimates for all fixed effects and, when used with lme4, the variance components and the residual ICC (see Supplement B in the supplemental online materials). The final estimates can be requested as given below.

```
# final parameter estimates (Rubin's rules)
testEstimates(fit, var.comp=TRUE)
```

The resulting estimates of the ICCs are presented in Table 4 along with the estimates from single-level MI and LD. Most

notably, multilevel MI (using pan) led to much larger estimates of the ICCs than single-level MI, especially for variables with large amounts of missing data (DPM and SES). This illustrates the importance of accounting for the multilevel structure when conducting MI for multilevel data. The estimates obtained from LD were closer to those of multilevel MI without any obvious pattern emerging. These results are consistent with previous research that was based on simulation studies (e.g., Taljaard et al., 2008; van Buuren, 2011).

Model of interest. The procedures outlined above can also be used for fitting the model of interest (Equation 1). Prior to fitting the model, the group means for DPM and SES must be calculated in each imputed data set, and the student-level variables must be centered around their respective group means. Such computations can be carried out using within as shown below.

```
# TRANSFORM: class means for MathDis and SES
impList <- within( impList, {

MathDis.CLS <- clusterMeans(MathDis,ID)
SES.CLS <- clusterMeans(SES,ID)
})
# TRANSFORM: center student-level predictors
impList <- within( impList, {

MathDis.STU <- MathDis - MathDis.CLS
SES.STU <- SES - SES.CLS
})
```

This results in a list of 100 imputed data sets, similar to the original list, but with the group means and the group-mean-centered variables added to each data set. Finally, the model of interest was fit as shown below using the lme4 package (using with).

```
# FIT: model of interest
fit <- with(impList, lmer(MathAchiev ~ 1 +
 SES.STU + SES.CLS + MathDis.STU +
 MathDis.CLS + (1|ID)))
```

As before, testEstimates returned the final parameter estimates and inferences.

```
# final parameter estimates (Rubin's rules)
testEstimates(fit, var.comp=TRUE)
```

The output of testEstimates includes the final parameter estimates, the MI standard errors, the degrees of freedom and value of the reference t distribution,[2] the fraction of missing information (FMI), and the relative increase in variance due to nonresponse (RIV). Even though the FMI is not frequently reported in empirical studies, it holds great value for the interpretation of results and has been recommended as a

```
Call:

testEstimates(model = fit, var.comp = TRUE)

Final parameter estimates and inferences obtained from 100 imputed data sets.

              Estimate    Std.Error    t.value         df    p.value     RIV     FMI
(Intercept)    502.498       19.254     26.098   1210.447      0.000   0.401   0.287
SES.STU          1.065        0.084     12.614    526.055      0.000   0.766   0.436
SES.CLS          2.150        0.267      8.065   1429.302      0.000   0.357   0.264
MathDis.STU    -20.736        2.032    -10.203    372.305      0.000   1.065   0.518
MathDis.CLS    -41.131        5.035     -8.169   1358.967      0.000   0.370   0.271


                     Estimate
Intercept~~Intercept|ID   655.957
Residual~~Residual       8318.936
ICC|ID                      0.073

Unadjusted hypothesis test as appropriate in larger samples.
```

diagnostic tool for analyzing multiply imputed data sets (Bodner, 2008). The FMI represents the amount of information about an estimand that is lost due to missing data (Allison, 2001; Enders, 2010). In other words, the FMI shows the loss of "efficiency" when estimating parameters from multiply imputed data sets (Savalei & Rhemtulla, 2012). Similar to the FMI, the RIV denotes the increase in sampling variability in an estimate that can be attributed to missing data (see Enders, 2010). The output for the model of interest is given above.

The results for multilevel MI, single-level MI, and LD are presented in Table 5. In general, a higher SES was associated with higher MA scores, whereas test scores tended to be lower if students reported disciplinary problems in class. The estimates at the class level were roughly twice as large as those at the student level. Single-level MI led to similar estimates of within-group effects, but the estimates of the between-group effects were consistently larger than those obtained from multilevel MI. LD produced larger standard errors (especially at the student level) and smaller estimates of class-level effects.

Researchers are often interested in estimating *contextual* effects, that is, group-level effects when controlling for effects at the student level. For example, the contextual effect of SES can be calculated simply by subtracting its within-group coefficient from its between-group coefficient (Kreft, de Leeuw, & Aiken, 1995). Effects constrained in such a way can be tested against zero using the testConstraints function as shown below.

```
# contextual effect via model constraints
testConstraints(fit, "SES.CLS - SES.STU")
```

Testing constrained parameters is based on the delta method (e.g., Casella & Berger, 2002; Raykov & Marcoulides, 2004), and the pooled test for multiply imputed data sets is based on the method by Li, Raghunathan, and Rubin (1991).[3] For further details, we refer to the package documentation. The output for testing the contextual effect of SES is printed below.

```
Call:

testConstraints(model = fit, constraints =
 "SES.CLS - SES.STU")

Hypothesis test calculated from 100 imputed
data sets. The following constraints were
specified:

SES.CLS - SES.STU

Combination method: D1

  F.value  df1        df2   p.value    RIV
   15.297    1   1292.993     0.000  0.365

Unadjusted hypothesis test as appropriate in
Larger samples.
```

In this example, the contextual effect of SES was statistically significant at $p < .001$ ($F = 15.297$, $df_1 = 1$, $df_2 = 1293.0$). Thus, it appeared that classes with a higher SES tended to have higher MA scores, even after controlling for SES at the student level.

Notice that, throughout this example, we used manifest group means as predictor variables in the multilevel

Table 5. Results From Multilevel MI, Single-Level MI, and Listwise Deletion for Example 1 (Random-Intercept Model).

	Multilevel MI			Single-level MI			Listwise deletion	
	Estimate	*SE*	FMI	Estimate	*SE*	FMI	Estimate	*SE*
Intercept	502.498	19.254	0.287	502.268	22.326	0.231	505.911	20.063
SES_{ij}	1.065	0.084	0.436	1.054	0.080	0.401	0.849	0.138
$\overline{SES}_j$	2.150	0.264	0.316	2.474	0.303	0.237	1.753	0.275
DPM_{ij}	−20.736	2.032	0.518	−21.609	1.816	0.440	−21.874	3.054
$\overline{DPM}_j$	−41.131	5.035	0.271	−47.165	5.764	0.187	−31.552	5.689
$Var(\upsilon_{0j})$	655.957			592.132			731.860	
$Var(\varepsilon_{ij})$	8,318.936			8,387.913			8,299.000	

Note. Estimates were significant at $p < .001$. MI = multiple imputation; FMI = fraction of missing information; SES = socioeconomic status; DPM = disciplinary problems in math classes; υ_{0j} = random intercepts; ε_{ij} = residuals at Level 1.

analyses. This is different from the imputation model, where the group-level portions of variables are represented as latent variables (i.e., random effects). In general, an imputation model based on latent group means (i.e., random effects) yields similar results as one that is based on manifest means, and both can be considered correct imputation models for multilevel data (Carpenter & Kenward, 2013; Lüdtke et al., in press; Mistler, 2015). However, when estimating the model of interest, the predictors' group means may again be considered as latent, and slightly different results are expected for such an analysis model (Asparouhov & Muthén, 2006; Lüdtke et al., 2008). A further discussion can be found in Supplement C in the supplemental online materials along with the M*plus* syntax files for fitting the latent analysis model. In this example, the two analysis models led to essentially the same conclusions.

Example 2: Random-Slope Model

In the second example, the model of interest examined the effect of student's CA and SES on students' MA scores. The effect of SES is assumed to be fixed, whereas the effect of CA is allowed to vary across groups.

$$\begin{aligned} MA_{ij} = {}& \beta_0 + \beta_1\left(CA_{ij} - \overline{CA}_j\right) + \beta_2\overline{CA}_j + \\ & \beta_3\left(SES_{ij} - \overline{SES}_j\right) + \beta_4\overline{SES}_j + \\ & \upsilon_{0j} + \upsilon_{1j}\left(CA_{ij} - \overline{CA}_j\right) + \varepsilon_{ij}. \end{aligned}$$

(2, revisited)

As discussed before, the imputation model must consider the model of interest. In this example, the effect of CA is assumed to vary across groups, which must be reflected in the imputation model. The full mixed-effects model was used for this task (see Table 3). Furthermore, we calculated the group means and the group-mean-centered CA scores so that we could use them in the imputation model. This was achieved using within as shown below.

```
# TRANSFORM: group mean centering (prior to
# performing MI)
dat <- within(dat, {
CognAbility.CLS <- clusterMeans(CognAbility,
 ID)
CognAbility.STU <- CognAbility -
CognAbility.CLS
})
```

Because CA scores are available for all students, it can be included on the right-hand side of the imputation model, which also allows the slope variance to be specified. The imputation model was set up as follows.

```
# SETUP: imputation model (random effects
# model)
fml <- MathAchiev + SES + ReadAchiev +
 MathDis + ReadDis + SchClimate ~ 1 +
 CognAbility.STU + CognAbility.CLS +
 (1+CognAbility.STU|ID)
```

The model includes MA, SES, and the auxiliary variables on the left-hand side of the equation. To include the slope variance, CA is featured on the right-hand side, where (1+CognAbility.STU|ID) allows the intercept and the effect of the group-mean-centered CA scores to vary across groups.

It is worth noting that the MLMM assumes the same random-effects structure for *all* dependent variables in the model. In other words, the full mixed-effects model includes not only the intercepts and slopes for the regressions of MA and SES on CA but also for the four remaining variables. Thus, users of pan should be wary of including too many variables if the model contains multiple random effects. The number of parameters can increase rapidly by adding dependent variables or predictors with random effects to the model, possibly requiring a large number of iterations for the model to converge.

As in the first example, the imputation procedure is started by using panImpute while referring to the data set and the model equation. In this example, we let pan perform 100,000
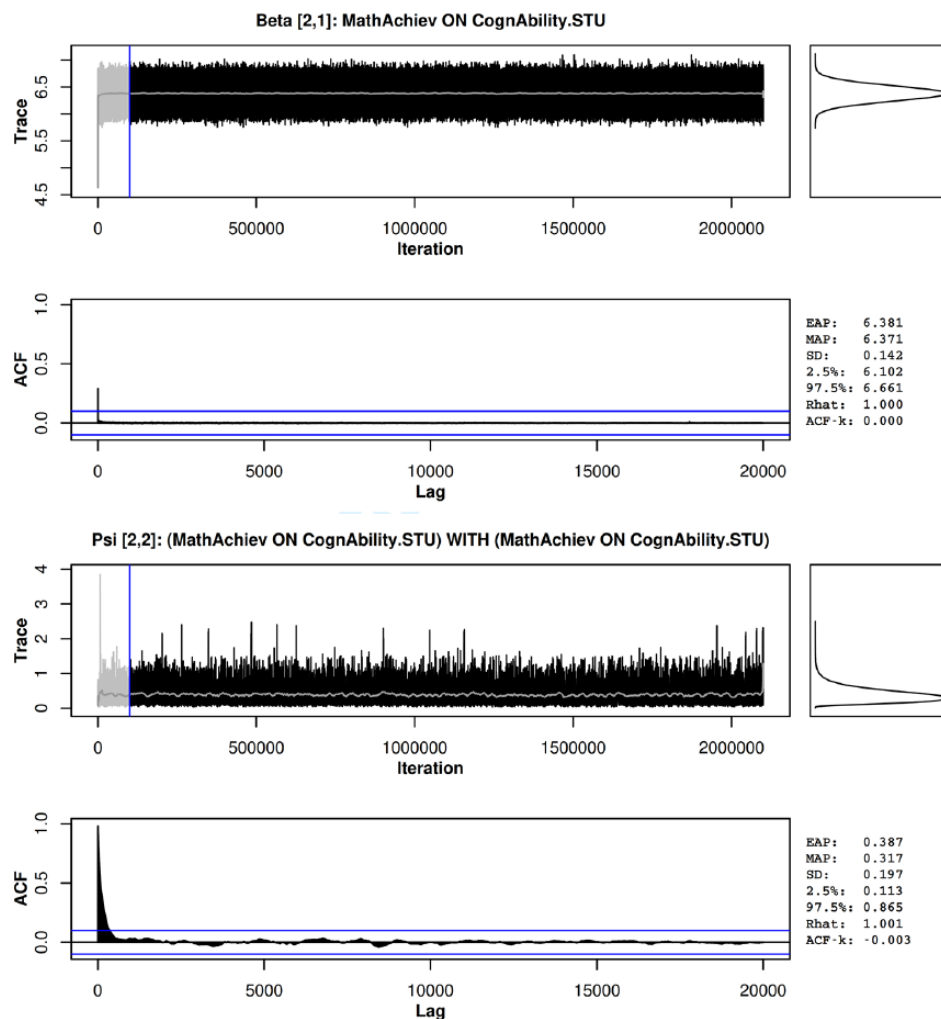
Figure 2. Diagnostic plots for the fixed effect (top) and the slope variance (bottom) for the regression of math achievement on cognitive ability in the full mixed-effects model.
Note. The trace plot includes all iterations from the burn-in and the imputation phase. The autocorrelation plot and the posterior summaries are calculated from the imputation phase.

burn-in iterations, after which we generated $m = 100$ imputed data sets, each spread 20,000 iterations apart. The respective command was as follows.

```
# IMPUTATION:
imp <- panImpute(dat, formula=fml,
 n.burn=100000, n.iter=20000, m=100,
 seed=1234)
```

As before, a list of imputed data sets was extracted using mitmlComplete. The code is not displayed here because it is identical to the previous example (see Supplement B in the supplemental online materials).

Convergence diagnostics. Before proceeding with the analysis, it must be ensured that the pan algorithm has converged during burn-in and that the interval between imputations was sufficiently large. Again, $\hat{R}$ gives an idea of possible problems with convergence and is accessed through the summary. The largest value of $\hat{R}$ was 1.001 in this case, indicating that the MCMC chain had become stationary for all parameters. Examining the diagnostic plots not only supported this impression but also indicated that some parameters were affected by autocorrelation. As shown in Figure 2, the parameters related to the variables of interest converged quickly and did not suffer greatly from autocorrelation. For some parameters, especially the group-level variance components,

Table 6. Results From Multilevel MI, Single-Level MI, and Listwise Deletion for Example 2 (Random-Slope Model).

	Multilevel MI			Single-level MI			Listwise deletion	
	Estimate	*SE*	FMI	Estimate	*SE*	FMI	Estimate	*SE*
Intercept	84.573	21.364	0.108	79.514	20.562	0.113	96.792	25.956
CA_{ij}	6.114	0.150	0.185	6.138	0.154	0.232	6.250	0.185
$\overline{CA}_j$	7.608	0.521	0.164	7.549	0.501	0.160	7.714	0.584
SES_{ij}	0.605	0.077	0.463	0.599	0.075	0.445	0.578	0.079
$\overline{SES}_j$	1.081	0.261	0.289	1.254	0.291	0.277	0.749	0.253
$Var(\upsilon_{0j})$	485.050			417.802			540.924	
$Var(\upsilon_{1j})$	1.528			1.381			1.572	
$Cov(\upsilon_{0j}, \upsilon_{1j})$	0.333			1.470			5.535	
$Var(\varepsilon_{ij})$	6,452.506			6,547.366			6,287.600	

Note. Estimates for the fixed effects were significant at $p < .001$. MI = multiple imputation; FMI = fraction of missing information; CA = cognitive ability; SES = socioeconomic status; υ_{0j} = random intercepts; υ_{1j} = random slopes; ε_{ij} = residuals at Level 1.

the autocorrelation was quite persistent but vanished for all parameters with a lag of 15,000 to 20,000 iterations.

Model of interest. To estimate the model of interest, the student-level variables were centered around their group means (using within), and the model was fit using the lme4 package (using with). We changed the method for estimating the multilevel model from restricted maximum likelihood (REML) to full information maximum likelihood (FIML) because the model comparison that was conducted as a later step in this analysis required that the analysis models were estimated using FIML. The code for fitting the model of interest is given below.

```
# FIT: model of interest
fit <- with (impList, lmer(MathAchiev ~ 1 +
 CognAbility.STU + CognAbility.CLS + SES.STU +
 SES.CLS + (1+CognAbility.STU|ID), REML=FALSE))
```

The final parameter estimates and inferences were obtained using testEstimates. These are presented in Table 6, along with the estimates from single-level MI and LD. Students with higher CA (as compared with their class average) tended to score higher on the MA test after controlling for SES. This relation appeared to vary substantially across groups. In comparison, single-level MI produced lower estimates of the intercept and slope variance and a slightly larger estimate of the class-level effect of CA. For LD, the estimates of the fixed effects and variance components were slightly different from those obtained with MI but comparable altogether. Results obtained using the H0 imputation in M*plus* yielded results similar to those produced by pan.[4]

When estimating multilevel models with random slopes, researchers are often interested in whether or not the regression coefficients vary substantially across groups. For this purpose, likelihood-ratio tests (LRTs), which compare the model

of interest with an alternative model that constrains the slope variance to zero, are often conducted (see Snijders & Bosker, 2012). A method for pooling the LRT across multiply imputed data sets was suggested by Meng and Rubin (1992). This procedure is accessible in mitml through the testModels function. The alternative model is similar to the model of interest, but only the intercept is allowed to vary across groups. The code for fitting the alternative model is given below.

```
# FIT: null model without random slopes
fit.null <- with (impList,

 lmer(MathAchiev ~ 1 + CognAbility.STU +
 CognAbility.CLS + SES.STU + SES.CLS +
 (1|ID),

REML=FALSE))
```

The two models can be compared using testModels, where method="D3" calls the procedure by Meng and Rubin (1992). The respective command was as follows.

```
# LRT for nonzero slope variance
testModels(fit, fit.null, method="D3")
```

The output of testModels for testing the slope variance is printed below.

```
Call:

testModels(model = fit, null.model =
 fit.null, method = "D3")

Model comparison calculated from 100 imputed
 data sets.

Combination method: D3

F.value   df1        df2     p.value    RIV
  5.119     2  10386.237       0.006  0.157
```

The pooled LRT was statistically significant at $p = .006$ ($F = 5.119$, $df_1 = 2$, $df_2 = 10386.2$) indicating that the slope variance was statistically different from zero. Thus, it appeared that students with different CA may differ more or less strongly in their MA scores, depending on the class to which they belong. It may be interesting to examine the determinants of this variation, for example, teachers' attributes or aspects of the learning environment. However, for the purpose of this article, we will not discuss these questions in detail. Research has shown that the LRT for variance components may suffer from low statistical power (see LaHuis & Ferguson, 2009; Stram & Lee, 1994). However, there are currently very few options for performing hypothesis tests for variance components with multiply imputed data sets other than Meng and Rubin's (1992) method.

Analyzing Imputations Generated by Alternative Software

As outlined above, there are a number of software alternatives for generating imputations for multilevel missing data, some of which are similar in scope to pan, and some of which provide further support for categorical, ordinal, or group-level variables. For example, if the model of interest also includes categorical variables with missing data, researchers may prefer using the R packages jomo or mice, or standalone software such as M*plus*. In general, the analysis steps presented here can be carried out on multiply imputed data sets irrespective of their origin. The requirement for using mitml's analysis functions is that the multiply imputed data sets are represented as a "list" of data sets in R. This can be achieved by either generating imputations using its wrapper functions, or by converting the imputed data into a list of data sets. The mitml package currently includes wrapper functions for pan (panImpute) and jomo (jomoImpute) as well as functions to convert imputed data sets generated by mice (mids2mitml.list). For other software packages, however, the conversion must be performed manually (e.g., using long2mitml.list, or as.mitml.list). The use of these functions is illustrated in the documentation of the package. In most applications, using the wrapper functions is recommended because it allows for using the tools for convergence diagnostics provided by mitml.

Discussion

Even though multilevel models are frequently used in psychology and the social sciences, MI of multilevel missing data is seldom discussed in the applied literature. As a result, LD, single-level MI, and ad hoc methods for representing the clustered data structure prevail in research practice (e.g., the dummy-indicator approach) even though research has shown that these methods can result in distorted parameter estimates in subsequent multilevel analyses. In the present

article, two empirical examples were used to illustrate the application of the two R packages pan and mitml to multilevel data. In Example 1, we discussed the application of pan to random-intercept models and for estimating between- and within-group effects. In Example 2, we focused on MI for multilevel models with random slopes and on estimating and testing the slope variance. We believe that researchers can benefit greatly from incorporating pan in their statistical analyses. Specifically, pan allows the special features of multilevel data to be preserved, a practice that is essential for obtaining reliable estimates from multilevel analyses and for understanding their results. Moreover, pan allows researchers to use all of the available information in the data and to include auxiliary information without altering the model of interest. By contrast, many interesting features of multilevel models may be distorted or even lost when using simpler methods for handling multilevel missing data. For example, the results from Example 1 showed that parameter estimates can be distorted if the imputation model ignores the multilevel structure of the data.

The field of statistical software is always in motion, and there continues to be a number of promising developments regarding multilevel MI. However, some problems still provide challenges for the future. For example, using multilevel MI can be difficult if missing data occur on predictor variables in models with random slopes or interaction effects. Graham (2012) suggested that MI for models with random slopes should be conducted separately for each group using single-level MI. Schafer (2001) proposed that incomplete predictor variables be treated as outcome variables in the imputation model, thus accepting a (possibly small) bias for the slope variance (see also Grund, Lüdtke, & Robitzsch, 2016). To mitigate this problem, it has been suggested to generate imputations for predictor variables in such a way that they are consistent with the model of interest (e.g., Goldstein, Carpenter, & Browne, 2014; Wu, 2010; see also Bartlett, Seaman, White, & Carpenter, 2015). These methods may provide an improvement over current implementations of multilevel MI in complex multilevel models with random slopes and missing values in predictor variables (Erler et al., in press). Unfortunately, they are currently not available in standard software.

Even though many algorithms exist for MI of multilevel data, the analysis often remains a challenge when software does not provide the tools for combining the results from multiply imputed data sets. Using the mitml package, we provided examples for combining simple parameter estimates, model comparisons, and model constraints with multiply imputed data sets. In addition to Rubin's (1987) rules, the package implements the procedures commonly referred to as D_1 (Li, Raghunathan, & Rubin, 1991; Reiter, 2007), D_2 (Li, Meng, Raghunathan, & Rubin, 1991), and D_3 (Meng & Rubin, 1992), which can be used for testing a variety of statistical hypotheses that potentially involve multiple parameters simultaneously (e.g., model comparisons). Nonetheless, open

questions remain about how some statistical quantities can be estimated from multiply imputed data sets. For example, it is not yet clear how researchers can obtain measures of the goodness-of-fit of multilevel models, which are often used for model selection (e.g., the model deviance, Akaike information criterion [AIC], or Bayesian information criterion [BIC]). Such procedures might be based on the methods by Li, Meng, et al. (1991) and Meng and Rubin (1992), or on variations thereof (Licht, 2010), but clear recommendations have not yet been made in the literature (see also Consentino & Claeskens, 2010; Grund, Lüdtke, & Robitzsch, in press).

The treatment of multilevel missing data offers many challenges, and state-of-the-art procedures are often not very accessible unless researchers are deeply familiar with missing data and MI. We hope that the present article will provide guidance for applied researchers and promote the use of modern missing data techniques such as MI. In general, we believe that pan is a powerful tool for treating multilevel missing data because many features of typical research questions can easily be represented in pan's MLMM. Future research should devote attention to increasing the accessibility of modern methods for handling and analyzing missing data. Currently, the use of MI in multilevel research, while largely desirable, is often hindered by the lack of accessible software and appropriate tools for analyzing multiply imputed data sets in real-world research scenarios. For future studies, the topic of multilevel missing data yields many interesting research questions that have yet to be explored.

Author Note

Declaration of Conflicting Interests

Funding

Notes

1. For Example 1, we used H1 imputation, which is equivalent to the multivariate empty model. For Example 2, we used H0 imputation because a model that was equivalent to the full mixed-effects model could not be specified using H1 imputation.
2. By default, testEstimates uses the standard t distribution proposed by Rubin (1987), which provides a test statistic that is appropriate in larger samples. Alternatively, the degrees of freedom may be adjusted for smaller samples as described in the package documentation (see also Barnard & Rubin, 1999; Reiter, 2007).
3. The method by Li, Raghunathan, and Rubin (1991) requires that the fractions of missing information (FMIs) are approximately equal across the parameters being tested (see also Licht, 2010). In the present case, the linear constraint being

tested has only one component and fulfills this requirement automatically.
4. These differences were negligible for most parameters, but M*plus* produced a large estimate of the slope variance, $Var(\upsilon_{1j}) = 2.277$. Despite the large similarities, there are some subtle differences between pan and the H0 imputation in M*plus*. For example, M*plus* uses "least informative" priors for H1 but improper priors for H0, which cannot be specified using pan. However, preliminary simulations could not replicate any difference between pan and M*plus*. A more in-depth exploration of these (relatively minor) differences was beyond the scope of this article and will be left for future research.

References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*, 1490-1528. doi:10.1177/0149206313478188

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, *53*, 57-74. doi:10.1002/bimj.201000140

Asparouhov, T., & Muthén, B. O. (2006). *Constructing covariates in multilevel regression* (Mplus Web Notes No. 11). Available from http://www.statmodel.com/

Asparouhov, T., & Muthén, B. O. (2010a). *Bayesian analysis using Mplus: Technical implementation* (Technical Appendix). Available from http://statmodel.com/

Asparouhov, T., & Muthén, B. O. (2010b). *Multiple imputation with Mplus* (Technical appendix). Available from http://statmodel.com/

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948-955. doi:10.1093/biomet/86.4.948

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*, 462-487. doi:10.1177/0962280214521348

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, *38*, 1845-1865. doi:10.1080/02664763.2010.529882

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 651-675. doi:10.1080/10705510802339072

Bos, W., Lankes, E. M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A., & Walther, G. (2005). *IGLU: Skalenhandbuch zur Dokumentation der Erhebungsinstrumente* [Scale handbook of the German PIRLS study]. Münster, Germany: Waxmann.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*(5), 1-14. Retrieved from http://www.jstatsoft.org/v45/i05

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Hoboken, NJ: Wiley.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Chen, G., & Bliese, P. D. (2002). The role of different levels of leadership in predicting self- and collective efficacy: Evidence for discontinuity. *Journal of Applied Psychology*, *87*, 549-556. doi:10.1037//0021-9010.87.3.549

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351. doi:10.1037/1082-989X.6.4.330

Consentino, F., & Claeskens, G. (2010). Order selection tests with multiply imputed data. *Computational Statistics & Data Analysis*, *54*, 2284-2295. doi:10.1016/j.csda.2010.04.009

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 883-904.

Culpepper, S. A., & Aguinis, H. (2011). R is for revolution: A cutting-edge, free, open source statistical package. *Organizational Research Methods*, *14*, 735-740. doi:10.1177/1094428109355485

Drechsler, J. (2015). Multiple imputation of multilevel missing data—Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*, 69-95. doi:10.3102/1076998614563393

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*, 39-55. doi:10.1037/a0035314

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. doi:10.1037/met0000063

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121-138. doi:10.1037/1082-989X.12.2.121

Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (in press). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*. doi:10.1002/sim.6944

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472. doi:10.1214/ss/1177011136

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *7*, 473-483. doi:10.1214/ss/1177011137

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (3rd ed.). Boca Raton, FL: CRC Press.

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*, 553-564. doi:10.1111/rssa.12022

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206-213. doi:10.1007/s11121-007-0070-9

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. doi:10.1037/1082-989X.11.4.323

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, *48*, 640-649. doi:10.3758/s13428-015-0590-3

Grund, S., Lüdtke, O., & Robitzsch, A. (in press). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*.

Grund, S., Robitzsch, A., & Lüdtke, O. (2016). mitml: Tools for multiple imputation in multilevel modeling (Version 0.3-2) [Computer software]. Retrieved from http://CRAN.R-project.org/package=mitml

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*, 623-641. doi:10.1177/014920639802400504

Hofmann, D. A., Morgeson, F. P., & Gerras, S. J. (2003). Climate as a moderator of the relationship between leader-member exchange and content specific citizenship: Safety climate as an exemplar. *Journal of Applied Psychology*, *88*, 170-178. doi:10.1037/0021-9010.88.1.170

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Hoboken, NJ: Wiley.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1-21. doi:10.1207/s15327906mbr3001_1

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, *17*, 494-509. doi:10.1016/j.learninstruc.2007.09.002

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*, 418-435. doi:10.1177/1094428107308984

Li, K. H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*, 65-92. Retrieved from http://www.stat.sinica.edu.tw/statistica/

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*, 1065-1073. doi:10.1080/01621459.1991.10475152

Licht, C. (2010). *New methods for generating significance levels from multiply-imputed data* (Doctoral dissertation, Universität Bamberg, Germany). Retrieved from http://d-nb.info/101104966X/34

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229. doi:10.1037/a0012869

Lüdtke, O., Robitzsch, A., & Grund, S. (in press). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*, 538-558. doi:10.1214/ss/1177010269

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*, 103-111. doi:10.1093/biomet/79.1.103

Mistler, S. A. (2013). A SAS macro for applying multiple imputation to multilevel data. In *Proceedings of the SAS Global Forum*. Retrieved from http://support.sas.com/

Mistler, S. A. (2015). *Multilevel multiple imputation: An examination of competing methods* (Doctoral dissertation, Arizona State University, Tempe). Available from http://repository.asu.edu/

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill, MA: Boston College.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*, 372-411. doi:10.1177/1094428114548590

Quartagno, M., & Carpenter, J. R. (2016). jomo: A package for multilevel joint modeling multiple imputation (Version 2.1-2) [Computer software]. Retrieved from http://CRAN.R-project.org/package=jomo

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 621-637. doi:10.1207/s15328007sem1104_7

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.1) [Computer software]. Available from http://www.R-project.org/

Reiter, J. P. (2007). Small-sample degrees of freedom for multicomponent significance tests with multiple imputation for missing data. *Biometrika*, *94*, 502-508. doi:10.1093/biomet/asm028

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*, 1462-1471. doi:10.1198/016214507000000932

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 477-494. doi:10.1080/10705511.2012.687669

Schafer, J. L. (2001). Multiple imputation with PAN. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 357-377). Washington, DC: American Psychological Association.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*, 19-35. doi:10.1111/1467-9574.00218

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177. doi:10.1037//1082-989X.7.2.147

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437-457. doi:10.1198/106186002760180608

Schafer, J. L., & Zhao, J. H. (2014). pan: Multiple imputation for multivariate panel or clustered data (Version 0.9) [Computer software]. Retrieved from http://CRAN.R-project.org/package=pan

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*, 1171-1177. doi:10.2307/2533455

Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, *50*, 329-345. doi:10.1002/bimj.200710423

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox (Ed.), *Handbook of advanced multilevel analysis* (pp. 173-196). New York, NY: Routledge.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67. Available from http://www.jstatsoft.org/

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377-399. doi:10.1002/sim.4067

Wu, L. (2010). *Mixed effects models for complex data*. Boca Raton, FL: CRC Press.

Author Biographies

Simon Grund graduated in Psychology at the Humboldt University in Berlin and joined the Leibniz Institute for Science and Mathematics Education as a PhD candidate. His research interests include missing data, causality, and multilevel modeling.

Oliver Lüdtke is a professor of Educational Measurement at the Leibniz Institute for Science and Mathematics Education and the University of Kiel. His main research interests include the application of multilevel modeling in psychological and educational research, international student achievement studies, and personality development in adolescence.

Alexander Robitzsch graduated in Mathematics at the University of Technology in Dresden and earned a PhD in Educational Science at the Humboldt-University in Berlin. His research interests include item response modelling, missing data, and multilevel modeling.