“Multiple Imputation of Missing Data in Multilevel Research”

Dissertation zur Erlangung des Doktorgrades der Philosophischen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Simon Grund

Kiel, den 12. September 2017

Erstgutachter: Prof. Dr. Oliver Lüdtke

Zweitgutachter: Prof. Dr. Olaf Köller

Drittgutachter: Prof. Dr. Benjamin Nagengast

Tag der mündlichen Prüfung: 15.11.2017

Durch den zweiten Prodekan, Prof. Dr. Elmar Eggert zum Druck genehmig am: 20.02.2018

Contents

Acknowledgments

My deepest gratitude goes to my supervisor, Oliver Lüdtke, for his guidance and support, for motivating me when things were slow and reassuring me when times were tough. I could not have wished for a better teacher. In that spirit, I also want to thank Alexander Robitzsch for his valuable advice and for showing me what is possible with diligence and dedication. *Bedanken möchte ich mich auch bei meiner Familie, meiner Schwester, meinen Großeltern und Eltern, die mir nicht nur das Leben geschenkt, sondern mich auf meinem Weg stets bekräftigt, in vielerlei Hinsicht unterstützt und immer an mich geglaubt haben. Ohne sie wäre diese Arbeit nicht möglich gewesen.*

I would also like to thank Steffen Zitzmann, with whom I shared not only an office but also many fruitful and inspiring discussions over the years. Furthermore, I want to thank (in no particular order) Julian, Dennis, and Jan for making me run the courts, Anne and Daniel for helping me overcome my fear of heights, Marlit and Christoph for lovely conversations and good food, Jakob and Nils for rolling the dice with me, Johanna, Fabi, Jenny, and Julia for making us all drink more than we should, and all the other PhD students, friends, and colleagues who made this time unforgettable. *Last but not least, I want to thank you, Bénédicte, for being my counsel, my pillar in need, my attorney against the odds, and my partner in crime. Your love and support was invaluable and has helped me in more ways than I can count.*

Abstract

Multilevel models have become one of the most frequently used statistical models for analyzing multilevel data. These types of data occur in many fields of psychology when observations (Level 1) are clustered within some higher-level collectives (Level 2). This includes, for example, students nested in schools, employees nested in work teams, patients nested in clinics, and longitudinal data, in which observations are nested within persons. Unfortunately, multilevel data often contain missing data, for example, when participants omit certain items in a questionnaire or they drop out before the end of a study. If treated improperly, missing data can severely distort parameter estimates and may compromise statistical decision making. For this reason, it is often recommended to rely on principled methods for dealing with missing data such as multiple imputation (MI) or maximum likelihood estimation (ML). These procedures have the advantage that they take all the available data into account, thus improving statistical power and the conclusions that can be drawn from the data.

In the present dissertation, I consider different procedures for the treatment of missing data with an emphasis on multilevel MI. In multilevel research, it is important that the imputation model takes the structure of the data and the features of the substantive analysis model into account. However, many open questions remain about how this can be achieved in practice. In the present dissertation, I consider a variety of applications of multilevel models as well as different implementations of multilevel MI. In multiple studies, I examined how the multilevel structure is represented in different implementations of multilevel MI, how different representations may

affect the results obtained from MI, and how missing data can be treated in multilevel models with random intercepts, random slopes, interaction effects, continuous and categorical data, and missing data at Level 2.

In addition, the present dissertation was concerned with the analysis of multiply imputed data sets. In this context, I examined different procedures for pooling the results obtained from multiply imputed data sets with an emphasis on multiparameter tests (e.g., model comparisons). This includes applications in traditional research designs with the analysis of variance (ANOVA) as well as applications in multilevel models with hypothesis tests about fixed effects and variance components. Finally, the dissertation presents the R package `mitml`, which is intended to provide researchers with a set of practical tools for conducting multilevel MI in research practice. This includes tools for the specification of the imputation model, convergence diagnostics, managing and analyzing multiply imputed data sets, and pooling methods for single- and multiparameter tests along with a tutorial article that illustrates these features and provides a nontechnical introduction to multilevel MI.

1
Introduction

Over the past years, multilevel models have become a standard tool for analyzing clustered data (e.g., Goldstein, 2011; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012b). Such data structures often occur in psychological research when observations (Level 1) are clustered within higher-level collectives (Level 2), for example, when students are nested within schools, employees are nested within enterprises, or in longitudinal or repeated-measures data when measurement occasions are nested within persons. In addition, psychological data are often incomplete, for example, when participants omit some of the items in a questionnaire or drop out before the end of a study. It is well known that simple methods for dealing with missing data such as listwise deletion (LD) can lead to biased parameter estimates and an inefficient use of the data (i.e., low statistical power). Fortunately, principled methods for the treatment of missing data such as multiple imputation (MI) or maximum likelihood estimation (ML) have become widely available (for an overview, see Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002).

Although a large body of research has concerned itself with missing data in general, the treatment of missing data in multilevel research is still not well understood. Consequently, missing data in multilevel research are most commonly treated with ad-hoc procedures such as LD instead of principled methods such as MI and ML (e.g., Diaz-Ordaz, Kenward, Cohen, Coleman, & Eldridge, 2014; Jelicic, Phelps, & Lerner, 2009; Nicholson, Deboeck, & Howard, 2017; Peugh & Enders, 2004). To provide an additional illustration, I conducted a software-

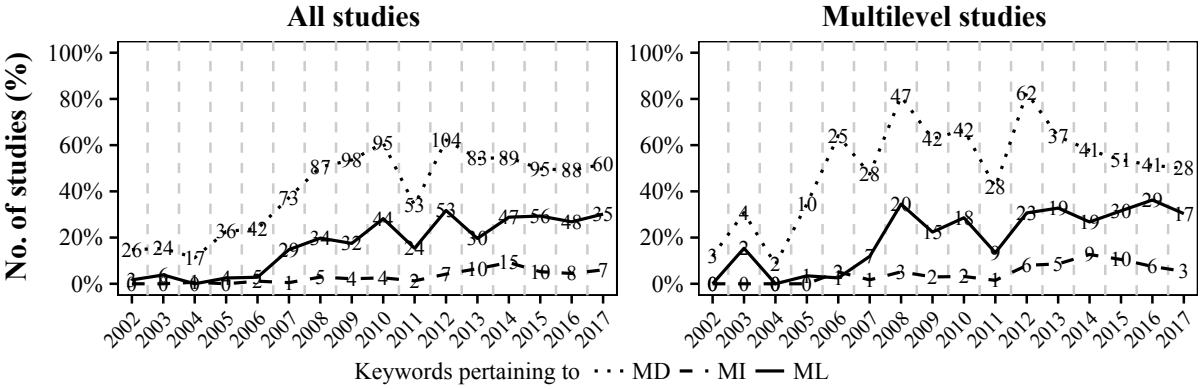

Figure 1.1: Number of studies (in %) identified in computer-assisted literature review with keywords pertaining to missing data (MD) and the treatment thereof using multiple imputation (MI) or maximum likelihood (ML). The numbers within each plot denote the absolute numbers of studies.

assisted literature review on the basis of the articles published in the *Journal of Education Psychology* and the *Journal of Applied Psychology* within the last 15 years ($n = 2,652$). Using a self-written program, I searched these articles for occurrences of keywords pertaining to multilevel models, missing data, as well as the treatment thereof with ML and MI. The results are shown in Figure 1.1. It is easy to see that the reporting of missing data has improved over the years; however, relatively few studies seem to use principled methods for dealing with them such as ML and, perhaps most noticeably, MI. The present dissertation considers this topic in detail and attempts to (a) contribute to the growing literature on missing data in multilevel research and (b) provide researchers with a set of clear-cut advice and practical tools for the application of multilevel MI.

The dissertation is structured as follows. Chapter 1 reviews the theoretical background of missing data, MI, and ML without particular emphasis on multilevel data. Chapter 2 then focuses on multilevel MI, considering applications in the context of (a) multilevel random intercept models, (b) the random coefficients model, with an emphasis of random slopes and cross-level interactions (CLIs), and (c) missing data at Level 2. Chapter 3 considers the analysis of multiply imputed data sets with an emphasis on multiparameter tests and model comparisons. Chapter 4 then introduces the R package `mitml`, which is intended to provide researchers with a simple and effective workflow for conducting and performing analyses with multilevel MI. Chapter 5 closes with a discussion and presents an outlook on possible topics for future research. The present dissertation also provides a motivation for the five research articles that have been

written as the dissertation progressed. These articles are provided in the Appendix.

1.1 Missing data

It is well known that an inadequate treatment of missing data can have adverse effects on statistical decision making (e.g., Allison, 2001; Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002). For example, when analyses are based on only the complete cases (LD), and data are missing in a systematic manner, then parameter estimates can be biased, statistical power can be low, and the generalizability of one's findings can be compromised. To gain a more thorough understanding of when and how missing data affect statistical analyses, it is useful to distinguish different *mechanisms* and different *patterns* of missing data, where missing data mechanisms describe the relation between the (hypothetical) complete data and the occurrence of missing data, and the patterns of missing data describe how missing data have manifested themselves in a given data set.

1.1.1 *Missing data mechanisms*

Rubin (1976) distinguished three broad classes of missing data mechanisms. Let $\mathbf{Y}$ denote the hypothetical complete data set, which can be decomposed in an observed and an unobserved portion, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, and let $\mathbf{R}$ be an indicator matrix that denotes which elements are missing in $\mathbf{Y}$. Rubin considered data to be missing at random (MAR), if the probability of missing data $P(\mathbf{R})$ is independent of the unobserved data $\mathbf{Y}_{mis}$ given the observed $\mathbf{Y}_{obs}$, that is, $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}|\mathbf{Y}_{obs})$. Put differently, under MAR, once the observed data are taken into account, there remains no link between the chance of observing data and the data themselves. As a special case, the data can be missing *completely* at random (MCAR) if missing data occur in a manner that is completely independent of both the observed and unobserved data $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$, that is, $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R})$. These two missing data mechanisms are often referred to as "ignorable"[1] because the missing data mechanism need not be known in order to obtain valid

[1]For simplicity, I use the term "ignorable" as equivalent with MAR. However, the formal definition of "ignorability" also requires that the missing data mechanism and the distribution of the data are governed by two distinct sets of parameters (Schafer, 1997).

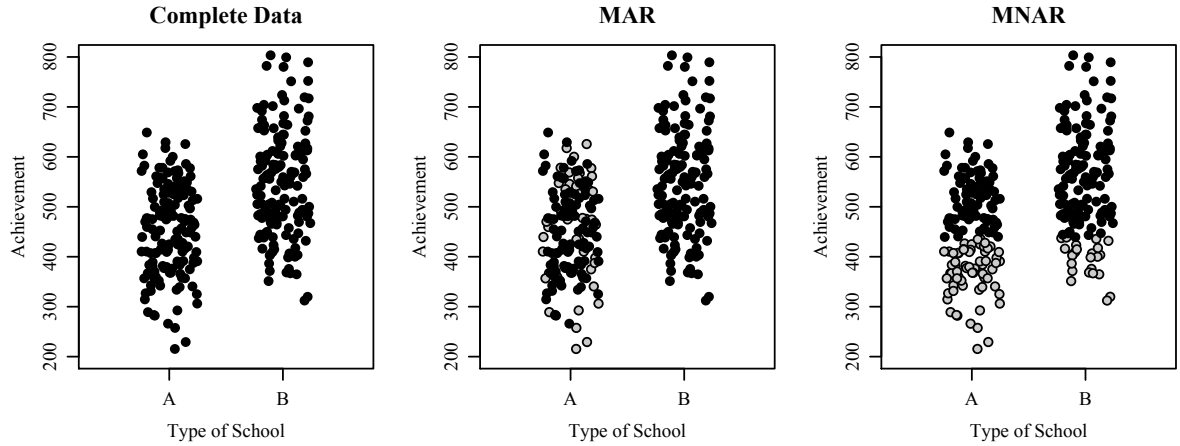

Figure 1.2: Example for different missing data mechanisms. Left panel = complete data. Middle panel = missing at random (conditional on school type). Right panel = missing not at random (conditional on achievement). Adapted from Carpenter and Kenward (2013).

statistical inferences from the observed data (see also Section 1.2). By contrast, Rubin considers data to be missing *not* at random (MNAR) when this condition is violated, that is, missing data occur in a manner that is dependent on the unobserved data $\mathbf{Y}_{mis}$ even after controlling for the observed $\mathbf{Y}_{obs}$. The notion of ignorability in missing data theory is conceptually similar to that in Rubin's causal model (Holland, 1986), where it refers to the mechanism for treatment assignment in nonrandomized observational studies (Rubin, 1977, 2005).

To provide an illustration, consider Figure 1.2. Assume that a researcher has obtained two samples of students from different school types (A and B), each of size $n_A = n_B = 150$, in order to estimate students' overall academic achievement μ across school types. The left panel of Figure 1.2 shows the complete data. Clearly, the two school types differ in terms of achievement, where achievement scores tend to be higher in school type B. Based on the complete data, an unbiased estimator of μ is the overall mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{300}\,(359.6 + 426.5 + \ldots) \approx 500.5\,, \tag{1.1}$$

where n is the total sample size. In the middle panel, achievement scores are missing at random (MAR) as a function of school type. Specifically, one third of the scores in school type A were deleted but those in school type B were complete. As a result, the propensity of missing data varies systematically with school types, and the overall mean is no longer unbiased

$$\bar{x}_{obs} = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} x_{i,obs} = \frac{1}{100 + 150} (426.5 + 481.6 \ldots) \approx 511.0 \, . \tag{1.2}$$

However, because the data are MAR given school type, that is, missing data occur completely at random within school types A and B, an unbiased estimator can be obtained on the basis of conditional distribution of achievement given school type. As implied by the factorization $P(x) = P(x|A)P(A) + P(x|B)P(B)$, the estimator is given by

$$\bar{x}'_{obs} = \frac{1}{2} \left(\bar{x}_{obs|A} + \bar{x}_{obs|B} \right) = \frac{1}{2} \left(\frac{1}{100} (426.5 + \ldots) + \frac{1}{150} (406.6 + \ldots) \right) \approx 501.6 \, . \tag{1.3}$$

Finally, in the right panel in Figure 1.2, achievement scores are missing not at random (MNAR) as a function of achievement (i.e., the bottom third of the achievement scores were deleted). I do not consider this case in detail; however, it should be immediately obvious that an unbiased estimate of the overall mean cannot be obtained with only the data at hand and without making specific assumptions about the missing data mechanism.

Perhaps more subtly, this example also illustrates that the consequences of the missing data mechanism may depend on the substantive analysis model (see also Carpenter & Kenward, 2013). Specifically, in the example above, student achievement is missing at random (MAR) given school type. As a result, the parameters of the conditional distribution of student achievement given school type (e.g., the regression coefficient) can be estimated without bias from only the observed data. In other words, even though the overall mean of academic achievement based only on the observed data, the same is not true for the regression of academic achievement on school type. Generally speaking, the consequences of the missing data mechanism (e.g., in terms of bias) depend on the substantive analysis model (for further discussion, see Carpenter & Kenward, 2013; see also Little, 1992; von Hippel, 2007).

In practice, the notion of missing data mechanism can be useful because it allows expressing conditions under which a treatment for missing data provides biased or unbiased parameter estimates. For example, LD provides generally unbiased estimates only under MCAR, whereas procedures such as MI and ML can provide unbiased estimates even under MAR. The example above also illustrates the need for *auxiliary* variables, that is, variables that are related to either the propensity of missing data or the missing data themselves, because the inclusion of such variables can increase the plausibility of the MAR assumption (Collins, Schafer, & Kam,

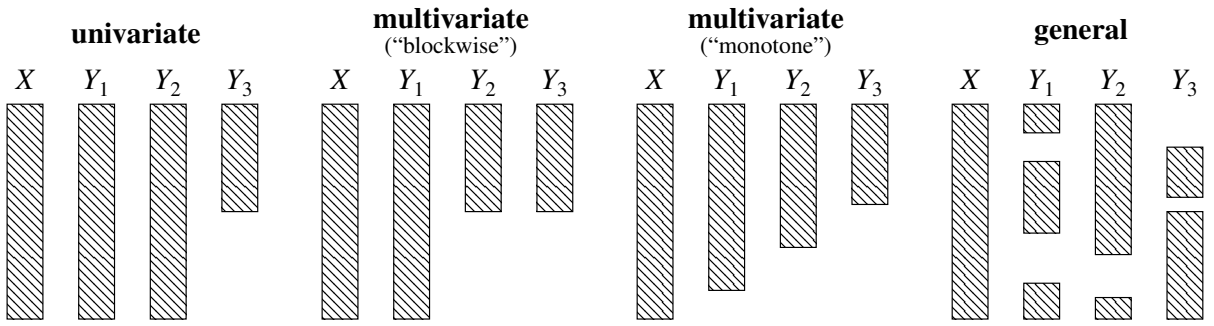

Figure 1.3: Illustration for different patterns of missing data.

2001; see also Section 1.1.2). Although this assumption can never be tested on the basis of the observed data alone (e.g., Enders, 2010), the data can be used to discern more from less plausible assumptions about the missing data mechanism. For example, auxiliary variables may be identified by conducting logistic regression analyses, where missing data indicators are regressed on (potential) auxiliary variables (e.g., Carpenter & Kenward, 2013; White, Royston, & Wood, 2011). In addition, graphical representations of missing data mechanisms can be used to express and evaluate potential mechanisms from a theoretical point of view (Thoemmes & Mohan, 2015; Thoemmes & Rose, 2014).

1.1.2 *Patterns of missing data*

In addition to missing data mechanisms, it is often useful to consider the patterns of missing data in a given data set. For example, Little and Rubin (2002) distinguish between univariate and multivariate patterns of missing data, in which one or several variables contain missing values. In addition, it is often useful to distinguish item and unit nonresponse, in which all data for a given unit are missing apart from some (known) background information (for a similar distinction, see also Newman, 2014). Finally, the scores of latent variables are sometimes considered a special case of missing data (see also Blackwell, Honaker, & King, 2017b; Mislevy, 1991). In multilevel research, in which variables can be measured at different levels of the sample, both the patterns and the adverse effects of missing data can extend to multiple levels (see Chapter 2).

Examples for common patterns of missing data are provided in Figure 1.3. In practice, missing data often follow a "general" pattern with missing values on multiple variables and

several sections of overlapping and non-overlapping "missingness" between variables (Figure 1.3, right panel). In applications of modern methods such as ML and MI, understanding the patterns of missing data in a given data set can be extremely helpful because it allows to identify further auxiliary variables, that is, variables that are (a) predictive of other variables with missing data and (b) observed when these variables are missing (i.e., non-overlapping "missingness"). For example, in the "general" pattern in Figure 1.3, both X and Y_2 may be considered as auxiliary variables for Y_3, provided that the observed values in these variables are predictive of the missing values in Y_3. By selecting a useful set of auxiliary variables, for example, by inspecting pairwise correlations between variables and patterns of missing data, modern methods for dealing with missing data can make better use of the information contained in the observed data, thus increasing the efficiency and statistical power of subsequent analyses (e.g., Collins et al., 2001).

1.2 Inference with missing data

The goal of statistical inference is to estimate population quantities on the basis of empirical data (e.g., Wasserman, 2004). However, in the presence of missing data, statistical inference can be challenging. For example, when data are missing in a systematic fashion (e.g., MAR) and only the complete cases are analyzed (LD), then parameter estimates can be biased, and statistical inferences may no longer apply to the entire target population (Little & Rubin, 2002). In other words, statistical inference is complicated by missing data because the observed data are no longer generated only by the parameters of the population model, say $\boldsymbol{\theta}$, but also by the mechanism that generated the missing data, say $\boldsymbol{\xi}$ (see above; see also Little & Rubin, 2002; Schafer, 1997).

An illustration is provided in Figure 1.4. With complete data, inference about $\boldsymbol{\theta}$ can conducted on the basis of the likelihood of the data given $\boldsymbol{\theta}$, $P(\mathbf{Y}|\boldsymbol{\theta})$. By contrast, with incomplete data, only $\mathbf{R}$ and $\mathbf{Y}_{obs}$ are observed, and the joint distribution of the data is governed by both $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. Specifically, the joint distribution can be written as

$$P(\mathbf{R}, \mathbf{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \int P(\mathbf{R}, \mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\xi})\, d\mathbf{Y}_{mis} \,. \tag{1.4}$$

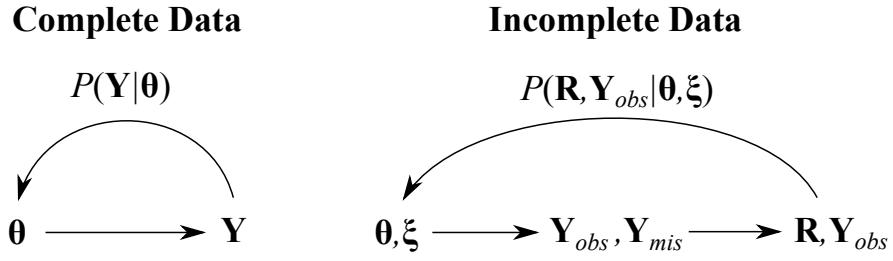

Figure 1.4: Illustration for statistical inference with complete and incomplete data.

This expression is difficult to evaluate in general. However, under the assumption that (a) the data are MAR and (b) $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ denote two distinct sets of parameters, it can be simplified to

$$\begin{aligned} P(\mathbf{R}, \mathbf{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{\xi}) &= \int P(\mathbf{R}|\mathbf{Y}_{obs}, \boldsymbol{\xi}) P(\mathbf{Y}|\boldsymbol{\theta})\, d\mathbf{Y}_{mis} \\ &= P(\mathbf{R}|\mathbf{Y}_{obs}, \boldsymbol{\xi}) \int P(\mathbf{Y}|\boldsymbol{\theta})\, d\mathbf{Y}_{mis} \end{aligned}, \tag{1.5}$$

where the first factor pertains to the missing data mechanism, and $\int P(\mathbf{Y}|\boldsymbol{\theta})\, d\mathbf{Y}_{mis} \equiv P(\mathbf{Y}_{obs}|\boldsymbol{\theta})$ is the likelihood of the observed data that is obtained by "integrating out" the missing data $\mathbf{Y}_{mis}$. This expression illustrates that, under MAR, inference about $\boldsymbol{\theta}$ can be carried out without considering the missing data mechanism (the missing data mechanism is "ignorable"). For this reason, $P(\mathbf{Y}_{obs}|\boldsymbol{\theta})$ is also referred to as the "likelihood ignoring the missing data mechanism" (e.g., Little & Rubin, 2002). In practice, there are two statistical procedures that are often considered as the "state of the art" for conducting statistical inferences on the basis of incomplete data: multiple imputation (MI) and maximum-likelihood estimation (ML). In the following, I provide a general introduction to MI and ML with an emphasis on single-level data. The application of MI (and to a lesser extent ML) to multilevel data is considered in detail thereafter.

1.3 Multiple imputation

The idea behind MI is to replace missing data with an "informed guess" by drawing repeatedly from the posterior predictive distribution of the missing data, given the observed data and a statistical model (Rubin, 1987). The data sets completed in this manner are then analyzed separately, and the results are pooled using the rules in Rubin (1987; see also Chapter 3). Multiple imputation is related to Bayesian inference with incomplete data (see also Carpenter & Kenward, 2013). In the Bayesian paradigm, inference about $\boldsymbol{\theta}$ can be conducted on the basis

of the observed-data posterior distribution, $P(\boldsymbol{\theta}|\mathbf{Y}_{obs})$. Because it is usually difficult to sample from $P(\boldsymbol{\theta}|\mathbf{Y}_{obs})$, the missing data $\mathbf{Y}_{mis}$ are often regarded as additional (nuisance) parameters. The joint posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{Y}_{mis}$ is given by

$$P(\boldsymbol{\theta}, \mathbf{Y}_{mis}|\mathbf{Y}_{obs}) = P(\boldsymbol{\theta}|\mathbf{Y}_{mis}, \mathbf{Y}_{obs})P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}) \ . \tag{1.6}$$

The marginal posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}_{obs}$ is then given by

$$P(\boldsymbol{\theta}|\mathbf{Y}_{obs}) = \int P(\boldsymbol{\theta}, \mathbf{Y}_{mis}|\mathbf{Y}_{obs}) \, d\mathbf{Y}_{mis} \ , \tag{1.7}$$

which can be regarded as the Bayesian equivalent to the observed-data likelihood in Equation 1.5 (see also Little & Rubin, 2002; Schafer, 1997). This idea to treat $\mathbf{Y}_{mis}$ as a set of nuisance parameters is also referred to as "data augmentation" (Tanner & Wong, 1987).

Data augmentation. The data augmentation algorithm is a Markov chain Monte Carlo (MCMC) technique that simulates from the distribution $P(\boldsymbol{\theta}, \mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ by iterating between a posterior or P-step and an imputation or I-step. At iteration t,

$$\begin{array}{ll} \boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(t)}) & \text{(P-step)} \\ \mathbf{Y}_{mis}^{(t+1)} \sim P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \boldsymbol{\theta}^{(t+1)}) & \text{(I-step)} \end{array} \ , \tag{1.8}$$

The resulting sequence converges in distribution to $P(\boldsymbol{\theta}, \mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ as $t \to \infty$.[2] This algorithm can be used to generate multiple, say M, imputations for the missing data, resulting in M copies of the original data with missing values "filled in" by the imputed data. Because small to moderate numbers of imputations (e.g., 5 to 100) are common, MI can be considered an approximation of Bayesian inference, based on only a small number of posterior draws (Carpenter & Kenward, 2013). However, MI also provides estimates with good frequentist properties (e.g., coverage) when analyzed with non-Bayesian methods (e.g., Rubin & Schenker, 1986). In this context, MI can be regarded as a sampling-based procedure for conducting inferences on the basis of incomplete data, that is, for "integrating out" the missing data by averaging over a predictive distribution of the missing data, given the observed data and a statistical model (e.g., Schafer,

[2]Note that different statistical models are often used for the analysis and the imputation of empirical data. If the two models differ, the technical requirement for inferences to remain valid is that the two models are "congenial" in the sense of Meng (1994; for further discussion, see Carpenter & Kenward, 2013; Schafer, 2003).

1999). In practice, MI can be implemented in a number of ways, where two implementations are particularly popular in current statistical software: joint modeling (JM) or the fully conditional specification (FCS).

1.3.1 *Joint modeling*

In the joint modeling (JM) approach, imputations are generated from a single statistical model for all variables simultaneously. For example, with multivariate normal data in $\mathbf{Y}$, imputations can be generated from the following model. For case i ($i = 1, \ldots, n$),

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{e}_i \,, \tag{1.9}$$

where $\boldsymbol{\mu}$ is a vector of means, and $\mathbf{e}_i$ is a vector of residuals which follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. This model allows for (linear) relations between all variables as implied by the multivariate normal distribution. In addition, it is possible to include completely observed predictor variables in the imputation model (for further details, see Schafer, 1997; Schafer & Olsen, 1998).

Categorical data. Even though the model above is restricted to continuous (i.e., multivariate normal) data, it can be extended to accommodate ordinal and (unordered) categorical variables with missing data. For example, for a categorical variable with c categories, the model may include a set of $c - 1$ latent continuous background variables that represent the differences between categories and which may be correlated with the other variables (Carpenter & Kenward, 2013; Schafer, 1997). Similarly, for an ordinal variable with c categories, it is possible to include a single background variable, where the differences between categories are represented by a set of $c - 1$ threshold parameters (Asparouhov & Muthén, 2010b). For individual variables, these model are equivalent to conventional generalized linear models for categorical data (e.g., Agresti, 2013; Fahrmeir & Tutz, 2010). To provide some general insight into JM, I briefly describe a sampling algorithm that can be used for generating imputations under the assumption of the multivariate normal distribution.

Sampling algorithm. In JM, imputations are generated in two steps. First, the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, are drawn from their posterior distributions, given $\mathbf{Y}_{obs}$ and current imputations for $\mathbf{Y}_{mis}$ (P-step). Second, new imputations for $\mathbf{Y}_{mis}$ are generated on the basis of

$\boldsymbol{\theta}$ and $\mathbf{Y}_{obs}$ (I-step). Specifically, let r denote the number of variables and let any $\mathbf{y}_i$ be missing in arbitrary patterns, $\mathbf{y}_i = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$. Under "flat" priors for $\boldsymbol{\mu}$, an inverse-Wishart prior $\boldsymbol{\Sigma} \sim W^{-1}(\nu, \boldsymbol{\Delta}^{-1})$ with $\nu \geq r$, and given a set of starting values, imputations are generated as follows. At iteration t,

1. P-step: Update $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as follows.
   i) Draw $\boldsymbol{\Sigma}^{(t+1)} \sim W^{-1}(\nu + n, \boldsymbol{\Delta}^{-1} + \mathbf{S}^{(t)})$, where $\mathbf{S}^{(t)} = \sum_{i=1}^{n}(\mathbf{y}_i^{(t)} - \bar{\mathbf{y}}^{(t)})^T(\mathbf{y}_i^{(t)} - \bar{\mathbf{y}}^{(t)})$ with mean vector $\bar{\mathbf{y}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^{(t)}$ and $\mathbf{y}_i^{(t)} = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,imp}^{(t)})$.
   ii) Draw $\boldsymbol{\mu}^{(t+1)} \sim N(\bar{\mathbf{y}}^{(t)}, \frac{1}{n}\boldsymbol{\Sigma}^{(t+1)})$ with $\bar{\mathbf{y}}^{(t)}$ as above .
2. I-step: Update $\mathbf{y}_{i,mis}$ as follows.
   i) Calculate the conditional mean of $\mathbf{y}_{i,mis}$ given $\mathbf{y}_{i,obs}$ as follows.
      a) $\tilde{\boldsymbol{\mu}}_{i,mis|obs}^{(t+1)} = \boldsymbol{\mu}_{i,mis}^{(t+1)} + \boldsymbol{\Sigma}_{i,mis,obs}^{(t+1)}\left[\boldsymbol{\Sigma}_{i,obs}^{(t+1)}\right]^{-1}(\mathbf{y}_{i,obs} - \boldsymbol{\mu}_{i,obs}^{(t+1)})$
      b) $\boldsymbol{\Sigma}_{i,mis|obs}^{(t+1)} = \boldsymbol{\Sigma}_{i,mis}^{(t+1)} - \boldsymbol{\Sigma}_{i,mis,obs}^{(t+1)}\left[\boldsymbol{\Sigma}_{i,obs}^{(t+1)}\right]^{-1}\boldsymbol{\Sigma}_{i,obs,mis}^{(t+1)}$
   ii) Draw $\mathbf{e}_{i,mis}^{(t+1)} \sim N(0, \boldsymbol{\Sigma}_{i,mis|obs}^{(t+1)})$ and impute $\mathbf{y}_{i,mis}^{(t+1)} = \tilde{\boldsymbol{\mu}}_{i,mis|obs}^{(t+1)} + \mathbf{e}_{i,mis}^{(t+1)}$.

This illustrates that imputations in JM indeed rely on the joint distribution of the data: Based on the (joint) multivariate normal distribution for $\mathbf{y}_i$, imputations for the missing data $\mathbf{y}_{i,mis}$ are drawn from the (conditional) normal distribution of the missing data, given the observed data $\mathbf{y}_{i,obs}$.

1.3.2 *Fully conditional specification*

As an alternative to JM, it has been suggested to approximate the joint distribution of the data with a sequence of univariate, conditional models (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren, 2012; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). This is referred to as the "fully conditional specification" (FCS) of MI and also known as "chained equations" or "sequential MI". For example, if the joint distribution of the data is multivariate normal, then imputations can be generated in a sequence of regression models with normally distributed residuals. Specifically, with multivariate normal data in $\mathbf{Y}$, imputations can be generated as follows. For case i ($i = 1, \ldots, n$) and variable p ($p = 1, \ldots, r$),

$$y_{ip} = \mathbf{y}_{i(-p)}\boldsymbol{\beta}_p + e_{ip}\,, \tag{1.10}$$

where $\mathbf{y}_{i(-p)}$ denotes the predictor variables in the p-th imputation model, including all variables other than y_{ip} and a "one" for the regression intercept, $\boldsymbol{\beta}_p$ denotes the regression coefficients for $\mathbf{y}_{i(-p)}$, and e_{ip} is a normally distributed residual with mean zero and variance σ_p^2. Note that the FCS approach differs from JM in that it considers only one variable at a time. To address multivariate patterns of missing data, the FCS approach iterates back and forth between variables, including the most recent imputations for missing values in $\mathbf{y}_{i(-p)}$ at each iteration. Similar to JM, the FCS approach also acknowledges the relations that exist between variables; however, it does so by repeatedly conditioning variables on one another, thus approximating the joint distribution used in JM. In principle, however, the imputation model for each variable in FCS may include only a subset of the variables in $\mathbf{y}_{i(-p)}$ as well as transformations of these variables or nonlinear effects.

Categorical data. Similar to JM, the FCS approach is able to accommodate ordinal and (unordered) categorical data using generalized linear models (e.g., Agresti, 2013). For example, if y_{ip} is a (unordered) categorical variable with missing data, the imputation model may be a multinomial logistic or probit model, conditional on $\mathbf{y}_{i(-p)}$. Similarly, if y_{ip} is ordinal, the imputation model may be an ordered logistic or probit model (Brand, 1999; van Buuren et al., 2006). To provide further insights into the FCS approach, I briefly describe a sampling algorithm that can be used for multivariate normal data (see also Rubin, 1987; van Buuren et al., 2006).

Sampling algorithm. In contrast to JM, the FCS approach iterates across variables with missing data, employing the P- and I-step separately for each variable. Specifically, let any y_{ip} be partially missing, $y_{ip} = (y_{ip,obs}, y_{ip,mis})$, and let $n_{p,obs}$ denote the number of cases with y_{ip} observed. Under "flat" priors for $\boldsymbol{\beta}_p$ and σ_p^2 (Box & Tiao, 1973)[3] and given a set of starting values, imputations are generated as follows. At iteration t, for variable p,

[3]Note that this choice of priors places a flat, uniform density on both $\boldsymbol{\beta}_p$ and $\log \sigma_p$ (see also Jeffreys, 1961). It is presented here mostly for consistency with the published literature (e.g., Rubin, 1987) and software implementations (van Buuren & Groothuis-Oudshoorn, 2011). As an alternative, any standard conjugate prior can be used, for example, the scaled inverse-χ^2 with user-defined prior parameters (see Chapter 2; for a general discussion, see also Gelman et al., 2014).

1. Estimate $\hat{\boldsymbol{\beta}}_p^{(t)}$ and $\hat{\sigma}_p^{2\,(t)}$ from the regression model $y_{ip,obs} = \mathbf{y}_{i(-p)}^{(t)}\boldsymbol{\beta}_p + e_{ip}$ using the cases with y_{ip} observed, with $\mathbf{y}_{i(-p)}^{(t)} = \left(\mathbf{y}_{i(-p),obs}, \mathbf{y}_{i(-p),mis}^{(t)}\right)$.
2. P-step: Update $\boldsymbol{\theta}_p = (\boldsymbol{\beta}_p, \sigma_p^2)$ as follows.
   i) Draw $\sigma_p^{2\,(t+1)} \sim \text{inv-}\chi^2(n_{p,obs} - k_p, \hat{\sigma}_p^{2\,(t)})$, where k_p is the number of variables in $\mathbf{y}_{i(-p)}^{(t)}$ (usually r).
   ii) Draw $\boldsymbol{\beta}^{(t+1)} \sim N(\hat{\boldsymbol{\beta}}_p^{(t)}, \sigma_p^{2\,(t+1)}\mathbf{V}^{(t)})$, where $\mathbf{V}^{(t)} = \left(\sum_{i=1}^{n_{p,obs}} \mathbf{y}_{i(-p)}^{(t)T}\mathbf{y}_{(-p)}^{(t)}\right)^{-1}$.
3. I-step: Update $y_{ip,mis}$ as follows.
   i) Draw $e_{ip,imp}^{(t+1)} \sim N(0, \sigma_p^{2\,(t+1)})$ and impute $y_{ip,imp}^{(t+1)} = \beta_{0p}^{(t+1)} + \boldsymbol{\beta}_{1p}^{(t+1)}\mathbf{y}_{i(-p)}^{(t)} + e_{ip,imp}^{(t+1)}$.

This illustrates that the FCS approach, like the JM, draws imputations for missing data $y_{ip,mis}$ from the conditional distribution of the missing data, given the observed data $\mathbf{y}_{i(-p),obs}$ and the most recent imputations for the missing data in other variables $\mathbf{y}_{i(-p),mis}^{(t)}$. However, it does so for each variable separately, thus implementing the data augmentation algorithm on a variable-by-variable basis.

Practical considerations. In comparison with one another, the JM approach tends to be easier to use in practice because it employs a single imputation model for all variables with missing data. Consequently, standard tasks such as the specification of the model and the assessment of convergence tend to be simpler under JM. By contrast, FCS tends to be more flexible with separate imputation models for each variable. This can be advantageous in applications with a larger number of variables or categorical variables with a large number of categories; in such cases it is often easier (and potentially more stable) to carry out the imputation in a sequential manner with FCS. In addition, each imputation model may include a different set of predictor variables, thus further reducing complexity (for a similar discussion, see also Carpenter & Kenward, 2013).

1.4 Model-based procedures

Missing data can also be treated using model-based procedures, which allow parameter estimates to be obtained directly on the basis of the incomplete data (e.g., Little & Rubin, 2002). In

practice, this is often achieved by employing maximum-likelihood (ML) or Bayesian estimation procedures (e.g., Enders, 2010; Little & Rubin, 2002). In the following, I briefly discuss these two model-based procedures for the treatment of missing data.

1.4.1 *Maximum-likelihood estimation*

The general approach of ML to the treatment of missing data is to directly estimate the parameters of the model of interest by maximizing the observed-data likelihood (e.g., Little & Rubin, 2002). However, because these estimates are seldom available in closed form with incomplete data, obtaining ML estimates often requires iteration (e.g., Schafer & Graham, 2002). In the following, I discuss two approaches for obtaining estimates in this manner: the expectation-maximization (EM) algorithm and full information maximum likelihood (FIML).

EM algorithm. The EM algorithm (Dempster, Laird, & Rubin, 1977) is an iterative procedure that consists of two steps. In the expectation or E-step, the complete-data likelihood function is calculated by replacing the missing elements in the likelihood with their expected values, given $\mathbf{Y}_{obs}$ and a current set of parameter estimates $\boldsymbol{\theta}^{(t)}$. In the maximization or M-step, a new estimate $\boldsymbol{\theta}^{(t+1)}$ is determined as the value that maximizes the complete-data likelihood; iterating these steps until convergence yields the ML estimate of $\boldsymbol{\theta}$. The EM algorithm bears resemblance with MI in multiple ways (Schafer, 1997). However, instead of replacing missing values, it "emulates" the expected contribution of the unobserved data to the complete-data likelihood. For example, with multivariate normal data in $\mathbf{y}_i$, the likelihood can be written in terms of the sufficient statistics $\sum_i \mathbf{y}_i$ and $\sum_i \mathbf{y}_i^T \mathbf{y}_i$. Given current estimates $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$, the E-step calculates the conditional mean and variance of $\mathbf{y}_{i,mis}$ given $\mathbf{y}_{i,obs}$ (see Step 3 during JM) and augments the sufficient statistics with the expected contributions of $\mathbf{y}_{i,mis}$. The M-step then computes a new estimate $\boldsymbol{\mu}^{(t+1)}$ and $\boldsymbol{\Sigma}^{(t+1)}$ from the sufficient statistics.

Full information ML. As an alternative to EM, the observed-data likelihood function can often be expressed and evaluated directly on the basis of the incomplete data. This is often referred to as "direct" (Allison, 2001; Yuan & Bentler, 2000) or "full information" ML (Arbuckle, 1996; Enders, 2001; Enders & Bandalos, 2001). Under FIML, each case contributes to the likelihood function to the extent to which it has data. For example, if the model of interest

is the multivariate normal distribution, the log-likelihood for each individual case i ($i = 1, \ldots, n$) can be written as

$$\log L(\boldsymbol{\theta}|\mathbf{y}_{i,obs}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_{i,obs}| - \frac{1}{2}(\mathbf{y}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \boldsymbol{\Sigma}_{i,obs}^{-1}(\mathbf{y}_{i,obs} - \boldsymbol{\mu}_{i,obs}) - \frac{r_{i,obs}}{2}\log(2\pi)\,, \quad (1.11)$$

where the subscripts simply refer to only the observed data for each case. The observed-data log-likelihood of $\mathbf{Y}_{obs}$ is then obtained by summing the individual likelihoods, $\log L(\boldsymbol{\theta}|\mathbf{Y}_{obs}) = \sum_i \log L(\boldsymbol{\theta}|\mathbf{y}_{i,obs})$. It is important to note that, in order to address missing data with FIML, the variables with missing data must be included in the likelihood function. For example, in a regression model, the likelihood is defined only in terms of the dependent variable. If missing data occur in explanatory variables, the likelihood function must be adjusted in such a way that it includes distributional assumptions about the explanatory variables with missing data, for example, by assuming that the variables follow a multivariate normal distribution (e.g., Enders, 2010; see also Anderson, 1957).

1.4.2 *Bayesian estimation*

Instead of ML, the model of interest can also be estimated using Bayesian methods (e.g., Gelman et al., 2014). In the Bayesian paradigm, inferences then focus on the observed-data posterior distribution of the parameters in the model of interest, where the missing data are regarded as an additional set of "nuisance" parameters (see Section 1.3). In practice, Bayesian estimation of the model of interest again requires the imposition of distributional assumptions through a statistical model for the variables with missing data (i.e., a predictive distribution for the missing data; see Little & Rubin, 2002). For example, if the model of interest is a regression model with missing data in the outcome variable, then Bayesian estimation could be carried out without additional assumptions, treating the missing outcomes as additional parameters to be simulated. If missing data occur in explanatory variables, the model must be extended, for example, by making assumptions about the joint distribution of the variables (e.g., Little & Rubin, 2002) or by factoring the (joint) posterior distribution into a sequence of conditionals (Ibrahim, Chen, & Lipsitz, 2002; Ibrahim, Chen, Lipsitz, & Herring, 2005).

1.5 Summary

In the present chapter, I provided a short introduction to different procedures for the treatment of missing data. Despite the growing body of research on missing data in general, relatively little is known about the treatment of missing data in multilevel research. For example, there are a number of open questions regarding the correct use of multilevel MI (i.e., with both JM or FCS), especially with missing data in variables of different types or at different levels, and if the model of interest includes random slopes or cross-level interactions (CLIs). Perhaps more subtly, it is still not clear how best to incorporate information located at different levels of the sample (Level 1 and 2) into the imputation model and how this may be achieved using multilevel JM and FCS. Finally, little is known about the correct application and the performance of FIML in multilevel analyses (however, see Black, Harel, & McCoach, 2011).

In the following chapters, I consider some of these problems in detail, thus motivating the research articles enclosed in this dissertation. First, I consider model-based procedures and multilevel MI in applications with random intercepts, random slopes, CLIs, and different types of variables (Articles 1 and 2) as well as the particular problem of including cluster-level information in the imputation model in multilevel MI (Article 3). Then, I consider the analysis of multiply imputed data sets with an emphasis on multiparameter tests and model comparisons (Article 4). Finally, I present the R package `mitml`, which is intended to simplify both the application of MI and the analysis of multiply imputed data sets, thus promoting a regular use of multilevel MI in practice (Article 5).

2

Multiple imputation of multilevel data

In the context of multilevel data, it can be challenging to treat missing data using MI. Previous research has shown that, in order for MI to yield valid results, the multilevel structure must be taken into account during the specification of the imputation model (Andridge, 2011; Drechsler, 2015; Enders, Mistler, & Keller, 2016; Lüdtke, Robitzsch, & Grund, 2017; Taljaard, Donner, & Klar, 2008; van Buuren, 2011). However, in multilevel research, what aspects of the multilevel structure need to be considered often depends on the research question. For example, a multilevel random intercept model may include explanatory variables at Level 1 and 2. In addition, it is possible to allow for *contextual* effects (Cronbach & Webb, 1975; Firebaugh, 1978) of variables at Level 1 by including the between-group components (e.g., the group means) as additional explanatory variable (e.g., Marsh, 1987). In other applications, the model may include random slopes of explanatory variables at Level 1, thus allowing for the relations between variables at Level 1 to differ across groups, as well as cross-level interactions (CLIs) to explain some of that variation (e.g., Hofmann, Morgeson, & Gerras, 2003).

Despite the growing interest in the problems associated with missing data in multilevel research, it is still unclear how these features of multilevel data can be addressed in multilevel MI. For example, although there is a consensus in the literature that relations between variables at different levels (i.e., contextual effects) should be taken into account (for a discussion, see

Enders et al., 2016), it is not yet fully understood how the cluster-level components of variables should enter the imputation model in multilevel MI (see also Carpenter & Kenward, 2013; Resche-Rigon & White, in press). Furthermore, it is currently unclear how random slopes and interactions effects (e.g., CLIs) should best be addressed (see also Gottfredson, Sterba, & Jackson, 2017; Grund, Lüdtke, & Robitzsch, 2016a). Consequently, the following chapter is dedicated the treatment of missing data in multilevel research and the specific challenges associated with multilevel MI. I start with reviewing the structure of multilevel data and then consider the treatment of missing data for (a) the multilevel random intercept model with missing data at Level 1 and 2, and (b) multilevel models with random slopes and CLIs.

2.1 Structure of multilevel data

In multilevel research, the data are characterized by a *clustered*, *nested*, or *hierarchical* structure (e.g., Raudenbush & Bryk, 2002), for example, with individuals (e.g., students) clustered within groups (e.g., schools). In these data, observations from individuals are not independent, for example, because members of the same group are more likely to share similar traits (e.g., motivation) or be exposed to similar influences (e.g., teacher characteristics; for further discussion, see Goldstein, 2011; Snijders & Bosker, 2012b). This non-independence can be regarded as a nuisance (e.g., Hedges, 2007); however, in multilevel research, the clustered structure is itself regarded as an interesting phenomenon because it allows observing variables and relations between them at different levels (Snijders & Bosker, 2012b).

Variables measured at Level 1 and 2. Consider the example above with students nested within schools. In such a case, variables can be measured at the level of students (Level 1) and the level of schools (Level 2), for example, with questionnaires handed out to students and school principals, respectively. Consequently, research questions in multilevel designs are often concerned with the relations between variables at different levels, for example, the effects of variables at Level 2 (e.g., school type) on outcome variables at Level 1 (e.g., academic achievement) and vice versa for outcome variables at Level 2 (e.g., Croon & van Veldhoven, 2007). Further examples include multilevel mediation analyses (e.g., Croon, van Veldhoven, Peccei, & Wood, 2014; Preacher, Zyphur, & Zhang, 2010).

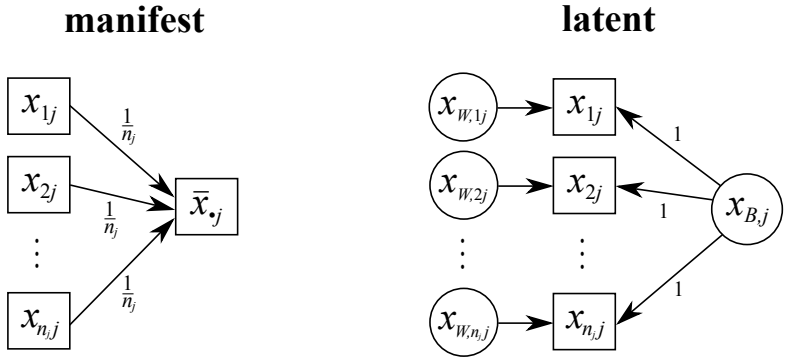

Figure 2.1: Illustration for manifest and latent specification of between-group components of variables at Level 1.

Between-group components of variables at Level 1. In multilevel data, variables measured at Level 1 can be decomposed into two independent components, where the first part varies only within groups (the within-group component) and the second part varies only between groups (the between-group component). The between- and within-group components can then be included in a multilevel analysis model, thus allowing for separate effects to be estimated at Level 1 and 2 or (alternatively) the estimation of *contextual* effects (Cronbach & Webb, 1975; see also Kreft, de Leeuw, & Aiken, 1995). In the multilevel literature, it is well known that the between-group components can be constructed in at least two different ways, as illustrated in Figure 2.1 (see also Asparouhov & Muthén, 2006; Kreft & de Leeuw, 1998; Lüdtke et al., 2008). For example, consider a single variable X, taking values x_{ij} for student i ($i = 1, \ldots, n_j$) in school j ($j = 1, \ldots, J$). This variable is typically decomposed as

$$x_{ij} = \bar{x}_{\bullet j} + (x_{ij} - \bar{x}_{\bullet j}) \,, \tag{2.1}$$

where the between-group component is represented by the group mean $\bar{x}_{\bullet j}$, and the within-group component is represented by the individual deviations from the mean $(x_{ij} - \bar{x}_{\bullet j})$. This is referred to as a *manifest* decomposition (Lüdtke et al., 2008) because the between-group component, $\bar{x}_{\bullet j}$, is directly observable in this specification: It is simply a summary measure (i.e., the average) of the individual values x_{ij} (see Figure 2.1). This represents the standard specification of between- and within-group components in multilevel analyses (e.g., Hox, 1994; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). As an alternative, the between-group components of X can be regarded as an unobservable, latent variable, for which the individual

values x_{ij} act as indicators (e.g., Croon & van Veldhoven, 2007; Grilli & Rampichini, 2011; Lüdtke et al., 2008). In such a case, the variable can be decomposed as

$$x_{ij} = x_{B,j} + x_{W,ij} \,, \tag{2.2}$$

where $x_{B,j}$ and $x_{W,ij}$ are normally distributed random variables denoting the between- and within-group components, respectively. This is illustrated in Figure 2.1 (see also Mehta & Neale, 2005). From this point of view, the between-group components $x_{B,j}$ can be regarded as a *latent* variable, of which only the indicators x_{ij} can be observed. This specification is identical to the decomposition in the one-way mixed-effects ANOVA, where $x_{B,j}$ is a random effect for group j at Level 2, and $x_{W,ij}$ is a residual at Level 1 (see also Searle, Casella, & McCulloch, 2009; Snijders & Bosker, 2012b).

In practice, it is often a matter of debate which specification of between-group components is more appropriate (for a discussion, see Lüdtke et al., 2008; Stapleton, Yang, & Hancock, 2016). In the latent specification, the observed values in each group are regarded as a finite sample from a potentially infinite population. This perspective is useful if interest lies primarily in a construct at Level 2 (e.g., school climate) that is measured at Level 1 (e.g., student ratings on school climate). In such a case, the true between-group component is unobserved (latent) and measured only through a finite number of observations at Level 1. By contrast, in the manifest specification, the observed values in each group are summarized by the group mean. This perspective is useful if interest lies primarily in a construct at Level 1 (e.g., gender), for which the group mean provides an exact summary of the construct at Level 2 (e.g., gender ratio). Critically, the two specifications provide different estimates of group-level effects: If the latent model holds in the population, the manifest mean provides only an unreliable measure of the true between-group component, whereas the latent mean corrects for that unreliability. In such a case, between-group effects calculated on the basis of manifest group means can be biased (and vice versa; see Lüdtke et al., 2008).

Consequences for multilevel MI. These aspect of multilevel data have important consequences for the treatment of missing data and multilevel MI. First, missing data may occur at both Level 1 and 2. Second, if the substantive analysis model allows for different relations

between variables between and within groups (e.g., contextual effects), it is important that the imputation model acknowledges that by including the between-group components of variables at Level 1 during the imputation of missing data at Level 1 and 2 (see also Enders et al., 2016). Finally, either manifest or latent group means can be used to represent the between-group components of variables at Level 1, and it is still not yet fully understood (a) how the between-group components are handled in current implementations of multilevel JM and FCS and (b) which option is to be preferred in a given scenario. In the following, I consider these issues in the context of the multilevel random intercept model with missing data at Level 1 and 2.

2.2 Random intercept models

To guide the following discussion, consider the following multilevel random intercept model (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012b). For individual i ($i = 1, \ldots, n_j$) in group j ($j = 1, \ldots, J$),

$$y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - \bar{x}_{\bullet j}) + \gamma_{01}\bar{x}_{\bullet j} + \gamma_{02}w_j + u_{0j} + e_{ij}\,, \tag{2.3}$$

where y_{ij} denotes the values of an outcome variable at Level 1, x_{ij} those of an explanatory variable at Level 1, $\bar{x}_{\bullet j}$ denotes the (manifest) group means of the x_{ij}, and w_j denotes the values of an explanatory variable at Level 2. In addition, u_{0j} denotes the random intercepts at Level 2, which is assumed to follow a normal distribution with mean zero and variance τ_0^2, and e_{ij} denotes residuals at Level 1, which is assumed to follow a normal distribution with mean zero and variance σ^2.

In this model, the outcome variable Y is allowed to vary at both Level 1 and 2. In addition, the model allows for different relations between Y and X at Level 1 and 2 by including different regression coefficients for $(x_{ij} - \bar{x}_{\bullet j})$ and $\bar{x}_{\bullet j}$. Finally, the model includes variables measured directly at Level 2, which do not vary within groups (W). If missing data occur in some or all of these variables, both joint modeling (JM) and the fully conditional specification (FCS) can be used for multilevel JM (Enders et al., 2016; Lüdtke et al., 2017). In the following, I consider these two approaches and describe how they incorporate the different features of the analysis model.

2.2.1 *Joint modeling*

In the JM approach, a single model is used to generate imputations for all variables simultaneously. In the context of multilevel data, the JM approach is based on a multivariate mixed-effects model (Carpenter & Kenward, 2013; Goldstein, Carpenter, Kenward, & Levin, 2009; Schafer & Yucel, 2002; Yucel, 2008). For a set of continuous variables measured at Level 1 and 2, the model can be written as follows. For student i ($i = 1, \ldots, n_j$) in school j ($j = 1, \ldots, J$)

$$\begin{aligned} \mathbf{y}_{1ij} &= \boldsymbol{\mu}_1 + \mathbf{u}_{1j} + \mathbf{e}_{ij} \quad \text{(Level 1)} \\ \mathbf{y}_{2j} &= \boldsymbol{\mu}_2 + \mathbf{u}_{2j}\,, \qquad \text{(Level 2)} \end{aligned} \tag{2.4}$$

where $\mathbf{y}_{1ij}$ denotes values for variables at Level 1 with means $\boldsymbol{\mu}_1$, random intercepts $\mathbf{u}_{1j}$ at Level 2, and residuals $\mathbf{e}_{1ij}$ at Level 1. Likewise, $\mathbf{y}_{2j}$ denotes values for variables at Level 2, with means $\boldsymbol{\mu}_2$ and residuals $\mathbf{u}_{2j}$ at Level 2. The random intercepts and residuals at Level 2 combined, $\mathbf{u}_j = (\mathbf{u}_{1j}, \mathbf{u}_{2j})$, are assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$. The residuals at Level 1, $\mathbf{e}_{1ij}$, are assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. If applied to the example above, $\mathbf{y}_{1ij}$ would comprise X and Y as well as auxiliary variables at Level 1, and $\mathbf{y}_{2j}$ would comprise W as well as auxiliary variables at Level 2. In addition, the model may include an additional set of completely observed predictor variables with associated fixed and random effects, an option that is not discussed here for simplicity (e.g., Schafer & Yucel, 2002).

In this formulation of multilevel JM, there are several points worth noting. First, multilevel JM separates between- and within-group components of variables at Level 1. Specifically, neglecting the overall means $\boldsymbol{\mu}_1$, the values $\mathbf{y}_{1ij}$ are decomposed into a vector of random effects $\mathbf{u}_{1j}$ specific to each group and a vector of residuals $\mathbf{e}_{1ij}$ specific to each individual. In doing so, multilevel JM automatically adopts a *latent* decomposition for the variables at Level 1, where $\mathbf{u}_{1j}$ represents the latent group means at Level 2, and $\mathbf{e}_{1ij}$ the individual deviations at Level 1. Moreover, multilevel JM allows for (a) between-group relations between all variables (i.e., variables both at Level 1 and 2) by allowing for the random intercepts and the residuals at Level 2 to be correlated ($\boldsymbol{\Psi}$) and (b) within-group relations between the variables at Level 1 by allowing the residuals at Level 1 to be correlated ($\boldsymbol{\Sigma}$).

Categorical data. The multilevel JM approach can also treat missing data in categorical variables. For example, for a categorical variable with c categories, the model can be extended to include $c - 1$ latent normal background variables that represent different categories, where c is the number of categories (Carpenter & Kenward, 2013; Goldstein et al., 2009). Ordinal data can be addressed in a similar manner with a single latent background variable and a set of $c - 1$ threshold parameters that represent different categories (Asparouhov & Muthén, 2010b). For categorical variables at Level 1, these strategies are equivalent to generating imputations from a (multivariate) generalized linear mixed-effects model with appropriate link functions (e.g., logistic or probit). Further details on the computational aspects of these models are given by Carpenter and Kenward (2013). In the following, I outline the sampling algorithm for the imputation of continuous data at Level 1 and 2 with multilevel JM (see also Carpenter & Kenward, 2013; Goldstein et al., 2009).

Sampling algorithm. Let $\mathbf{y}_{1ij}$ and $\mathbf{y}_{2j}$ be missing in arbitrary patterns so that they can be partitioned into observed and unobserved parts, $\mathbf{y}_{1ij} = (\mathbf{y}_{1ij}^{obs}, \mathbf{y}_{1ij}^{mis})$ and $\mathbf{y}_{2j} = (\mathbf{y}_{2j}^{obs}, \mathbf{y}_{2j}^{mis})$. Note that the covariance matrices at Level 1 and 2, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, can be partitioned according to the missing and observed parts of $\mathbf{y}_{1ij}$ and $\mathbf{y}_{2j}$, respectively. For each individual, $\boldsymbol{\Sigma}$ can be partitioned as $\begin{bmatrix} \boldsymbol{\Sigma}_{ij}^{obs} & \boldsymbol{\Sigma}_{ij}^{mis,obs} \\ \boldsymbol{\Sigma}_{ij}^{obs,mis} & \boldsymbol{\Sigma}_{ij}^{mis} \end{bmatrix}$ and for each group, $\boldsymbol{\Psi}$ can be partitioned as $\begin{bmatrix} \boldsymbol{\Psi}_{j}^{obs} & \boldsymbol{\Psi}_{j}^{mis,obs} \\ \boldsymbol{\Psi}_{j}^{obs,mis} & \boldsymbol{\Psi}_{j}^{mis} \end{bmatrix}$. In addition, $\boldsymbol{\Psi}$ can also be partitioned according to whether variables were measured at Level 1 and 2, namely $\begin{bmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_2 \end{bmatrix}$. Similarly, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be partitioned as $(\boldsymbol{\mu}_{1ij}^{obs}, \boldsymbol{\mu}_{1ij}^{mis})$ and $(\boldsymbol{\mu}_{2j}^{obs}, \boldsymbol{\mu}_{2j}^{mis})$, respectively.

The procedure seeks to find plausible imputations for $\mathbf{y}_{1ij}^{mis}$ and $\mathbf{y}_{2j}^{mis}$ on the basis of the observed data and the parameters of the imputation model, $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. Given flat priors for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and inverse-Wishart priors $W^{-1}(\nu_1, \boldsymbol{\Delta}_1^{-1})$ and $W^{-1}(\nu_2, \boldsymbol{\Delta}_2^{-1})$ for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, respectively, as well as a set of starting values for the missing data and $\boldsymbol{\theta}$, the procedure can be summarized as follows. For notational convenience, I describe the P- and I-steps in reversed order. Then, at iteration t,

1. I-step (missing data at Level 2): Update $\mathbf{y}_{2j}^{imp}$ as follows.

   i) Draw $\mathbf{u}_{1j}^{(t+1)} \sim N(\tilde{\mathbf{u}}_{1j}^{(t)}, \mathbf{U}_{1j}^{(t)})$ with mean and covariance matrix as follows.

a) $\tilde{\mathbf{u}}_{1j}^{(t)} = (\mathbf{I} - \mathbf{\Lambda}_{1|2j}^{(t)})\boldsymbol{\mu}_{1|2j}^{(t)} + \frac{1}{n_j}\mathbf{\Lambda}_{1|2j}^{(t)} \sum_i (\mathbf{y}_{1ij}^{(t)} - \boldsymbol{\mu}_1^{(t)})$, where $\mathbf{\Lambda}_{1|2j}^{(t)} = \mathbf{\Psi}_{1|2}^{(t)} \left[\mathbf{\Psi}_{1|2}^{(t)} + \frac{1}{n_j}\mathbf{\Sigma}^{(t)}\right]^{-1}$ is the reliability of the group means, $\boldsymbol{\mu}_{1|2j}^{(t)} = \mathbf{\Psi}_{12}^{(t)} \left[\mathbf{\Psi}_2^{(t)}\right]^{-1} \mathbf{u}_{2j}^{(t)}$ is the expected value of $\mathbf{u}_{1j}$, and $\mathbf{\Psi}_{1|2}^{(t)} = \mathbf{\Psi}_1^{(t)} - \mathbf{\Psi}_{12}^{(t)} \left[\mathbf{\Psi}_2^{(t)}\right]^{-1} \mathbf{\Psi}_{21}^{(t)}$ is the variance of $\mathbf{u}_{1j}$, given $\mathbf{u}_{2j}$

b) $\mathbf{U}_{1j}^{(t)} = \frac{1}{n_j}\mathbf{\Lambda}_{1|2j}^{(t)}\mathbf{\Sigma}^{(t)}$ with $\mathbf{\Lambda}_{1|2j}^{(t)}$ as above

ii) Calculate $\mathbf{u}_{2j}^{obs,(t+1)} = \mathbf{y}_{2j}^{obs} - \boldsymbol{\mu}_{2j}^{obs,(t)}$ for the observed data at Level 2.

iii) Draw $\mathbf{u}_{2j}^{imp,(t+1)} \sim N(\tilde{\boldsymbol{\mu}}_{2j}^{mis|obs,(t)}, \mathbf{\Psi}_j^{mis|obs,(t)})$ with mean and covariance matrix as follows.

a) $\tilde{\boldsymbol{\mu}}_{2j}^{mis|obs,(t)} = \mathbf{\Psi}_j^{obs,mis,(t)} \left[\mathbf{\Psi}_j^{obs,(t)}\right]^{-1} \mathbf{u}_j^{obs,(t+1)}$, where $\mathbf{u}_j^{obs,(t+1)} = (\mathbf{u}_{1j}^{obs,(t+1)}, \mathbf{u}_{2j}^{obs,(t+1)})$

b) $\mathbf{\Psi}_j^{mis|obs,(t)} = \mathbf{\Psi}_j^{mis,(t)} - \mathbf{\Psi}_j^{obs,mis,(t)} \left[\mathbf{\Psi}_j^{obs,(t)}\right]^{-1} \mathbf{\Psi}_j^{mis,obs,(t)}$

iv) Form $\mathbf{u}_{2j}^{(t+1)} = (\mathbf{u}_{2j}^{obs,(t+1)}, \mathbf{u}_{2j}^{imp,(t+1)})$ and impute $\mathbf{y}_{2j}^{(t+1)} = \boldsymbol{\mu}_2^{(t)} + \mathbf{u}_{2j}^{(t+1)}$.

2. I-step (missing data at Level 1): Update $\mathbf{y}_{1ij}^{imp}$ as follows.

i) Calculate $\mathbf{e}_{1ij}^{obs,(t+1)} = \mathbf{y}_{1ij}^{obs} - \mathbf{u}_{1j}^{obs,(t+1)} - \boldsymbol{\mu}_{1ij}^{obs,(t)}$ for the observed data at Level 1.

ii) Draw $\mathbf{e}_{1ij}^{imp,(t+1)} \sim N(\tilde{\boldsymbol{\mu}}_{1ij}^{mis|obs,(t)}, \mathbf{\Sigma}_{ij}^{mis|obs,(t)})$ with mean and covariance matrix as follows

a) $\tilde{\boldsymbol{\mu}}_{1ij}^{mis|obs,(t)} = \mathbf{\Sigma}_{ij}^{obs,mis,(t)} \left[\mathbf{\Sigma}_{ij}^{obs,(t)}\right]^{-1} \mathbf{e}_{1ij}^{obs,(t+1)}$

b) $\mathbf{\Sigma}_{ij}^{mis|obs,(t)} = \mathbf{\Sigma}_{ij}^{mis,(t)} - \mathbf{\Sigma}_{ij}^{obs,mis,(t)} \left[\mathbf{\Sigma}_{ij}^{obs,(t)}\right]^{-1} \mathbf{\Sigma}_{ij}^{mis,obs,(t)}$

iii) Form $\mathbf{e}_{1ij}^{(t+1)} = (\mathbf{e}_{1ij}^{obs,(t+1)}, \mathbf{e}_{1ij}^{imp,(t+1)})$ and impute $\mathbf{y}_{1ij}^{(t+1)} = \boldsymbol{\mu}_1^{(t)} + \mathbf{u}_{1j}^{(t+1)} + \mathbf{e}_{1ij}^{(t+1)}$.

3. P-step: Update $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Psi}, \mathbf{\Sigma})$ as follows.

i) Draw $\boldsymbol{\mu}_1^{(t+1)} \sim N(\tilde{\mathbf{y}}_1^{(t+1)}, \frac{1}{N}\mathbf{\Sigma}^{(t)})$, where $\tilde{\mathbf{y}}_1^{(t+1)} = \frac{1}{N}\sum_{i,j}(\mathbf{y}_{1ij}^{(t+1)} - \mathbf{u}_{1j}^{(t+1)})$ and N is the total sample size.

ii) Draw $\boldsymbol{\mu}_2^{(t+1)} \sim N(\bar{\mathbf{y}}_2^{(t+1)}, \frac{1}{J}\mathbf{\Psi}^{(t)})$, where $\bar{\mathbf{y}}_2^{(t+1)} = \frac{1}{J}\sum_j \mathbf{y}_{2j}^{(t+1)}$.

iii) Draw $\mathbf{\Sigma}^{(t+1)} \sim W^{-1}(\nu_1 + N, \mathbf{\Delta}_1^{-1} + \mathbf{S}_1^{(t+1)})$, where $\mathbf{S}_1^{(t+1)} = \sum_{i,j} \mathbf{e}_{1ij}^{(t+1)T}\mathbf{e}_{1ij}^{(t+1)}$.

iv) Draw $\mathbf{\Psi}^{(t+1)} \sim W^{-1}(\nu_2 + J, \mathbf{\Delta}_2^{-1} + \mathbf{S}_2^{(t+1)})$, where $\mathbf{S}_2^{(t+1)} = \sum_j \mathbf{u}_j^{(t+1)T}\mathbf{u}_j^{(t+1)}$ and $\mathbf{u}_j^{(t+1)} = (\mathbf{u}_{1j}^{(t+1)}, \mathbf{u}_{2j}^{(t+1)})$.

The procedure acknowledges both the between- and within-group relations between variables: Missing data at Level 1 are imputed conditionally on the observed data at Level 1; missing data at Level 2 are imputed conditionally on the observed data at Level 2 as well as the random effects of the variables at Level 1. This also illustrates how exactly the latent group means (i.e., random effects) of variables at Level 1 are used in multilevel JM.

2.2.2 *Fully conditional specification*

As an alternative to multilevel JM, the joint distribution of the variables can be approximated with a sequence of conditional models by multilevel FCS. For example, in the context of the multilevel random intercept model with continuous data, missing data at Level 1 can be addressed with univariate mixed-effects models, and missing data at Level 2 with regression models (e.g., van Buuren, 2011; Yucel, Schenker, & Raghunathan, 2007; see also Gelman & Hill, 2006). Specifically, with missing data at Level 1 and 2, multilevel FCS can be based on the following set of models. For the p-th variable with missing data at Level 1,

$$y_{1ijp} = \mathbf{y}_{ij(-p)}\boldsymbol{\beta}_{1p} + u_{1jp} + e_{ijp}\,, \tag{2.5}$$

where $\mathbf{y}_{ij(-p)}$ denotes all variables at Level 1 and 2 other than y_{1ijp} (or a subset of these) as well as the between-group components of the variables at Level 1, and $\boldsymbol{\beta}_{1p}$ is a vector of regression coefficients. For he q-th variable with missing data at Level 2,

$$y_{2jq} = \mathbf{y}_{j(-q)}\boldsymbol{\beta}_{2q} + u_{2jq}\,, \tag{2.6}$$

where $\mathbf{y}_{j(-q)}$ denotes all variables at Level 2 other than y_{2jq} as well as the between-group components of the variables at Level 1 (or a subset of these), and $\boldsymbol{\beta}_{2q}$ is a vector of regression coefficients. The random intercepts u_{1jp} as well as the residuals u_{2jq} and e_{ijp} are each assumed to follow independent normal distributions with mean zero and variances ψ^2_{1p}, ψ^2_{2q}, and σ^2_p, respectively.

To address multivariate patterns of missing data, the FCS approach iterates across all variables with missing data. Once imputations have been generated for variables at Level 1, their between-group components must be updated in order to reflect the most recent imputations (e.g., with "passive imputation"; Royston, 2004; van Buuren & Groothuis-Oudshoorn, 2011). By including all other variables at Level 1 and 2 (or a subset) as well as the between-group components of variables at Level 1 as predictor variables in each variable's imputation model, the multilevel FCS approach—like multilevel JM—allows for all between- and within-group relations between variables to be included during MI. However, in contrast to multilevel JM, the current implementations of multilevel FCS use *manifest* group means as the between-group

components of variables at Level 1 (e.g., van Buuren, 2011; see also Enders, Keller, & Levy, in press; Enders et al., 2016).

Categorical data. Because multilevel FCS employs a sequence of univariate models to handle missing data adapting categorical variables with missing data is relatively straightforward. For example, categorical and ordinal variables with missing data at Level 1 can be imputed on the basis of multinomial and ordered logistic (or probit) mixed-effects models, respectively (Asparouhov & Muthén, 2010b; Carpenter & Kenward, 2013; Enders et al., in press; see also W. Wu, Jia, & Enders, 2015). For categorical and ordinal variables with missing data at Level 2, multinomial and ordered logistic (or probit) regression models can be used as in the single-level case (see Chapter 1). In the following, I outline the sampling algorithm for the imputation of continuous data at Level 1 with multilevel FCS. The algorithm for missing data at Level 2 is identical with that in single-level data (Chapter 1) and is formally described in Article 3.

Sampling algorithm. For the variables with missing data at Level 1 (y_{1ijp}), each y_{1ijp} can be partitioned as $y_{1ijp} = (y_{1ijp}^{obs}, y_{1ijp}^{mis})$. Furthermore, let $N_{p,obs}$ and $J_{p,obs}$ denote the number of individuals and groups, respectively, for which y_{1ijp} is observed. Under "flat" priors for $\boldsymbol{\beta}_{1p}$, with priors $\psi_{1p}^2 \sim \text{inv-}\chi^2(\nu_{1p}, \tau_{1p}^2)$ and $\sigma_p^2 \sim \text{inv-}\chi^2(\nu_p, \tau_p^2)$, and given a set of starting values, imputations are generated as follows. At iteration t, for variable p,

1. Estimate $\hat{\boldsymbol{\beta}}_{1p}^{(t)}$, $\hat{\psi}_{1p}^{2,(t)}$, and $\hat{\sigma}_p^{2,(t)}$ from the multilevel random intercept model $y_{1ijp}^{obs} = \mathbf{y}_{ij(-p)}^{(t)}\boldsymbol{\beta}_{1p} + u_{1jp} + e_{ijp}$, where $\mathbf{y}_{ij(-p)}^{(t)} = (\mathbf{y}_{ij(-p)}^{obs}, \mathbf{y}_{ij(-p)}^{imp,(t)})$.
2. P-step: Update $\boldsymbol{\theta} = (\boldsymbol{\beta}_{1p}, \psi_{1p}^2, \sigma_p^2)$ as follows.
   i) Draw $\psi_{1p}^{2,(t+1)} \sim \text{inv-}\chi^2(\nu_{1p} + J_{p,obs}, \frac{\nu_{1p}\tau_{1p}^2 + J_{p,obs}\hat{\psi}_{1p}^{2,(t)}}{\nu_{1p} + J_{p,obs}})$.
   ii) Draw $\sigma_p^{2,(t+1)} \sim \text{inv-}\chi^2(\nu_p + N_{p,obs}, \frac{\nu_p\tau_p^2 + N_{p,obs}\hat{\sigma}_p^{2,(t)}}{\nu_p + N_{p,obs}})$.
   iii) Draw $\boldsymbol{\beta}_{1p}^{(t+1)} \sim N(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\mathbf{V}}^{(t)})$, where $\hat{\mathbf{V}}^{(t)}$ is the estimated variance-covariance matrix of the regression coefficients.
3. I-step: Update y_{1ijp}^{mis} as follows.
   i) Draw $u_{1jp}^{(t+1)} \sim N(\tilde{u}_{1jp}^{(t)}, U_{1jp}^{(t)})$ with mean and covariance matrix as follows.

a) $\tilde{u}_{1jp}^{(t)} = \lambda_{pj}^{(t+1)} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y_{1ijp}^{obs} - \mathbf{y}_{ij(-p)}^{(t)} \boldsymbol{\beta}_{1p}^{(t+1)} \right)$, where $\lambda_{pj}^{(t+1)} = \frac{\psi_{1p}^{2,(t+1)}}{\psi_{1p}^{2,(t+1)} + \sigma_p^{2,(t+1)}/n_j}$ is the conditional reliability of the group means of y_{1ijp}^{obs}, given $\mathbf{y}_{ij(-p)}^{(t)}$

b) $U_{1jp}^{(t)} = \lambda_{pj}^{(t+1)} \cdot \frac{\sigma^{2,(t+1)}}{n_j}$ with $\lambda_{pj}^{(t+1)}$ as above

ii) Draw $e_{1ijp}^{imp,(t+1)} \sim N(0, \sigma_p^{2,(t+1)})$ for cases with missing y_{1ijp}.

4. Impute $y_{1ijp}^{imp,(t+1)} = \mathbf{y}_{ij(-p)}^{(t)} \boldsymbol{\beta}_{1p}^{(t+1)} + u_{1jp}^{(t+1)} + e_{ijp}^{(t+1)}$.

It is important to note that, in current implementations of multilevel FCS, the predictors $\mathbf{y}_{ij(-p)}$ may include (a) all variables at Level 1 and 2 other than y_{1ijp} and (b) the between-group components—specifically, the manifest group means—of the variables at Level 1. Once y_{1ijp} has been imputed, the group means of that variable have to be recalculated in a passive imputation step (e.g., Royston, 2005) in order to be used in subsequent steps of the algorithm for the imputation of missing data at Level 1 and 2.

2.2.3 *Maximum-likelihood estimation*

In addition to multilevel MI, it is also possible to treat missing data in multilevel analyses with FIML. Consider the multilevel random intercept model in Equation 2.3. Using FIML, the model parameters are estimated by evaluating the likelihood function directly on the basis of the incomplete data. However, because the likelihood function encompasses only the dependent variable Y, only missing data in Y are addressed by FIML (see also Allison, 2012; Hox, van Buuren, & Jolani, 2016). In order to extend the treatment of missing data to explanatory variables, the analysis model must be altered in such a way that the likelihood function includes the explanatory variables with missing data. For this purpose, software for structural equation modeling (SEM) can be used, which allows introducing additional distributional assumption (e.g., multivariate normal) for the variables with missing data while estimating the parameters of interest in the structural part of the model (see also Enders, 2010).

However, in multilevel data, this strategy can have unintended effects. For example, when estimating the model in Equation 2.3 in the statistical software M*plus* (L. K. Muthén & Muthén, 2012), missing data in X may be accommodated by treating X as a multilevel continuous

variable that is correlated with W at Level 2 and with Y at both Level 1 and 2. However, in this case, the model adopts a latent decomposition of the between- and within-group components of X, thus changing the substantive analysis model and the interpretation of between-group effects. The model given in Equation 2.3 can only be estimated directly by calculating the group means of X beforehand while introducing distributional assumptions only for the within-group components of X. However, although this strategy leaves the specification of between-group components unchanged, the estimates of the group means may be biased if the values in X are missing in a systematic manner (e.g., MAR).

2.2.4 *Comparison of different procedures*

Despite the broad selection of procedures for treating missing data in multilevel research, little is known about how they compare with each other. Consequently, the present dissertation devoted two articles to providing a comparison of these methods and guidance for how they might be used in practice (see Articles 1 and 2). In this context, Article 1 provided an introduction to the treatment of missing data in the context of the multilevel random intercept model using both multilevel JM and FCS as well as FIML. It was shown that when the analysis model used a latent specification of between-group effects (Lüdtke et al., 2008), all procedures provided accurate results. However, when the analysis model used a manifest specification of between-group effects (i.e., manifest means), only multilevel JM and FCS provided satisfactory results. By contrast, the standard strategy for implementing FIML led to strongly biased parameter estimates due to the unintentional change in the analysis model; this bias was reduced (but not fully eliminated) by calculating the manifest group means beforehand.

In Article 2, these comparisons were extended in two different ways: First, missing data in categorical variables and in variables at Level 2 were considered in detail. Second, the comparison was extended to include multilevel models with random slopes and CLIs. For the random intercept model, the results mimicked those in Article 1. In addition, both multilevel JM and FCS were shown to provide accurate results in various conditions with missing data in categorical variables and in variables at Level 2. However, the results presented in Article 2 raised two additional points, which I will consider in the following. First, the question remained

whether there is a formal equivalence in the treatment of missing data with multilevel JM and FCS, that is, whether their use of the between-group components of variables at Level 1 constitutes an equivalent treatment of missing data at Level 1 and 2. This is discussed in the following in the context of missing data at Level 2, in Section 2.3. In addition, Article 2 pointed out several limitations of current procedures for multilevel MI in the context of multilevel models with random slopes and CLIs. This is discussed in Section 2.4.

2.3 Missing data at Level 2

Relatively few studies have considered the treatment of missing data at Level 2 (for recent discussions, see Enders et al., 2016; van Buuren, 2011; see also Black, Harel, & Matthews, 2013; Gelman & Hill, 2006). Moreover, these studies often considered ad-hoc procedures, for example, single-level MI or to restrict the MI procedure to include variables only at Level 2 (Cheung, 2007; Gibson & Olejnik, 2003). For this reason, Article 3 was concerned with the treatment of missing data at Level 2 and the role of between-group components in variables at Level 1. In the following, I argue that the use of manifest group means in multilevel FCS, while usually safe in practice, is not strictly equivalent with multilevel JM. Furthermore, I outline a computational procedure developed in Article 3, which allows including latent group means in multilevel FCS by using the method of *plausible values* (Mislevy, 1991).

2.3.1 *(Non-) Equivalence of manifest and latent group means*

In order to treat missing data at Level 2, the between-group components of variables at Level 1 often need to be taken into account, for example, because they are (a) featured in the analysis model, thus leading to bias in parameter estimates if they were omitted (Meng, 1994; Schafer, 2003), or (b) related to the variables with missing data or the propensity of missing data, thus improving the performance of MI (Collins et al., 2001; Schafer & Graham, 2002). However, it is currently an open question how best to do this. In the multilevel literature, it is well known that the latent and manifest models tend to provide different estimates of group-level effects (e.g., Lüdtke et al., 2008). For that reason, it may be hypothesized that the choice between

the two specifications of between-group components may also affect parameter estimates when they are used in MI.

Regarding the case with balanced data, it has been argued that FCS with manifest group means provides imputations consistent with the joint model (i.e., with multilevel JM) because the two models imply identical variance and covariance structures (Carpenter & Kenward, 2013; see also Mistler, 2015; Mistler & Enders, in press). However, regarding the case with *un*balanced data and missing values at Level 1, Resche-Rigon and White (in press) argued that the conditional distribution implied by the joint model depend not only on the manifest group means but also on group size and recommended a two-step variant of multilevel FCS that allows for heteroscedasticity in Level-1 variances across groups (Audigier & Resche-Rigon, 2017). In the present dissertation, I extend this line of reasoning and show that the use of manifest group means in multilevel FCS can lead to biased estimates of covariances and regression coefficients at Level 2 in unbalanced data.

Consider the case with two variables, where Y is measured at Level 1, and Z is measured at Level 2 (for a more general expression, see Article 3). Under the assumption that Z is MCAR and the number of groups goes to infinity ($J \rightarrow \infty$), it can be shown that (a) the variance of Z is preserved under FCS with manifest group means of Y, but (b) the covariance of Y with Z is biased in the case with unbalanced data. Specifically, the bias (in %) can be shown to be

$$\%Bias(\hat{\sigma}_{yz}) = p \left[\sum_{k \in \mathcal{S}} \left(\frac{k}{\bar{n}} - 1 \right) \pi_k \left(\tau_y^2 + \frac{\sigma_y^2}{k} \right) \right] \left[\sum_{k \in \mathcal{S}} \pi_k \left(\tau_y^2 + \frac{\sigma_y^2}{k} \right) \right]^{-1},$$

where $\hat{\sigma}_{yz}$ is the estimator for the covariance of Y with Z at Level 2 (B. O. Muthén, 1994), p is the probability of missing data, S is the set of unique group sizes in the sample, k is one of the fixed group sizes in $\mathcal{S}$, π_k is the frequency of each k in S, $\bar{n}$ is the average group size, and σ_y^2 and τ_y^2 are the variances of Y at Level 1 and 2. The bias is zero in balanced samples (i.e., when $\frac{k}{\bar{n}} = 1$ for all $k \in \mathcal{S}$) but tends to be negative in unbalanced samples (i.e., toward zero). Though the bias is usually very small in most practical scenarios, this illustrates that multilevel FCS with manifest group means is not fully equivalent to multilevel JM and may introduce bias into estimates of regression coefficients at Level 2 if the joint model holds (for further details, see Article 3).

2.3.2 *Latent group means in multilevel FCS*

As an alternative to the standard implementations of multilevel FCS, latent means may be included in in the imputation model for missing data at Level 2. Recall that, in multilevel FCS, missing data at Level 2 can be treated by regressing each variable with missing data at Level 2 on a set of predictor variables that may include any other variable at Level 2 and the between-group components of variables at Level 1 (Equation 2.6). In order to include in latent group means in the set or predictors, the method of *plausible values* can be used (Mislevy, 1991), which provides a general framework for generating imputations for latent quantities (see also Little & Rubin, 2002; for a similar application, see Yang & Seltzer, 2016). To this end, the latent means are drawn from their posterior distribution, given the variables at Level 2 and between-group components of all the other variables at Level 1. The sampling algorithm is given below (see also Article 3).

Sampling algorithm. Recall that latent group means can be regarded as random effects in a multilevel random intercept model. Therefore, the latent means can be sampled using standard "empirical Bayes" methods for sampling random effects (e.g., Efron & Morris, 1973; see also Raudenbush & Bryk, 2002). At iteration t, the latent group means $y_{B,1jp}$ for the p-th variable at Level 1 are sampled as follows,

1. Estimate $\hat{\boldsymbol{\beta}}_{1p}^{(t)}$, $\hat{\psi}_{1p}^{2,(t)}$, and $\hat{\sigma}_{p}^{2,(t)}$ from the multilevel random intercept model $y_{1ijp}^{(t)} = \mathbf{y}_{j(-p)}^{(t)}\boldsymbol{\beta}_{1p} + u_{1jp} + e_{ijp}$, where $\mathbf{y}_{j(-p)}^{(t)} = (\mathbf{y}_{j(-p)}^{obs}, \mathbf{y}_{j(-p)}^{mis,(t)})$ contains the variables at Level 2, the between-group components of other variables at Level 1, and a constant (for the intercept).
2. Draw $b_{1jp}^{(t)} \sim N(\tilde{b}_{1jp}^{(t)}, B_{1jp}^{(t)})$ with mean and variance are calculated as follows.
   i) $\tilde{b}_{1jp}^{(t)} = (1 - \lambda_{pj}^{(t)}) \cdot \mu_{pj}^{(t)} + \lambda_{pj}^{(t)} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} y_{1ijp}^{(t)}$, where $\lambda_{pj}^{(t)} = \frac{\psi_{1p}^{2,(t)}}{\psi_{1p}^{2,(t)} + \sigma_{p}^{2,(t)}/n_j}$ is the conditional reliability of the group means, and $\mu_{pj}^{(t)} = \mathbf{y}_{j(-p)}^{(t)}\hat{\boldsymbol{\beta}}_{1p}^{(t)}$ is the conditional mean of $y_{1ijp}^{(t)}$, given $\mathbf{y}_{j(-p)}^{(t)}$
   ii) $B_{1jp}^{(t)} = \lambda_{pj}^{(t)} \cdot \frac{\hat{\sigma}^{2,(t)}}{n_j}$ with $\lambda_{pj}^{(t)}$ as above
3. Impute $y_{B,1jp}^{(t)} = b_{1jp}^{(t)}$.

Overall, the sampling procedure can be compared to the conventional sampling of random effects in multilevel FCS (Yucel et al., 2007). For this reason, multilevel FCS approach with latent group means becomes very similar to multilevel JM. Note that the latent group means must be updated with a new posterior draw at each iteration of the procedure even if the underlying variable is completely observed; this is required to preserve the uncertainty associated with the (unobserved) latent means. Finally, it is worth pointing out that the procedure is not settled on how the parameter estimates are obtained, that is, the procedure may be based on ML estimates, Bayesian estimates, or a fully Bayesian approach, in which the point estimates $\hat{\boldsymbol{\beta}}_{1p}$, $\hat{\psi}^2_{1p}$, and $\hat{\sigma}^2_p$ are replaced by draws from the posterior distributions of these parameters. The simulation results of Article 3 indicated that multilevel FCS with latent cluster means provides results that are asymptotically identical to those of multilevel JM and unbiased with both balanced and unbalanced data. Both the implementation with Bayesian estimates and the fully Bayesian procedure were found to yield adequate results, where the fully Bayesian procedure appeared to be the most accurate overall (for further details, see Article 3).

2.4 Random coefficient models

If the model of interest includes additional random effects or non-linear terms such as cross-level interaction effects (CLIs), the application of multilevel MI is less well understood. In the following, I consider multilevel MI in the context of the random coefficients model (e.g., Snijders & Bosker, 2012b), which allows for relations between variables at Level 1 to vary across groups and may include explanatory variables at Level 2 to account for some of that variation. For example, assume a researcher is interested in the following model

$$\begin{aligned} y_{ij} = {} & \gamma_{00} + \gamma_{10}(x_{ij} - \bar{x}_{\bullet j}) + \gamma_{01}\bar{x}_{\bullet j} + \gamma_{02}w_j + \gamma_{11}(x_{ij} - \bar{x}_{\bullet j})w_j \\ & + u_{0j} + u_{1j}(x_{ij} - \bar{x}_{\bullet j}) + e_{ij} \,, \end{aligned} \tag{2.7}$$

where γ_{10} denotes the main (fixed) effect of $(x_{ij} - \bar{x}_{\bullet j})$, u_{1j} denotes the random effect of $(x_{ij} - \bar{x}_{\bullet j})$ that varies across groups, γ_{11} denotes the cross-level interaction (CLI) associated with the product term $(x_{ij} - \bar{x}_{\bullet j})w_j$. The two random effects, u_{0j} and u_{1j}, are assumed to follow a multivariate normal distribution and can be interpreted as random (i.e., unexplained) variation

in the regression coefficients. In that context, the CLI denotes the extent to which the effect of $(x_{ij} - \bar{x}_{\bullet j})$ varies systematically with W. If missing data occur only in the dependent variable Y, both multilevel MI and FIML can be used to estimate the model of interest. Unfortunately, the treatment of missing data is much less straightforward if missing data in explanatory variables (e.g., X). In order to preserve the relevant features of the analysis model, the imputation model must allow for the effect of $(x_{ij} - \bar{x}_{\bullet j})$ on Y to vary both unsystematically and as a function of W. Below, I provide a short discussion about the problems one faces with standard implementations of multilevel JM and FCS. In this context, I summarize the results of Article 2, which evaluated these procedures in the context of the multilevel random coefficient model with and without CLIs, and I present an alternative procedure for estimating the model of interest that relies on Bayesian estimation.

2.4.1 *Challenges with multilevel MI*

In multilevel research, the random coefficients model is frequently used to gain a better understanding how the relations between individual-level variables (Level 1) vary across groups (at Level 2) and which explanatory variables account for some of that variance (CLIs; for a discussion, see Aguinis & Culpepper, 2015). From the viewpoint of multilevel MI, the challenges associated with missing data in the random coefficients model are twofold and concern both random slopes of explanatory variables with missing data and the presence of non-linear terms such as the CLI.

Random slopes. If missing data occur only in the dependent variable Y, the treatment of missing data is straightforward. Specifically, missing data in Y can be imputed conditionally on the explanatory variables with a univariate mixed-effects model that mimics the model of interest (potentially with additional auxiliary variables) using either multilevel JM or FCS (Schafer & Yucel, 2002; see also Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016a) or simply estimated using FIML. However, if missing data occur in X, it is currently not possible to directly include the random slope of $(x_{ij} - \bar{x}_{\bullet j})$ in the imputation model in multilevel MI; it is possible to "reverse" the imputation model (e.g., with multilevel FCS; see Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016a) such that the imputation model for X contains a

random slope for $(y_{ij} - \bar{y}_{\bullet j})$, but this does not directly correspond to the relation specified in the model of interest. Consequently, this strategy has been shown to induce bias into the estimates of fixed effects and to underestimate the slope variance in subsequent analyses (Grund, Lüdtke, & Robitzsch, 2016a; see also (Gottfredson et al., 2017)).

In the simulations conducted in Article 2, the performance of multilevel MI and FIML in the context of the random coefficients model was evaluated in a broad range of settings both with and without CLIs. The results for conditions without CLIs are largely in line with previous findings: Both multilevel MI and FIML provide asymptotically unbiased parameter estimates when missing data are confined to the dependent variable. However, if the explanatory variable is affected by missing data, "reversing" the imputation model (e.g., using multilevel FCS) results in slightly biased estimates of the fixed effect and the slope variance. If the random slope is omitted from the imputation model (e.g., using multilevel JM or FCS), the bias in the fixed effects becomes smaller at the cost of larger bias in the slope variance.

Cross-level interactions (CLIs). Similar to the challenges associated with random slopes, the treatment of missing data in models with CLIs is only straightforward if missing data are confined to the outcome variable Y. By contrast, if the explanatory variables that partake in the interaction contain missing data (e.g., X and W), it is currently an open question how best to perform MI. In single-level MI, several ad-hoc procedures have been proposed to accommodate interaction effects. For example, the interaction term may be regarded as "just another variable" (JAV) and imputed without further constraints (von Hippel, 2009; White et al., 2011). This approach has produced mixed results overall but has also been termed "the best of a set of imperfect methods" (Seaman, Bartlett, & White, 2012). However, in multilevel MI, using JAV may not be straightforward if the model relies on group mean centering to separate between- and within-group effects because the group means are themselves subject to uncertainty (see Article 2). As an alternative, it has been recommended to simply impute the variables underlying the interaction effect (e.g., X and W), after which the product term can be updated using "passive imputation" (e.g., White et al., 2011). This approach is particularly attractive in multilevel MI because it is easy to implement and available in standard software; however, it has been shown to be more prone to bias as compared with other methods (e.g., S. Kim, Sugar, & Belin, 2015;

Seaman et al., 2012; von Hippel, 2009).

In Article 2, particular emphasis was placed on multilevel FCS with passive imputation of the CLI. In this evaluation, multilevel MI still provided reasonable estimates of the main effects, but estimates of the CLI and the (residual) slope variance were noticeably biased. If the random slope was omitted, the bias in main effects and the CLI became slightly smaller at the cost of larger bias in the slope variance. By contrast, LD provided biased estimates of the main effects but the least biased estimates of the CLI; single-level MI generally performed worse than did multilevel MI. Taken together, the results of the simulations in Article 2 indicated that, although LD and single-level MI tended to perform worse than multilevel MI, current implementations of multilevel MI are still not perfectly suited for dealing with missing data in the multilevel random coefficients model. For that reason, Article 2 also includes a list of recommendations and fully reproducible examples for the use of current software for multilevel MI.

2.4.2 *New methods for accommodating random slopes and CLIs*

In the context of single-level MI, it has previously been acknowledged that the conditional distributions for the imputation of incomplete explanatory variables employed under passive imputation and JAV are misspecified when the true model includes quadratic or interaction effects (Seaman et al., 2012; see also S. Kim, Belin, & Sugar, in press; S. Kim et al., 2015). As an alternative, Bartlett, Seaman, White, and Carpenter (2015) poposed an adjusted procedure for single-level FCS which factorizes the joint posterior distribution into separate components pertaining to the model of interest and the explanatory variables with missing data, thus taking nonlinear and interaction effects in the model of interest into account during MI. Similar approaches have also been applied in the context of regression (Zhang & Wang, 2016) and multilevel analyses (Erler et al., 2016), which used Bayesian estimation to estimate the model of interest directly from the incomplete data but can also be used to perform MI. In addition, Goldstein, Carpenter, & Browne, 2014 proposed a procedure for fitting the joint model using Bayesian methods and generating imputations with multilevel JM while accommodating random slopes and interaction effect in the model of interest. First implementations of this approach are currently becoming available but still require further evaluation (Quartagno & Carpenter, 2017).

Finally, similar methods have also been proposed for ML estimation, which likewise rely on a factorization of the joint likelihood into separate components pertaining to the model of interest and the explanatory variables with missing data (Ibrahim, Chen, & Lipsitz, 2001; Stubbendick & Ibrahim, 2003). In the context of Article 2, preliminary simulations have demonstrated that these procedures possess great potential for the treatment of missing data in the random coefficients model. For this reason, I consider these approaches in some additional detail with an emphasis on Bayesian estimation.

2.5 Bayesian estimation of the random coefficients model

The presence of interaction effects complicate matters because they imply a complex joint distribution for the variables of interest (e.g., Seaman et al., 2012). Consider the model in Equation 2.7 with missing data in X, in which case imputations would usually be generated from the conditional distribution $P(X|Y, W)$ using multilevel JM or FCS. However, it has been shown that this distribution is not strictly linear in Y and W but also includes the interaction between the two and higher-order effects of W (S. Kim et al., 2015). Then, according to Bayes' theorem, the conditional distribution of X, given Y and W, can be expressed in the alternative factorization

$$P(X|Y, W) \propto P(Y|X, W)P(X|W)P(W)\,, \tag{2.8}$$

where $P(Y|X, W)$, which is the model of interest, and $P(X|W)P(W)$, which represents a conditional model for the (missing) covariates. In other words, instead of sampling directly from $P(X|Y, W)$, samples can be obtained equivalently by sampling from the factorization on the right-hand side of the equation. In practice, this expression will not belong to a standard family of distributions so that sampling can be achieved by rejection sampling or Metropolis-Hastings (MH) steps with a suitable proposal distribution for missing $x_{i,mis}$, after which the proposed value $x^{*}_{i,mis}$ is rejected or accepted based on the joint likelihood of the data (see also Bartlett et al., 2015; Goldstein et al., 2014). This provides the advantage that interactions and nonlinear terms enter the joint density only through the model of interest but not the model for X, which can now take simpler parametric forms.

In the context of Bayesian estimation with missing data, this approach can be used to factorize the joint posterior distribution in a similar manner (Erler et al., 2016). For example, for the model of interest in Equation 2.7 with missing data in X, the joint posterior distribution can be written as

$$P(\theta, \xi, X_{mis}|Y, X_{obs}, W) \propto P(Y|X_{obs}, X_{mis}, W, \theta)P(\theta)P(X_{obs}, X_{mis}|W, \xi)P(\xi)\,, \qquad (2.9)$$

which comprises the model of interest with parameters θ, and a conditional model for the missing values in X, given the observed values for the explanatory variables, with parameters ξ. Similar approaches have been used to obtain ML estimates for multilevel models with missing data (e.g., Ibrahim, Chen, & Lipsitz, 1999; Ibrahim et al., 2001).

Computer code for Bayesian estimation with JAGS. The model can be fitted with standard software for Bayesian estimation such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2016), or Stan (Stan Development Team, 2016). Given below is compute code needed to specify the random coefficients model with missing data in X in the statistical software JAGS.

```
model{

  for(n in 1:N){

    y[n] ~ dnorm( yhat[n] , tau1y )
    x[n] ~ dnorm( xhat[n] , tau1x )

    # model of interest
    yhat[n] <- b0y + b1y * ( x[n] - xgm[group[n]] ) + b2y * xgm[group[n]] + b3y * z[n] +
               b4y * ( x[n] - xgm[group[n]] ) * z[n] + uy[group[n],1] +
               uy[group[n],2] * ( x[n] - xgm[group[n]] )

    # "imputation" model for x
    xhat[n] <- b0x + b1x * z[n] + ux[group[n]]

  }

  for(g in 1:G){

    ux[g] ~ dnorm( 0 , tau2x )
    uy[g,1:2] ~ dmnorm(  M[1:2] , Tau[1:2,1:2] )

    # group mean
    xgm[g] <- mean( x[ ( (g-1)*ng + 1 ):( (g-1)*ng + ng ) ] )

  }

  # fixed effects
```

```
  b0x ~ dnorm( 0 , .001 )
  b1x ~ dnorm( 0 , .001 )
  b0y ~ dnorm( 0 , .001 )
  b1y ~ dnorm( 0 , .001 )
  b2y ~ dnorm( 0 , .001 )
  b3y ~ dnorm( 0 , .001 )
  b4y ~ dnorm( 0 , .001 )

  # variances components
  tau1x <- pow( sigma1x , -2 )
  tau2x <- pow( sigma2x , -2 )
  tau1y <- pow( sigma1y , -2 )
  sigma1x ~ dunif( 0 , 10 )
  sigma2x ~ dunif( 0 , 10 )
  sigma1y ~ dunif( 0 , 10 )

  Tau[1:2,1:2] ~ dwish( Id[1:2,1:2] , scaleId )
  Sigma[1:2,1:2] <- inverse( Tau[1:2,1:2] )

}
```

It is easy to see that this procedure includes both (a) the model of interest and (b) an "imputation" model for X, that is, a random intercept model in which W is used as a predictor at Level 2. Despite its focus on estimation, this procedure can also be used to generate imputations for X, which may be preferable in some settings (see also Goldstein et al., 2014). In the following, I present the results of a simulation study which evaluates the performance of the Bayesian estimation approach.

Evaluation of performance. To evaluate the performance of the Bayesian estimation approach, I conducted a simulation study using the design from Article 2. The data were generated from the random coefficients model in Equation 2.7, where all variables were simulated with mean zero and unit total variance. The simulated conditions included different sample sizes at Level 1 (n = 5, 10) and Level 2 (J = 50, 100, 200, 500), and different levels of the ICC for the variables at Level 1 ($\rho_{I,X} = \rho_{I,Y}$ = .10, .50). In addition, the simulation included different levels of the CLI (γ_{11} = 0, .20) and the total slope variance (i.e., including the contribution of the CLI; $Var(\beta_{1j}) = \gamma_{11}^2 + \tau_1^2$ = .05, .10, .20). Missing data were induced in X according to a MAR mechanism depending on Y, leading to 25% missing data in X. To provide a comparison with other methods, missing data were also treated with the procedures evaluated in Article 2: listwise deletion (LD), single-level FCS (FCS-SL), multilevel FCS passive imputation of the CLI (FCS-CLI/RS), and multilevel JM. In order to allow a fair comparison between the procedures, Bayesian estimation (JAGS) was used to generate imputations for X so that parameter

estimates were obtained in the same way as under MI, that is, by fitting the model of interest using `lme4` (Bates, Maechler, Bolker, & Walker, 2016) and combining the results according to Rubin (1987). The parameter estimates under each method and each condition were then evaluated according to bias, RMSE, and the coverage rate of the 95% confidence interval.

The simulation results are summarized in Table 2.1 for selected conditions. In conditions without CLI ($\gamma_{11} = 0$), the regression coefficients were estimated approximately without bias under JM and JAGS. In addition, FCS-CLI/RS provided imperfect but reasonable estimates of these parameters with bias below 6%. By contrast, FCS-SL and LD provided biased estimates of the regressions coefficients, particularly under FCS-SL with large ICC (.50) and under LD with small ICC (.10). In conditions with CLI ($\gamma_{11} = .20$), the results for the main effects were similar; however, estimates for the CLI ($\hat{\gamma}_{11}$) were biased under FCS-SL, FCS-CLI/RS, JM, and (to a lesser extent) LD; only JAGS provided approximately unbiased estimates of the CLI. Finally, the slope variance ($\hat{\tau}_1^2$) was estimated without bias only under JAGS, whereas the estimates were biased downward under FCS-SL, FCS-CLI/RS, JM, and (to a lesser extent) LD. Regarding the RMSE and the coverage rates of the 95% confidence interval, the results were mostly in line with the bias. However, under FCS-CLI/RS and JM, the coverage rates of the 95% confidence intervals for the within-group regression coefficient of X ($\hat{\gamma}_{10}$) were below the nominal value of 95% despite relatively low bias in these conditions, which may be attributed to the fact that these methods underestimated (FCS-CLI/RS) or ignored (JM) the slope variance, thus underestimating the uncertainty associated with the fixed effect ($\hat{\gamma}_{10}$). Taken together, Bayesian estimation and imputations generated on that basis tended to outperform current implementations of multilevel MI in terms of both accuracy and efficiency. For that reason, Bayesian estimation and related procedures appear to be promising approaches for estimating the multilevel random coefficients model with missing data.

2.6 Summary

In this chapter, I attempted to (a) provide an overview of the treatment of missing data and multilevel MI and (b) motivate the research conducted as part of the present dissertation. In

Table 2.1: Bias (in %), RMSE, and Coverage (in %) of the 95% Confidence Interval for the Within-Group Regression Coefficient of X, the Between-Group Regression Coefficient of W, the CLI of X with W, and the Residual Slope Variance (Medium Group Size, n = 10, Total Slope Variance = .20)

	LD			FCS-SL			FCS-CLI/RS			JM			JAGS		
	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.
Random coefficients model without CLI (γ_{11} = 0)															
k = 100							$\rho_{I,X} = \rho_{I,Y} = .10$								
$\hat{\gamma}_{10}$	−6.2	0.06	91.4	−8.5	0.07	85.5	−5.0	0.06	90.1	−1.4	0.05	90.7	−0.2	0.05	94.1
$\hat{\gamma}_{02}$	−10.2	0.05	80.0	−1.7	0.03	93.5	−0.0	0.03	94.7	−0.3	0.03	94.5	−0.0	0.03	94.7
$\hat{\gamma}_{11}$	−0.3[a]	0.06	92.8	−3.2[a]	0.04	96.4	−2.2[a]	0.05	94.3	−0.3[a]	0.05	96.6	−0.0[a]	0.06	94.1
$\hat{\tau}^2_1$	−10.8	0.05	—	−37.7	0.08	—	−22.2	0.06	—	−33.6	0.07	—	1.2	0.04	—
k = 500															
$\hat{\gamma}_{10}$	−6.0	0.04	75.1	−8.2	0.05	51.7	−4.1	0.03	85.9	−1.1	0.02	91.6	0.2	0.02	95.1
$\hat{\gamma}_{02}$	−10.4	0.04	34.0	−1.9	0.02	90.5	−0.2	0.01	93.7	−0.5	0.01	94.1	−0.2	0.01	93.9
$\hat{\gamma}_{11}$	−1.3[a]	0.02	94.7	−4.0[a]	0.02	96.0	−2.1[a]	0.02	95.6	−1.0[a]	0.02	96.8	−0.5[a]	0.03	94.1
$\hat{\tau}^2_1$	−11.3	0.03	—	−37.3	0.08	—	−24.4	0.05	—	−33.1	0.07	—	0.1	0.02	—
k = 100							$\rho_{I,X} = \rho_{I,Y} = .50$								
$\hat{\gamma}_{10}$	−4.6	0.06	93.7	−34.6	0.18	3.6	−5.5	0.06	90.9	−2.0	0.05	93.5	−0.8	0.05	95.2
$\hat{\gamma}_{02}$	−5.0	0.07	93.9	−1.8	0.07	94.7	1.1	0.07	95.2	1.0	0.07	95.2	1.0	0.07	95.2
$\hat{\gamma}_{11}$	−0.7[a]	0.05	96.2	−12.9[a]	0.04	95.4	−2.4[a]	0.05	95.8	−0.3[a]	0.04	97.1	−0.0[a]	0.05	94.7
$\hat{\tau}^2_1$	−6.0	0.05	—	−53.3	0.11	—	−21.5	0.06	—	−32.8	0.07	—	1.0	0.04	—
k = 500															
$\hat{\gamma}_{10}$	−4.1	0.03	85.3	−34.5	0.17	0.0	−4.2	0.03	82.7	−1.4	0.03	89.9	−0.3	0.02	95.4
$\hat{\gamma}_{02}$	−5.7	0.03	88.8	−3.0	0.03	93.5	−0.0	0.03	95.8	−0.0	0.03	95.8	−0.0	0.03	95.8
$\hat{\gamma}_{11}$	−0.4[a]	0.02	95.8	−12.6[a]	0.03	79.2	−1.6[a]	0.02	96.6	−0.5[a]	0.02	97.1	−0.1[a]	0.02	95.8
$\hat{\tau}^2_1$	−7.5	0.02	—	−54.6	0.11	—	−24.2	0.05	—	−33.3	0.07	—	−0.4	0.02	—
Random coefficients model with CLI (γ_{11} = .20)															
k = 100							$\rho_{I,X} = \rho_{I,Y} = .10$								
$\hat{\gamma}_{10}$	−5.8	0.06	89.1	−10.0	0.07	80.6	−6.1	0.06	90.5	−3.0	0.05	92.0	0.3	0.05	94.5
$\hat{\gamma}_{02}$	−18.9	0.08	50.3	−1.6	0.03	93.1	−0.3	0.03	92.6	−0.5	0.03	93.5	−0.3	0.03	93.7
$\hat{\gamma}_{11}$	−6.8	0.05	93.5	−27.3	0.07	80.0	−15.8	0.05	91.6	−21.2	0.06	87.8	−1.0	0.05	96.2
$\hat{\tau}^2_1$	−10.2	0.04	—	−36.7	0.06	—	−22.4	0.05	—	−32.9	0.06	—	2.1	0.04	—
k = 500															
$\hat{\gamma}_{10}$	−6.6	0.04	67.5	−10.7	0.06	27.6	−6.3	0.04	66.7	−3.8	0.03	82.3	−0.4	0.02	93.5
$\hat{\gamma}_{02}$	−18.7	0.07	0.8	−1.2	0.01	94.3	0.1	0.01	96.2	−0.1	0.01	94.9	0.1	0.01	95.6
$\hat{\gamma}_{11}$	−6.6	0.02	93.2	−26.8	0.06	23.0	−15.1	0.04	72.8	−20.7	0.04	47.7	−0.5	0.02	97.7
$\hat{\tau}^2_1$	−10.3	0.02	—	−36.5	0.06	—	−24.3	0.04	—	−32.5	0.05	—	1.0	0.02	—
k = 100							$\rho_{I,X} = \rho_{I,Y} = .50$								
$\hat{\gamma}_{10}$	−4.4	0.06	93.1	−37.3	0.19	1.5	−7.1	0.06	88.0	−4.2	0.06	91.4	−0.9	0.05	94.5
$\hat{\gamma}_{02}$	−11.2	0.08	88.0	−2.5	0.08	92.8	0.3	0.07	94.7	0.3	0.07	94.9	0.3	0.07	94.9
$\hat{\gamma}_{11}$	−2.1	0.05	95.6	−52.5	0.11	28.4	−14.3	0.05	93.3	−19.0	0.06	89.7	1.3	0.05	94.1
$\hat{\tau}^2_1$	−8.8	0.04	—	−51.4	0.08	—	−23.0	0.05	—	−33.6	0.06	—	−0.6	0.04	—
k = 500															
$\hat{\gamma}_{10}$	−3.7	0.03	86.9	−37.1	0.19	0.0	−5.5	0.04	74.1	−3.2	0.03	87.1	0.0	0.02	95.8
$\hat{\gamma}_{02}$	−11.9	0.05	71.3	−2.7	0.03	94.9	−0.1	0.03	94.9	−0.1	0.03	95.1	−0.1	0.03	94.9
$\hat{\gamma}_{11}$	−3.5	0.02	93.9	−53.2	0.11	0.0	−14.3	0.04	72.8	−19.6	0.04	53.0	0.2	0.02	94.1
$\hat{\tau}^2_1$	−6.3	0.02	—	−50.8	0.08	—	−23.5	0.04	—	−31.5	0.05	—	1.0	0.02	—

Note. $\hat{\gamma}_{10}$ = within-group regression coefficient of X; $\hat{\gamma}_{02}$ = between-group regression coefficient of W; $\hat{\gamma}_{11}$ = CLI; $\hat{\tau}^2_1$ = residual slope variance; LD = listwise deletion; FCS-SL = single-level FCS; FCS-CLI/RS = multilevel FCS including product terms and random slopes; JM = joint modeling; JAGS = model-based MI (via Bayesian estimation).
[a] If the true CLI was zero, the scale of the bias was adjusted to mimic conditions with CLI = .20.

general, multilevel MI is a powerful approach for dealing with missing data in many application of multilevel analysis models. In the context of the multilevel random intercept model, it is important that the imputation model allows for different relations between variables at Level 1 and 2, which can be accomplished in a very general manner with standard implementations of both multilevel JM and FCS (for a detailed discussion, see Articles 1 and 2). However, in applications of multilevel JM and FCS, it is important to acknowledged that they tend to use the between-group components of variables at Level 1 in different ways (i.e., latent vs. manifest group means). The two approaches tend to provide equivalent answers only in balanced but not in unbalanced data; in order for multilevel FCS to be fully consistent with JM, it is possible to simulate latent group means with the method of plausible values as outlined in Article 3 (for a related discussion, see also Article 1).

In the context of the multilevel random coefficients model, it can be challenging to use multilevel MI. Specifically, current implementations of multilevel MI are facing problems when the model of interest includes random slopes or interaction effects (e.g., CLIs) and missing data occur in explanatory variables (for a discussion, see Article 2). As an alternative, it is possible to explicitly take the model of interest into account during model estimation or multilevel MI, for example, using extensions of standard multilevel JM and FCS or by relying on specialized Bayesian or ML estimation procedures. These procedures, though not yet widely available in standard software for multilevel MI, provide promising results in the simulations studies included in the present dissertation and should further be considered in future studies. Further details and a comprehensive set of recommendations regarding the treatment of missing data in multilevel research in various settings are provided in Article 2. In the following chapters, I consider the analysis of multiply imputed data sets and the use of multilevel MI in research practice.

3

Analysis of multiply imputed data

Naturally, the application of MI involves not only the imputation itself but also analyzing the imputed data sets. To this end, the imputed data sets are analyzed separately with regular complete-data methods, and the results are pooled into a final set of parameter estimates and inferences (Rubin, 1987). In the missing data literature, several procedure have been proposed for this task, including procedures for scalar estimands (e.g., individual regression coefficients) as well as for complex statistical hypotheses that involve multiple parameter simultaneously (e.g., model comparisons; for an overview, see Reiter & Raghunathan, 2007). However, relatively little is known about how these procedures perform in practice (e.g., Allison, 2001; Enders, 2010; Schafer, 1997; van Buuren, 2012). In the following chapter, I discuss the procedures available for the pooling of parameter estimates in MI. In that context, I summarize the results of Article 4, which was concerned with the evaluation of different methods for conducting multiparameter tests in the context of the analysis of variance (ANOVA). In addition, I present the results of two simulation studies, which evaluated these methods for testing hypotheses about (a) fixed effects and (b) variance components in multilevel analyses.

3.1 Pooling of scalar estimands

In order to obtain final estimates and inferences for scalar estimands (e.g., regression coefficients), pooling is most frequently achieved with the procedure outlined by Rubin (1987). Let

Q be the quantity of interest, which is estimated in the complete data by $\hat{Q}$ with variance $\hat{U}$. Furthermore, assume that we obtained a number of M ($m = 1, \ldots, M$) imputed data sets. From each of the imputed data sets, we obtain an estimate $\hat{Q}_m$ of the quantity of interest as well as an estimate $\hat{U}_m$ of its variance (e.g., its squared standard error). Then, according to Rubin, the final estimate of Q under MI is

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m \,, \tag{3.1}$$

the average of the individual estimates $\hat{Q}_m$. Under the assumption that $\hat{Q}$ is distributed normally around Q in the compete data, the (total) sampling variance of $\bar{Q}$ can be written as

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B \,, \tag{3.2}$$

where $\bar{U}$ is known as the *within*-imputation variance, which is simply the average of the individual variance estimates $\hat{U}_m$,

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} \hat{U}_m \,, \tag{3.3}$$

and B is known as the *between*-imputation variance and calculated as

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\bar{Q} - \hat{Q}_m)^2 \,. \tag{3.4}$$

In the calculation of the total sampling variance T, the component B/M can be regarded as a penalty term that accounts for the fact that the variance tends to be estimated less precisely if the number of imputations is low.[4]

Statistical hypotheses about Q can be tested similar to complete-data analyses by comparing $\bar{Q}/\sqrt{T}$ against a t distribution with v degrees of freedom,

$$v = (M-1)\left[1 + \frac{1}{\text{RIV}}\right]^2 \,, \tag{3.5}$$

where

$$\text{RIV} = \frac{(1 + 1/M)B}{\bar{U}} \tag{3.6}$$

[4]More formally, Rubin (1987) derives the variance T as an approximation to the posterior variance of Q with $M \to \infty$ (e.g., with hypothetical estimates $\bar{Q}_\infty$, $\bar{U}_\infty$, and B_∞) but based on only a finite number of imputations M. In that context, B/M is an estimator of the variance of $\bar{Q}$ around $\bar{Q}_\infty$ (e.g., Schafer, 1997; see also Carpenter & Kenward, 2013).

is the relative increase in variance due to nonresponse. The idea behind this expression for the degrees of freedom ν is to "widen" the reference distribution in the complete data (i.e., the standard normal) to account for the loss of information that is due to the missing data (Schafer, 1997). On the basis of the RIV, the fraction of missing information (FMI) can be estimated as

$$\widehat{\text{FMI}} = \frac{\text{RIV} + 2/(v+3)}{\text{RIV} + 1}, \tag{3.7}$$

which can provide a useful diagnostic tool in practice because it quantifies the extent to which the available information about the quantity of interest is affected by missing data (see also Andridge & Thompson, 2015; Bodner, 2008).

Small-sample modification. Barnard and Rubin (1999) proposed an alternative expression for the degrees of freedom more suitable for applications in smaller samples. Instead of Rubin's original large-scale approximation, Barnard and Rubin recommend using the adjusted degrees of freedom

$$\tilde{\nu} = \left(\frac{1}{\nu} + \frac{1}{\nu_{obs}} \right)^{-1}, \tag{3.8}$$

where ν_{obs} are the observed-data degrees of freedom, calculated as

$$\nu_{obs} = \frac{\nu_{com} + 1}{\nu_{com} + 3} \nu_{com} \left(\frac{1}{1 + \text{RIV}} \right), \tag{3.9}$$

and ν_{com} are the complete-data degrees of freedom.[5] In practice, using $\tilde{\nu}$ can be useful because it (a) never exceeds ν_{com} and (b) is always smaller than ν. For this reason and because $\tilde{\nu}$ is only slightly conservative in larger samples, Barnard and Rubin recommend using this expression regardless of sample size.

Other modifications. Further modifications and alternatives to Rubin's rules for scalar quantities include variance estimator for complex sampling designs of J. K. Kim, Brick, Fuller, and Kalton (2006), the rules for nested imputations of Shen (2000), and the alternative variance estimator of Robins and Wang (2000), which has been shown to be robust against certain types of misspecification of the imputation model (for further discussion, see Reiter & Raghunathan,

[5] The observed-data degrees of freedom ν_{obs} can be better understood by observing that the estimated FMI can be written as $\widehat{\text{FMI}} = 1 - \frac{\nu+1}{\nu+3} \frac{\bar{U}}{T}$. Therefore, ν_{obs} can be understood as $\nu_{obs} = \nu_{com}(1 - \widehat{\text{FMI}}_{com})$, where $\widehat{\text{FMI}}_{com}$ is an estimate of the FMI based on ν_{com} (Equation 3.7). From this perspective, ν_{obs} reflects a reduced sample size, where the reduction is determined by the FMI.

2007). In the following, I focus on the procedures available for pooling multidimensional estimands (i.e., multiparameter tests).

3.2 Pooling of multidimensional estimands

In research practice, statistical hypotheses often involve more than one parameter. For example, when including a set of explanatory variables in a regression model, it is often interesting to test for the simultaneous contribution of these variables. In the complete data, this test can be carried out, for example, with a Wald-test on the vector of regression coefficients, by testing the difference in variance explained that is attributable to the explanatory variables, or with likelihood-ratio test (LRT). In the following, I discuss the procedures available for multiparameter tests and model comparisons with multiply imputed data sets.

3.2.1 *Moment-based procedure (D_1)*

Similar to the scalar case, let $\mathbf{Q}$ be the K-dimensional quantity of interest (e.g., a vector of regression coefficients), and let $\hat{\mathbf{Q}}_m$ and $\hat{\mathbf{U}}_m$ denote the estimates of the parameter vector and its variance-covariance matrix obtained from M ($m = 1, \ldots, M$) imputed data sets. Then, according to Rubin (1987) and Li, Raghunathan, and Rubin (1991), hypotheses about $\mathbf{Q}$ can be tested with the following test statistic

$$D_1 = \frac{(\bar{\mathbf{Q}} - \mathbf{Q}_0)^T \bar{\mathbf{U}}^{-1} (\bar{\mathbf{Q}} - \mathbf{Q}_0)}{K(1 + \text{ARIV}_1)} , \tag{3.10}$$

where $\bar{\mathbf{Q}}$ is the pooled estimate $\mathbf{Q}$ (i.e., the average $\hat{\mathbf{Q}}_m$), $\mathbf{Q}_0$ is the hypothesized value of $\mathbf{Q}$ under the null hypothesis (e.g., a vector of zeros), $\bar{\mathbf{U}}$ is the within-imputation variance (i.e., the average $\hat{\mathbf{U}}_m$), and ARIV_1 is an estimate of the average relative increase in variance (ARIV),

$$\text{ARIV}_1 = \frac{(1 + M^{-1})\text{tr}(\mathbf{B}\bar{\mathbf{U}}^{-1})}{K} , \tag{3.11}$$

where $\mathbf{B}$ is the between-imputation variance. Li, Raghunathan, and Rubin (1991) recommended comparing D_1 against an F distribution with K numerator and v_1 denominator degrees of freedom. For $a = K(M - 1)$,

$$v_1 = \begin{cases} 4 + (a - 4)[1 + (1 - 2a^{-1})\,\text{ARIV}_1^{-1}]^2 & \text{if } a > 4 \\ (K + 1)(M - 1)(1 + \text{ARIV}_1^{-1})^2/2 & \text{otherwise} \end{cases} . \tag{3.12}$$

Small-sample modification. Reiter (2007) proposed an alternative for the degrees of freedom to be used in smaller samples (henceforth referred to as D_1^*). Conceptually, this modification is similar to that of Barnard and Rubin (1999) and derived using Taylor series expansion. For brevity, the expression is not given in detail here (for further discussion, see Reiter, 2007; Reiter & Raghunathan, 2007). Further adaptions of D_1 which allow for different RIVs across components of the parameter vector are given by Licht (2010).

Current recommendations regarding D_1 and D_1^*. In the missing data literature, D_1 and D_1^* are often recommended because they utilize a (near) maximum of the information provided by the imputed data (e.g., Enders, 2010; Schafer, 1997; van Buuren, 2012). Previous research, though limited in scope, has shown that both D_1 and D_1^* perform well as long as their assumptions are not severely violated (e.g., Li, Raghunathan, & Rubin, 1991; Licht, 2010). Similarly, in Article 4 of the present dissertation, D_1 and D_1^* were always among the most reliable procedures for testing hypotheses in the context of the ANOVA, especially in conditions with larger FMIs, in which other procedures tended to be less robust. In smaller samples, D_1 tended to be slightly liberal, whereas D_1^* provided Type I error rates close to the nominal level across throughout the study (for further details, see Article 4; see also Reiter, 2007; van Ginkel & Kroonenberg, 2014).

3.2.2 *Procedure based on individual χ^2 statistics (D_2)*

In some cases, it may not be feasible to use D_1, for example, because estimates for the variance-covariance matrix of the quantity of interest are not available. Therefore, Li, Meng, Raghunathan, and Rubin (1991) proposed a simple alternative which requires only a χ^2-distributed test statistic W_m (or a p-value, equivalently) from each of the imputed data sets. The pooled test statistic D_2 is then calculated as

$$D_2 = \frac{\overline{W}K^{-1} - (M+1)(M-1)^{-1}\text{ARIV}_2}{1 + \text{ARIV}_2}\,, \tag{3.13}$$

where $\overline{W}$ is the average of the W_m, K is the number of parameters being tested, and ARIV_2 is another estimate of the ARIV,

$$\text{ARIV}_2 = (1 + M^{-1})\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\sqrt{W}_m - \overline{\sqrt{W}}\right)^2\right]\,. \tag{3.14}$$

To conduct hypothesis tests on the basis of D_2, Li, Meng, et al. recommended to compare it against an F distribution with K numerator and v_2 denominator degrees of freedom

$$v_2 = K^{-3/M}(M - 1)(1 + \text{ARIV}_2^{-1})^2 \ . \tag{3.15}$$

Because D_2 requires only the test statistics W_m, it tends to be very flexible and can be used for pooling both Wald-like hypothesis tests and LRTs.

Current recommendations regarding D_2. In the literature, D_2 has been both praised for its simplicity (e.g., Allison, 2001; Snijders & Bosker, 2012b) and criticized because (a) it tends to be less reliable than D_1 when the number of parameters K to be tested is large, and (b) it may provide overly conservative or liberal inferences depending on the FMI, producing Type I error rates well above or below the nominal value (e.g., Enders, 2010; van Buuren, 2012; see also Li, Meng, et al., 1991). However, much of the previous research focused on applications of D_2 with relatively few imputations (e.g., as little as $M = 3$). In Article 4, these results were only partially replicated. Specifically, with $M = 10$ or fewer imputations, D_2 was far more conservative (and less powerful) than other procedures. However, with a larger number of imputations (e.g., $M = 100$), it provided Type I error rates close to the nominal value with good statistical power as long as the FMI was not too large (i.e., larger than 35%). In conditions with larger FMI, D_2 became increasingly liberal with Type I error rates above the nominal value. Based on these findings, D_2 may very well be used in many conditions that are likely to occur in psychological research, particularly if the number of missing values is not too large or auxiliary variables are available that reduce the FMI (for further details, see Article 4).

3.2.3 *Likelihood ratio tests (D_3)*

Finally, hypothesis tests for multiple parameters can be conducted by comparing two nested statistical models using the LRT. For example, when testing for the contribution of a set of explanatory variables, then the LRT may be used to compare the full model with a reduced model that does not include the variables in question. For that reason, Meng and Rubin (1992) proposed a procedure for pooling the LRT, which relies only on the individual LRTs and parameter estimates from each of the M imputed data sets. Let L_m denote the individual LRT

statistics. Then, the pooled LRT can be calculated as

$$D_3 = \frac{\tilde{L}}{K(1 + \text{ARIV}_3)} , \tag{3.16}$$

where $\tilde{L}$ is the average LRT statistic evaluated at the average parameter estimates, K is the number of parameters being tested, and ARIV_3 is another estimate of the ARIV

$$\text{ARIV}_3 = \frac{M + 1}{K(M - 1)}(\bar{L} - \tilde{L}) , \tag{3.17}$$

where $\bar{L}$ is the average LRT statistic (i.e., the average L_m). Meng and Rubin (1992) recommended comparing D_3 against an F distribution with K numerator and v_3 denominator degrees of freedom. For $a = K(M - 1)$,

$$v_3 = \begin{cases} 4 + (a - 4)[1 + (1 - 2a^{-1})\,\text{ARIV}_3^{-1}]^2 & \text{if } a > 4 \\ (K + 1)(M - 1)(1 + \text{ARIV}_3^{-1})^2/2 & \text{otherwise} \end{cases} . \tag{3.18}$$

The D_3 procedure can be useful in practice because, similar to D_2, is does not require an estimate of the variance-covariance matrix for the quantity of interest. However, because D_3 requires multiple evaluations of the likelihood function, it can be difficult to implement and is currently used primarily in software for structural equation modeling (SEM; see also Asparouhov & Muthén, 2008).

Current recommendations regarding D_3. In the missing data literature, D_3 is often recommended because it is asymptotically equivalent to D_1 (e.g., Schafer, 1997; van Buuren, 2012). However, relatively few studies have evaluated the performance of D_3 (Enders, 2010); the results of those that have suggest that the procedure performs well but that it tends to be more conservative than D_1 in conditions with smaller samples and larger FMIs (Y. Liu & Enders, in press; Meng & Rubin, 1992). Similarly, in Article 4, the performance of D_3 was often comparable with that of D_1 and D_1^*. However, D_3 tended to be more conservative than the other procedures in smaller samples with slightly lower statistical power. In addition, the procedure became more conservative in conditions with very large FMIs (e.g., larger than 60%). Taken together, these results indicated that D_3 is relatively robust against conditions with larger FMIs but slightly worse in terms of statistical power when compared with D_1 and D_1^*.

3.3 Conducting multiparameter tests and model comparisons in multilevel analyses

In the context of multilevel MI, the evaluation of multiparameter tests is particularly interesting because the different test statistics may be used in a more or less flexible manner for various kinds of hypothesis tests. For example, hypotheses about fixed effects are usually tested with Wald-like hypothesis tests (D_1, D_1^*, and D_2; for a discussion, see Manor & Zucker, 2004; Snijders & Bosker, 2012b). By contrast, variance components are often tested with LRTs (D_2 and D_3), especially with multilevel software that constrain [6] the variance components to be positive (Snijders & Bosker, 2012b). In the following, I provide the results from two additional simulation studies, which evaluated the performance of different procedures for multiparameter tests with respect to (a) fixed effects and (b) variance components in multilevel MI.

3.3.1 *Fixed effects*

The first simulation study examined the performance of the different pooling methods—D_1, D_1^*, D_2, and D_3—for testing a subset of the fixed effects in a multilevel random intercept model. Specifically, the data were generated from the following multilevel model. For individual i ($i = 1, \ldots, n$) in group j ($j = 1, \ldots, J$),

$$y_{ij} = \gamma_{10}(x_{ij} - \bar{x}_{\bullet j}) + \gamma_{01}\bar{x}_{\bullet j} + \gamma_{02}^{(1)}d_j^{(1)} + \ldots + \gamma_{02}^{(K)}d_j^{(K)} + u_{0j} + e_{ij}\,, \tag{3.19}$$

where $(x_{ij} - \bar{x}_{\bullet j})$ and $\bar{x}_{\bullet j}$ denote the within- and between-group components of a continuous explanatory variable X at Level 1 with regression coefficients γ_{10} and γ_{01}, and $\mathbf{d}_j = (d_j^{(1)}, \ldots, d_j^{(K)})$ denotes the values of a set of dummy indicator variable representing the $K + 1$ levels of a categorical explanatory variable D at Level 2 with regression coefficients $\boldsymbol{\gamma}_{02} = (\gamma_{02}^{(1)}, \ldots, \gamma_{02}^{(K)})$. The multiparameter test was concerned with the effect of the categorical variable D, that is, with the simultaneous test of the regression coefficients $\boldsymbol{\gamma}_{02}$ against zero. For example, D may represent school types in educational research. In this case the multiparameter test can be used

[6] The use of D_1 and D_1^* for variance components may be inappropriate because their sampling distribution may be skewed due to the constrained estimation procedures in some multilevel software (e.g., `lme4`). However, with *un*constrained estimation (e.g., M*plus*), D_1 and D_1^* may perform well; however, this option is not further explored here in detail (for a discussion, see Savalei & Kolenikov, 2008; Stoel, Garre, Dolan, & van den Wittenboer, 2006).

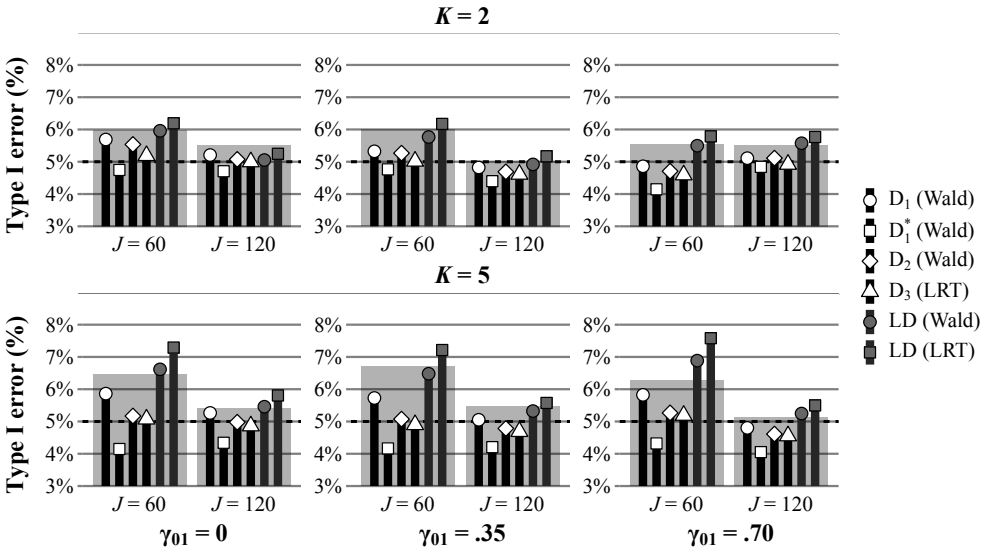

Figure 3.1: Type I error rates for different pooling methods and LD ($\alpha = 5\%$) depending on the sample size at Level 2 (J), the number of parameters being tested (K), and the effect of X (γ_{10} and γ_{01}). The shaded areas represent the results in complete-data. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

to test for overall differences between school types in the outcome variable Y, where the null hypothesis states that there are no differences between school types.

Missing values were induced in Y as a function of X (MAR, 25%). For simplicity, the sample size at Level 1 was fixed at $n = 10$, and the ICCs of X and Y were fixed at .20. I varied the number of groups ($J = 60, 120$) and the number of parameters being tested ($K = 2, 5$), which corresponds to conditions with 3 and 6 levels of the categorical variable D, respectively. The overall effect of X was varied ($\gamma_{10} = \gamma_{01} = 0, .35, .70$) in order to allow for conditions with different FMIs. The parameters $\boldsymbol{\gamma}_{02}$ were specified in such a way that either (a) the differences between the categories in D were all zero or (b) the coefficients representing one third of the categories were set to .35. The missing values were treated with LD and multilevel JM (number of imputations, $M = 100$). In line with current recommendations in the multilevel literature (e.g., Snijders & Bosker, 2012b), the multiparameter test for the fixed effects was carried out using both Wald-tests with standard errors based on restricted maximum likelihood estimation (REML; applicable with LD, D_1, D_1^*, and D_2) and LRTs on the basis of ML estimation (applicable with LD and D_3). To calculate D_1^*, the complete-data degrees of freedom were set to $J - (K + 1) - 1$ (see also Manor & Zucker, 2004).

The results are summarized in Figures 3.1 and 3.2. As can be seen, the Type I error rates

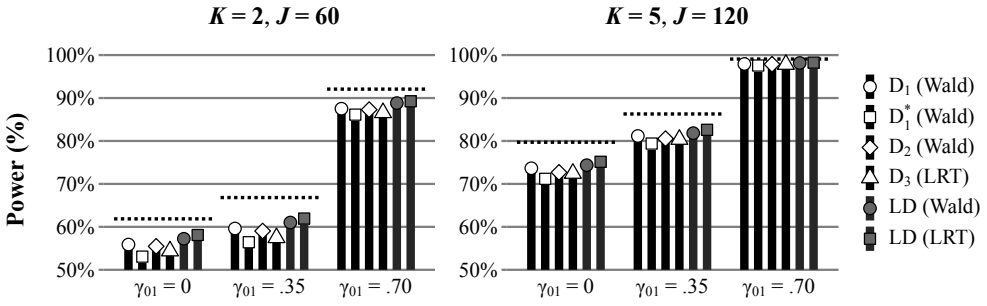

Figure 3.2: Power for different pooling methods and LD (α = 5%) depending on the sample size at Level 2 (J), the number of parameters being tested (K), and the effect of X (γ_{10} and γ_{01}). The dotted lines indicate the power obtained in complete data. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

usually remained within reasonable bounds under MI (i.e., between 2.5% and 7.5%, given α = 5%). However, the procedures tend to differ in comparison with one another. First, the Type I error rates tended to be slightly higher in conditions with smaller samples at Level 2 (J = 60) under LD, which further reduced sample size, as well as D_1 and (to a lesser extent) D_2. By contrast, D_1^* and (to a lesser extent) D_3 tended to be more conservative, especially in smaller samples (J = 60), a larger number of parameters being tested (K = 5), and larger FMIs ($\gamma_{10} = \gamma_{01} = 0$). The results for the statistical power essentially matched the differences in Type I error rates with slightly higher power under D_1, D_2, and LD as compared with D_1^* and D_3. The power was strongly increased in conditions with larger effects of X, which may be attributed both to the differences in the expected power (i.e., by reducing the residual variance at Level 2), and a reduction in the FMI (i.e., by increasing the information available about missing Y). Due to the simulation design, the reduction of the FMI was also beneficial for statistical power under LD; however, this may not be the case if auxiliary variables are available that are not included in the analysis model (see also Article 4). In summary, the results under MI maintained Type I error rates close to the nominal value with only small differences in statistical power, this lending support to the use of all procedures, including D_2, within the scope of the simulated conditions (see also Article 4).

3.3.2 *Variance components*

The second simulation study examined the performance of the different methods conducting multiparameter hypothesis tests for testing variance components, that is, the slope variance, in a

multilevel random coefficients model. The data were generated from the following model. For individual i ($i = 1, \ldots, n$) in group j ($j = 1, \ldots, J$),

$$y_{ij} = \gamma_{10}(x_{ij} - \bar{x}_{\bullet j}) + \gamma_{01}\bar{x}_{\bullet j} + u_{0j} + u_{1j}(x_{ij} - \bar{x}_{\bullet j}) + e_{ij} \,, \tag{3.20}$$

where u_{1j} denotes the random slopes pertaining to the within-group components of X, and the random effects (u_{0j}, u_{1j}) were assumed to follow a multivariate normal distribution. In multilevel research, the slope variance is often tested using the LRT by comparing the model of interest with a reduced model which included the same fixed effects but only the random intercept (see Snijders & Bosker, 2012b). This test involves $K = 2$ parameters: the slope variance and the intercept-slope covariance. Therefore, the test can be carried out by comparing the LRT statistic with a χ^2 distribution with two degrees of freedom. However, if the variances are constrained to be larger than zero during model estimation, it has been argued that the LRT statistic under the null hypothesis follows a 50/50 mixture of two χ^2 distributions with two and one degrees of freedom, respectively, resulting in overly conservative inferences with the standard procedure (Self & Liang, 1987; Stram & Lee, 1994; see also LaHuis & Ferguson, 2009).

Missing data were induced in Y dependent on X (MAR, 25%). The sample size at Level 1 was fixed at $n = 10$, and the ICCs of X and Y were fixed at .20. Similar to the previous study, I varied the number of groups ($J = 50, 100$), the slope variance ($\tau_1^2 = 0, .05, .10$), and the overall effect of X ($\gamma_{10} = \gamma_{01} = 0, .35, .70$) to allow for conditions with different FMIs. The missing data were treated using LD and multilevel JM (number of imputations, $M = 100$). In contrast to before, I implemented multilevel JM in two different ways: with standard "least-informative" priors and with data-dependent priors[7] on the basis of LD (Grund, Lüdtke, & Robitzsch, 2016a). The slope variance was tested using LRTs on the basis of ML using LD, D_2, and D_3, each compared with both the standard and the mixture-χ^2 distribution. To this end, D_2 and D_3 were multiplied by K, where the transformed statistics $D_2 \cdot K$ and $D_3 \cdot K$ asymptotically

[7]The use of data-dependent priors is a controversial topic in the statistical literature (e.g., Gelman et al., 2014). However, it has been shown that least-informative priors can lead to biased estimates of the variance components in multilevel models, especially when the variance components are small (e.g., McNeish, 2016). In practice, the use of data-dependent priors may be avoided by specifying a "prior guess" for the variance-covariance matrix of the random effects (see Grund, Lüdtke, & Robitzsch, 2016a; Schafer & Yucel, 2002).

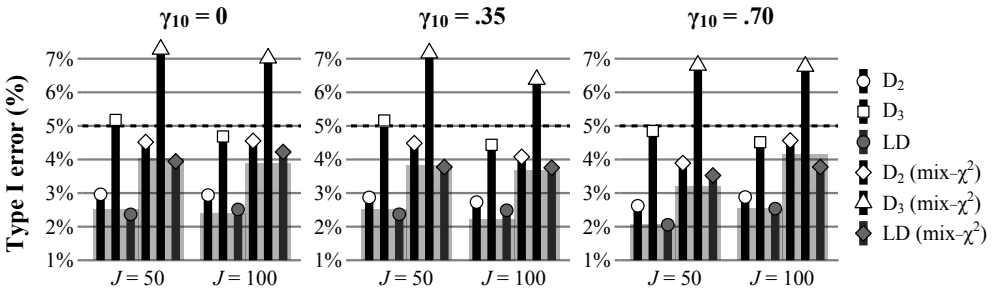

Figure 3.3: Type I error rates for different pooling methods and LD (α = 5%) depending on the sample size at Level 2 (J) and the effect of X (γ_{10} and γ_{01}). The shaded areas represent the results in complete-data for the standard test (left) and the mixture-χ^2 (right). D_2, D_3 = pooling methods; LD = listwise deletion.

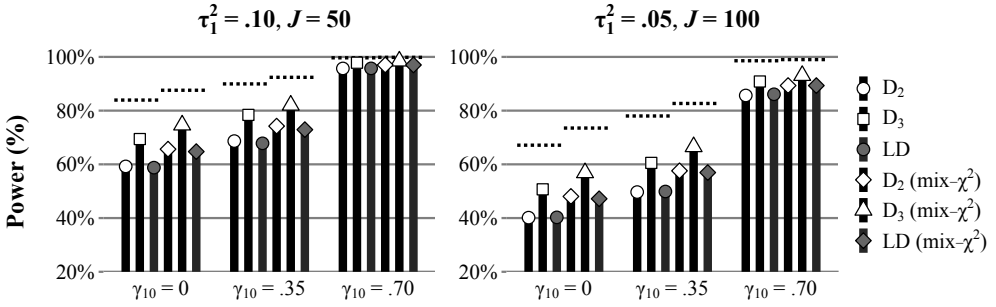

Figure 3.4: Power for different pooling methods and LD (α = 5%) depending on the sample size at Level 2 (J), the true slope variance (τ_1^2), and the effect of X (γ_{10} and γ_{01}). The dotted lines indicate the power obtained in complete data for the standard test (left) and the mixture-χ^2 (right). D_2, D_3 = pooling methods; LD = listwise deletion.

follow a distribution to a χ^2 distribution with K degrees of freedom—or a mixture-χ^2 under the null—as the denominator degrees go to infinity (see also Asparouhov & Muthén, 2008).

The results are summarized in Figures 3.3 and 3.4. For simplicity, I do not consider the results obtained with least-informative priors because they introduced large biases into the estimates of the slope variance in some conditions, leading to inflated Type I error rates under MI. In contrast with the results for the fixed effects, D_3 tended to provide a more liberal test of the slope variance as compared with D_2 and LD. Specifically, Type I error rates under D_3 were relatively close to the nominal level (α = 5%) with the standard test but exceeded the nominal level with the mixture-χ^2. By contrast, the Type I error rates with D_2 and LD were more similar to those in the complete data and were closer to the nominal when compared with the mixture-χ^2. The results for the statistical power mostly resembled the differences in Type I error rates, with slightly larger power with D_3 as compared with D_2 and LD and slightly larger power with the mixture-χ^2 as compared with the standard test. In summary, the

differences between pooling methods were again relatively small in the selected conditions. However, the results obtained from D_3 were surprisingly liberal, whereas D_2—though slightly more conservative—maintained Type I error rates close to the nominal value and benefited from the mixture-χ^2 in a manner similar to the complete data.

3.4 Summary

In the present chapter, I summarized the most widely known procedures for obtaining parameter estimates and inferences from multiply imputed data set. These procedures include relatively general procedures for scalar and multidimensional estimands (e.g., one or several regression coefficients) and methods for testing statistical hypotheses about these parameters (e.g., Wald tests, LRTs). However, even though recommendations tend to favor some procedures (D_1, D_1^*, and D_3) over others (D_2), little is known about the performance of these methods in research practice. The present dissertation contributes to this topic in a number of ways. First, in Article 4, the performance of these procedures was compared in the context of the ANOVA, including a large number of simulated conditions and a varying number of imputations of each method. Second, the same procedures were evaluated in two additional simulation studies that were concerned with multilevel analyses. The results indicated that (a) D_1 and D_1^* are often the most reliable among the procedures but also that (b) D_2 performs well in a surprisingly large range of conditions provided that the number of imputations is reasonably large.

In addition to these findings, many open questions remain about the analysis of multiply imputed data sets. For example, researchers often wish to assess how well statistical models "fit" the data by examining "goodness-of-fit" indices (Bentler, 1990; Bentler & Bonett, 1980); however, not much is known about how these measures can be obtained from the imputed data and how the resulting statistics perform in practice (see also Enders & Mansolf, in press; Kientoff, 2011). Furthermore, very few studies have investigated the use of procedures for model selection such as the LASSO (Tibshirani, 1996) or the elastic net (Zou & Hastie, 2005) as well as resampling-based methods such as the bootstrap (Efron, 1981) when missing data are treated with MI (Q. Chen & Wang, 2013; Claeskens & Consentino, 2008; Geronimi & Saporta,

2017; Heymans, van Buuren, Knol, van Mechelen, & de Vet, 2007). Finally, the analysis of multiply imputed data sets is further complicated if imputations are generated in multiple stages, resulting in "nested imputations" of missing data (e.g., Rubin, 2003). Such applications are common in educational large-scale assessments and require the use of an adjusted set of rules for pooling parameter estimates and inferences (Reiter & Raghunathan, 2007; Shen, 2000). Because many of these methods are frequently used in educational and psychological research, they should be considered in future studies.

Despite these interesting theoretical questions, the challenges associated with conducting analyses under MI are often of a very practical nature. Specifically, relatively few statistical software packages—especially in the context of multilevel modeling—include an implementation of procedures for multiparameter tests and model comparisons under MI. This is problematic because using complex statistical procedures such as D_3 then requires programming skills and formal statistical knowledge. For this reason, the following chapter introduces the R package `mitml`, which is intended to provide a more user-friendly interface for specifying the imputation model as well as a fully automated set of tools for analyzing multiply imputed data sets with particular emphasis on multilevel MI.

4

The R package `mitml`

Thus far, the present dissertation has been focused on the theoretical aspects of multilevel MI. However, despite the tremendous advances in the methodological and statistical literature on multilevel MI over the past years, few studies appear to be using it to treat missing data in research practice (Diaz-Ordaz et al., 2014; Jelicic et al., 2009; Nicholson et al., 2017; Peugh & Enders, 2004; see also Chapter 1). I argue that this is, at least in part, because (a) the literature is still lacking accessible introductory articles about multilevel MI, and (b) the current implementations of multilevel MI tend to be technically very sophisticated and often require programming skills or advanced statistical knowledge to use them effectively. For this reason, one of the goals of the present dissertation was (a) to provide a comprehensive tutorial on the use of multilevel MI, and (b) the development of the R package `mitml`, which provides simple and automated procedures for the imputation of missing data and the analysis of multiple imputed data sets. In this chapter, I provide a brief overview of `mitml` with the aid of an illustrative data example. In addition, the chapter includes a summary of Article 5 of the present dissertation, which provides an in-depth tutorial for multilevel MI using `mitml` in the statistical software R.

4.1 Multiple imputation in practice

Despite the theoretical appeal of MI, conducting multilevel MI can be a daunting task because researchers need to incorporate a number of additional steps in their analytical efforts. For

example, a typical application of multilevel MI requires specifying the imputation model, running the procedure to generate imputations, checking convergence, fitting the analysis model to each data set, and pooling the results to obtain final parameter estimates and inferences. The `mitml` package attempts to provide a comprehensive set of tools for each of these steps, enabling users to follow a simple workflow that requires only a minimum of computer programming skills. In the present chapter, I outline the core features of the package with a number of examples and notes on their implementation.

4.1.1 *Features of* `mitml`

The `mitml` package is available in the statistical software R (R Core Team, 2016) and can be installed from the Comprehensive R Archive Network (CRAN). To illustrate the implementation of the different features in `mitml` and the intended workflow, I make use of the `leadership` data set, which contains artificial data from 750 employees (Level 1) in 50 work teams (Level 2), including data on the teams' cohesion and the employees' work load (categorical, high/low), job satisfaction, and ratings on negative leadership style. As illustrated below, all variables contain missing data.

```
library(mitml)    # load package
data(leadership)  # load data
```

```
#     GRPID          JOBSAT            COHES            NEGLEAD          WLOAD
# Min.   : 1.0   Min.   :-7.32934   Min.   :-3.4072   Min.   :-3.13213   low :416
# 1st Qu.:13.0   1st Qu.:-1.61932   1st Qu.:-0.4004   1st Qu.:-0.70299   high:248
# Median :25.5   Median :-0.02637   Median : 0.2117   Median : 0.08027   NA's: 86
# Mean   :25.5   Mean   :-0.03168   Mean   : 0.1722   Mean   : 0.04024
# 3rd Qu.:38.0   3rd Qu.: 1.64571   3rd Qu.: 1.1497   3rd Qu.: 0.79111
# Max.   :50.0   Max.   :10.19227   Max.   : 2.5794   Max.   : 3.16116
#                NA's   :69         NA's   :30        NA's   :92
```

To generate imputations, `mitml` builds on two existing packages that implement the JM approach to multilevel MI: the `pan` package (Schafer & Yucel, 2002), which can be used to address missing data in continuous variables at Level 1, and the `jomo` package (Quartagno, 2016), which extends this functionality to mixed continuous and categorical variables at both Level 1 and 2. In the present case, I will use `jomo`.

Specifying the imputation model. With `mitml`, the imputation model can be specified in two different ways. First, the imputation model can be specified as a formula similar to the R

package lme4 (Bates, Mächler, Bolker, & Walker, 2015). In the present case, the imputation model comprises two components, one pertaining to variables at Level 1 and the other to variables at Level 2 (see Chapter 2). Specifically, with missing data in all variables in the data set, the imputation model is specified as follows.

```
fml <- list( JOBSAT + NEGLEAD + WLOAD ~ 1 + (1|GRPID) , # Level-1 model
             COHES ~ 1 )                                # Level-2 model
```

The first entry in the list denotes the imputation model for missing data at Level 1. The second entry denotes the imputation model for the variable at Level 2. The ~ symbol separates the target and predictor variables in the model. Here, the predictor side includes only a 1 for the intercept. The | operator denotes the clustering variable as well as the random effects to be included in the model (i.e., random intercepts and slopes). Here, the data are clustered by GRPID and the model includes only a random intercept; further random effects may be included for completely predictor variables should these be available. Categorical variables are recognized automatically, provided that they are formatted as factors in R.

As an alternative, the imputation model can be specified using an integer vector denoting the "type" of each variable (i.e., its role in the imputation model). The corresponding type vectors equivalent to the formulas above are as follows.

```
type <- list( c( -2, 1, 0, 1, 1) , # Level-1 model
              c( -2, 0, 1, 0, 0) ) # Level-2 model
```

The integer values -2 and 1 denote the cluster variable and the target variables in the model, respectively. In addition, the values 2 and 3 can be used to include predictor variables with fixed and random effects, respectively (see the package documentation). The type interface can be helpful when dealing with large data sets, where writing formulas can be tedious. By contrast, the formula interface is more convenient and easy to understand; it is also be very flexible because, similar to general formulas in R, it can be used to include functions of predictor variables as additional predictors in the model (e.g., group means, squared terms, or interactions).

Running MI. Given the imputation model represented as a model formula or type vector, the imputation procedure can be run by calling one of the wrapper functions panImpute (for

pan) or `jomoImpute` (for `jomo`). This requires the specification of the number of iterations and imputations for which the procedure should run. In the present example, the imputation is run as follows.

```
imp <- jomoImpute(data=leadership, formula=fml, n.burn=5000, n.iter=250, m=20)
```

The total number of iterations is determined by the number of burn-in iterations (`n.burn`), which are performed before any imputations are generated, the number of iterations between imputations (`n.iter`), and the number of imputations (`m`). In addition, the user has the option to designate an additional grouping variable, in which case imputations are generated separately within the levels of that variable, to specify the Bayesian prior distributions for the parameters of the imputation model, and to pass other parameters to the function that can be used to influence the behavior of the procedure (see the package documentation). In particular, with `jomoImpute`, it is possible to allow for heterogeneity in the Level-1 residual covariance matrix across groups (see Quartagno & Carpenter, 2016b; Yucel, 2011).

Convergence diagnostics. Once the imputation is completed, users are required to ensure that the parameter chains of the imputation procedure converged to stationary distributions (Gelman et al., 2014). The `mitml` package offers two options to do so. First, convergence statistics can be calculated with the `summary` function, which includes the $\hat{R}$ criterion (Gelman & Rubin, 1992) as well as (optionally) the autocorrelation at lag k and $2k$, where k is the number of iterations between imputations, and a measure for the goodness of approximation in the central tendency of the posterior distribution (see Hoff, 2009). For example, the default command which requests only $\hat{R}$ is as follows.

```
summary(imp)
```

```
# Call:
#
# jomoImpute(data = leadership, formula = fml, n.burn = 5000, n.iter = 250,
#                m = 20)
#
# Level 1:
#
# Cluster variable:         GRPID
# Target variables:         JOBSAT NEGLEAD WLOAD
# Fixed effect predictors:  (Intercept)
# Random effect predictors: (Intercept)
#
```

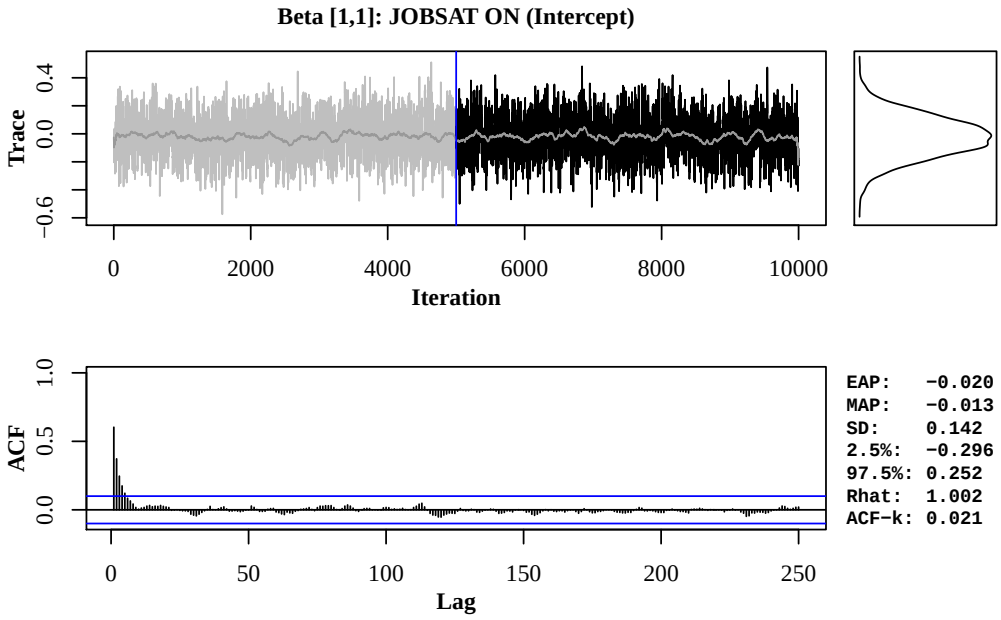

Figure 4.1: Convergence plot for the fixed intercept of job satisfaction (JOBSAT), including the trace plot (top left), the autocorrelation plot (bottom left), the density plot (top right), and the posterior summary (bottom right).

```
# Level 2:
#
# Target variables:         COHES
# Fixed effect predictors:  (Intercept)
#
# Performed 5000 burn-in iterations, and generated 20 imputed data sets,
# each 250 iterations apart.
#
# Potential scale reduction (Rhat, imputation phase):
#
#          Min   25%  Mean Median   75%   Max
# Beta:  1.002 1.002 1.003  1.003 1.004 1.004
# Beta2: 1.000 1.000 1.000  1.000 1.000 1.000
# Psi:   1.000 1.000 1.001  1.001 1.001 1.004
# Sigma: 1.000 1.002 1.008  1.004 1.010 1.025
#
# Largest potential scale reduction:
# Beta: [1,2], Beta2: [1,1], Psi: [4,3], Sigma: [3,1]
```

As can be seen, the output is separated by parameter class and includes a reference to the parameter with the most problematic value (e.g., the largest $\hat{R}$) in order to assist users in finding the source of convergence problems, should they occur. As a second option, the `mitml` package offers diagnostic plots to assess convergence in a graphical manner (see also Schafer & Olsen, 1998). The plots can be requested as follows, with an example given in Figure 4.1.

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

The diagnostic plots include the trace plot for each parameter in the imputation model, a plot for the autocorrelation in the parameter chain, a density plot, and a summary of the posterior distribution (for a general discussion of convergence in Bayesian data analysis, see also Gelman et al., 2014; Gill, 2014; Hoff, 2009; Jackman, 2009).

Transforming and analyzing data. Having assessed that the algorithm converged, a list containing the imputed data sets can be extracted from the imputed data object as follows.

```
implist <- mitmlComplete(imp, "all")
```

In order to manipulate and analyze the data, `mitml` implements additional methods for the generic functions `within` and `with` from base R. First, the `within` function can be used to evaluate a given expression in each of the imputed data sets, thus creating transformations of the imputed data. For example, the following code illustrates this for the calculation of group means and the group mean centering of employees' ratings of negative leadership style.

```
implist <- within( implist,{
  M.NEGLEAD <- clusterMeans(NEGLEAD, GRPID)  # calculate group means
  I.NEGLEAD <- NEGLEAD - M.NEGLEAD           # group mean centering
})
```

Second, the analysis of the imputed data sets can be carried out with the `with` function. Formally, `with` also evaluates an expression in each data set but returns the result of the evaluated expression instead. For example, the following command fits a multilevel model to each of the imputed data sets using the R package `lme4`, where job satisfaction is explained by negative leadership style, work load, and cohesion.

```
library(lme4)
fit <- with( implist, lmer(JOBSAT ~ 1 + I.NEGLEAD + M.NEGLEAD + WLOAD + COHES + (1|GRPID)) )
```

This results in a list of fitted models, one for each of the imputed data sets, the results of which can be pooled in subsequent steps.

Pooling. The `mitml` package offers several function for pooling the results obtained from multiply imputed data sets. In the simplest case, parameter estimates and inferences can be obtained by pooling the individual (i.e., scalar) estimands in the fitted models. This can be achieved with the `testEstimates` function as follows.

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

# Call:
#
# testEstimates(model = fit, var.comp = TRUE)
#
# Final parameter estimates and inferences obtained from 20 imputed data sets.
#
#              Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
# (Intercept)     0.246     0.141     1.740  1084.906     0.082     0.153     0.134
# I.NEGLEAD      -0.536     0.088    -6.112   553.853     0.000     0.227     0.188
# M.NEGLEAD      -1.516     0.352    -4.303   625.680     0.000     0.211     0.177
# WLOADhigh      -0.828     0.190    -4.352   698.799     0.000     0.197     0.167
# COHES           0.234     0.094     2.481  1691.155     0.013     0.119     0.107
#
#                             Estimate
# Intercept~~Intercept|GRPID    0.315
# Residual~~Residual            4.966
# ICC|GRPID                     0.060
```

By default, `testEstimates` employs Rubin's rules for pooling individual parameters, but the small-sample correction by Barnard and Rubin (1999) can be applied by providing the complete-data degrees of freedom as an additional argument to the function call (`df.com`). This method is automatic for all model classes that define methods for the generic functions `coef` and `vcov`, which includes most of the standard models in R (e.g., linear and generalized linear models) as well as some additional[8] model classes (e.g., multilevel models, generalized estimating equations).

In addition, pooling methods for more complex statistical hypothesis are provided with `mitml`. This includes the procedures for multiparameter tests and model comparisons discussed in Chapter 3. Specifically, the function `testModels` allows the comparison of different statistical models using either Wald-tests (D_1, D_1^*, and D_2) or LRTs (D_2 and D_3). The procedures for pooling Wald-tests (D_1, D_1^*, and D_2) are again automatic for model classes that define methods for `coef` and `vcov` and some additional models. In addition, D_2 provides an implementation for pooling LRTs that is automatic for model classes that define methods for the generic function `logLik` as well as some additional models (e.g., the Cox proportional hazards model). For D_3, which cannot be implemented in a generic manner, `mitml` includes implementations specific

[8]It is worth noting that the framework of generic functions in R provides a simple method for extending the pooling methods to additional model classes because the required methods for `coef` and `vcov` (or similar functions) can simply be defined by `mitml` if they are not defined by the model classes.

to linear models and (linear) multilevel models with a single level of clustering (i.e., two-level models). The `anova` function provides a convenience wrapper for likelihood-based comparisons using D_2 and D_3. Finally, the `testConstraints` function includes an implementation of the delta method for multiply imputed data sets which can be used to test contrasts and constraints on the model parameters (e.g., Casella & Berger, 2002; Fox, 2008). This procedure is based on the rules for pooling multiparameter Wald-tests (D_1, D_1^*, and D_2) and is thus automatic provided that `coef` and `vcov` methods exist.

4.1.2 *Practical guidelines on multilevel MI*

It may be argued that, in addition to comprehensive tools in statistical software, the missing data literature is still in need of accessible tutorials on conventional and multilevel MI. For this reason, Article 5 of the present dissertation includes an in-depth tutorial about multilevel MI. Specifically, the article includes examples for multilevel MI using the data from the German subsample of primary school students in the Progress in International Reading Literacy Study (PIRLS). In two examples—one concerning the multilevel random intercept model, the other concerning a model with random slopes—the tutorial illustrates multilevel JM using the R packages `pan` and `mitml`. In that context, particular emphasis is placed on the correct specification of the imputation model, which is often crucial in multilevel MI (see also Enders et al., 2016; Lüdtke et al., 2017). In addition, the article discusses the assessment of convergence and the analysis of multiply imputed data sets, including examples for multiparameter tests, that is, model comparisons (e.g., inclusion of random slopes) and tests for constraints on the model parameters (e.g., contextual effects).

4.2 Summary

In the present chapter, I focused on the practical aspects of multilevel MI in greater detail. In this context, I summarized the main features of the R package `mitml`, which is intended to provide a user-friendly interface to multilevel MI, as well as the contents of Article 5, which provides an introduction in the form of a tutorial article for researchers who are not yet familiar

with multilevel MI. In line with my original argument stating that the current literature on multilevel MI is still lacking clear introductions to the topic and accessible implementations in statistical software, these contributions can be regarded as an effort to bridge the divide between theory and practice of multilevel MI, thus promoting a more consistent use of these methods in psychological research.

Over the past years, many interesting developments have been proposed in the literature on missing data and multilevel MI. However, as outlined earlier, there are still a number of open questions regarding the imputation of multilevel data (see Chapter 2) and the analysis of multiply imputed data sets (see Chapter 3). For both of these reasons, it is important that new and improved procedures are made available to a wide audience of researchers in the form of free and accessible statistical software. For the future, there are many ways in which `mitml` can be improved and extended. This includes technical aspects of the package as well as the continued development of improvements and extensions to the interface and the implementation of new procedures for the imputation of different types of multilevel data (e.g., as implemented in the R package `jomo`; see also Quartagno, 2016) and different methods for analyzing multiply imputed data sets (e.g., with respect to "nested" MI and procedures for model fit, resampling, or model selection). In addition, the existing procedures (e.g., pooling methods) should be further extended to accommodate a wider variety of statistical models, especially in the context of multilevel analyses.

5
Conclusion

The present dissertation was concerned with the treatment of missing data in multilevel research with particular emphasis on multilevel MI. It included five articles that addressed some of the theoretical and practical challenges associated with multilevel MI and also featured an R package that was developed in the context of these studies. In the following conclusion, I provide a short summary of the topics considered in the dissertation and the articles provided with it and a critical discussion of their limitations and weaknesses. In addition, I will discuss some of the more general points regarding missing data and the utility and challenges associated with MI. In this context, I also provide an outlook on possible extensions and topics for future research.

In the first part of the dissertation, I considered the theoretical basis of missing data and MI and discussed the theoretical challenges associated with multilevel MI (Chapters 1 and 2). In particular, I discussed the treatment of missing data in two broad classes of multilevel analysis models: the random intercept model and the random coefficients model with and without CLIs. First, in the context of the multilevel random intercept model, I outlined two computational paradigms—multilevel JM and FCS—for conducting multilevel MI and discussed how these approaches relate to one another both conceptually and computationally (Articles 1 and 2). A major point in this discussion was related to the structure of multilevel data, that is, the components in and relations between the variables in multilevel data. In this context, I argued that the statistical models employed by multilevel JM and FCS are very similar on a conceptual level but not entirely equivalent due to the different use of between-group components in

multilevel variables if the data are unbalanced (Article 3). The second class of multilevel analysis models included multilevel models with random slopes and CLIs. In this context, I provided a discussion about the intricacies associated with multilevel MI when missing data occur in the explanatory variables of the model. Specifically, it can be shown that the current procedures for multilevel MI, though very powerful in a general manner, still have room for improvement in models with random slopes and CLIs; in these cases, modern methods that explicitly take the model of interest into account appear to provide essentially unbiased results (see also Article 2).

In the second part of the dissertation, I considered the analysis of multiply imputed data sets with an emphasis on multiparameter tests and model comparisons. In this context, I outlined some of the more frequently-used procedures for conducting multiparameter tests, namely D_1, D_2, and D_3 (and variants thereof). Although current recommendations in the missing data literature clearly convey that D_1 and D_3 should always be preferred over D_2, several simulation studies that were conducted in the course of this dissertation suggest a more nuanced view. Specifically, D_2 appeared to be reliable and often comparable to the other procedures as long as the number of imputations was appropriately high and the FMI did not become too extreme; a finding that was present both in the context of the ANOVA as well as in multilevel analyses with multiparameter tests of fixed effects and variance components (see also Article 4). Contrary to previous findings, these results illustrate that D_2 may be a useful alternative in many (albeit not all) plausible research scenarios.

Finally, in the third part of the dissertation, I considered the practical aspects of multilevel MI. In line with earlier research regarding the use of missing data methods in psychological research (see also Chapter 1), it appears that multilevel MI is currently not in widespread use; a potential reason for this may be that its application can introduce new problems and may add several steps to conventional statistical analyses. For this reason, the present dissertation made a two-fold attempt to improve the accessibility of multilevel MI. First, the dissertation features a tutorial on multilevel MI, illustrating its use in two examples from educational research (Article 5). Second, it features the R package `mitml`, which aims to provide a user-friendly interface to multilevel MI. The package is free, open-source, and includes a number

of statistical tools to enable a user-friendly workflow without requiring sophisticated statistical expertise or programming skills. Taken together, these contributions build on the previous parts of the dissertations and attempt to make the procedures discussed therein available to a wider audience.

The works in this dissertation come with several limitations and points to consider. For example, the articles enclosed in this dissertation all focused on multilevel data with a two-level structure (i.e., Level 1 and 2). By contrast, multilevel data in general may include multiple levels of nesting, cross-classified random effects, or multiple-membership structures (e.g., Goldstein et al., 2014; Raudenbush & Bryk, 2002; see also Baayen, Davidson, & Bates, 2008; Browne, Goldstein, & Rasbash, 2001; Quené & van den Bergh, 2008). From a theoretical point of view, these applications face similar challenges as the two-level case, and similar procedures have been proposed for the treatment of missing data in three-level (e.g., Yucel, 2008) and cross-classified data (e.g., Clayton & Rasbash, 1999; Rasbash & Browne, 2008; Yucel, Ding, Uludag, & Tomaskovic-Devey, 2008). However, on the practical side, these procedures are not widely available in statistical software, and little is known about their performance in plausible research scenarios. For that reason, a further evaluation of these methods in future studies seems to be warranted.

Furthermore, the present dissertation was focused on applications in which the missing data mechanism is ignorable (i.e., MCAR or MAR). However, in research practice, the missing data mechanism can also be nonignorable (i.e., MNAR), for example, in longitudinal research when dropout can reasonably be assumed to depend directly on the outcome of interest (e.g., Diggle & Kenward, 1994; Molenberghs, Kenward, & Lesaffre, 1997) or in educational achievement tests when missing item responses can be directly related to student ability (e.g., Glas, Pimentel, & Lamers, 2015; Holman & Glas, 2005; Rose, von Davier, & Nagengast, 2017). In such a case, statistical inferences are no longer possible without introducing additional assumptions, for example, by extending the statistical models in such a way that it includes a model for the propensity of the missing data (see also Little & Rubin, 2002; Moustaki & Knott, 2000). In multilevel data, similar questions arise in the context of missing data mechanisms that depend on random effects or other latent variables, which may be regarded as nonignorable (e.g., latent

group means; see also Gottfredson et al., 2017; Little, 1995; Skrondal & Rabe-Hesketh, 2004).

Finally, the works presented in this dissertation focused on the treatment of missing data in conventional multilevel models (e.g., Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012a), whereas multilevel modeling is often conducted in a structural equation modeling (SEM) framework (e.g., Skrondal & Rabe-Hesketh, 2004). It is important to note that the procedures discussed in the present dissertation can also be used for the treatment item-level missing data in multilevel SEM (Gottschall, West, & Enders, 2012; Mazza, Enders, & Ruehlman, 2015; Schafer & Graham, 2002). However, there are a number of interesting questions regarding the use of multilevel MI in the context of SEM that warrant further investigation, for example, the assessment of "goodness-of-fit" (e.g., Enders & Mansolf, in press) or the combined treatment of missing data and measurement error using MI (Blackwell, Honaker, & King, 2017a; Blackwell et al., 2017b). Similar problems also arise in the context of educational large-scale assessment, when latent proficiency scores are estimated (i.e., imputed) with missing data in the context questionnaire (Assmann, Gaasch, Pohl, & Carstensen, 2015; Kaplan & Su, 2016; Weirich et al., 2014).

Future research may consider a number of possible topics. First, the application of multilevel MI is still challenging in certain scenarios and for certain types of research questions, for example, if the model of interest includes random slopes or nonlinear effects such as CLIs. Although previous studies have indicated that model-based procedures that rely on Bayesian or ML estimation can provide accurate results in these cases, these methods are still not well understood, and general implementations in standard software are not yet widely available (however, see Quartagno, 2016). Procedures for multilevel MI may benefit from these developments, thus allowing a wider range of statistical models to be accommodated with multilevel MI (see also Chapter 2). Regarding the analysis of multiply imputed data sets, future studies may consider a wider range of statistical methods that are commonly used in research practice. This includes, for example, the use of "nested" imputations in educational research (e.g., latent proficiency scores and missing data), the calculation and assessment of "goodness-of-fit" indices under MI as well as procedures for model selection (e.g., the LASSO) and nonparametric statistical inference (e.g., the bootstrap; see also Chapter 3). Finally, I believe that it is imperative that future

research continues to provide recommendations and accessible introductions to the increasingly sophisticated methods that are used for multilevel MI. This includes not only the presence of tutorial articles and educational resources on these topics but also a continued implementation of these procedures in user-friendly and accessible statistical software. In conclusion, although the literature on the treatment of missing data in multilevel research has brought fourth many exciting developments during the past years, there are a number of interesting questions that are yet to be considered. In this spirit, I hope that the present dissertation and the works enclosed therein will contribute to the fascinating though ever growing literature on missing data in multilevel research.

Zusammenfassung

In der psychologischen Forschung haben sich Mehrebenenanalysen zu einem der am meisten verwendeten Verfahren zur Analyse hierarchischer Daten entwickelt (z.B. Goldstein, 2011; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012b). Solch hierarchische Daten treten auf, wenn Beobachtungen (Ebene 1) in übergeordneten Einheiten (Ebene 2) organisiert sind, zum Beispiel Schülerinnen und Schüler organisiert in Schulen, Angestellte in Arbeitsgruppen oder Unternehmen sowie in längsschnittlichen Untersuchungen, in denen für alle Teilnehmenden mehrere Beobachtungen vorliegen (für weitere Beispiele siehe z.B. Goldstein et al., 2014; Snijders & Bosker, 2012b). Weiterhin können empirische Daten häufig nicht vollständig erhoben werden und enthalten dadurch fehlende Werte, zum Beispiel weil einige der Teilnehmenden einen Fragebogen nicht vollständig ausgefüllt haben oder einem Teil der Untersuchung gänzlich fernbleiben. In der psychologischen und statistischen Literatur ist bekannt, dass einfache Verfahren zum Umgang mit fehlenden Werten (z.B. listenweiser Fallausschluss) zur Verzerrung statistischer Analysen und der daraus gezogenen Schlussfolgerungen führen können. Demgegenüber werden modernere Verfahren wie die multiple Imputation (MI) sowie die Schätzung mit Maximum-Likelihood (ML) Verfahren allgemein zum Umgang mit fehlenden Werten empfohlen (z.B. Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002).

Die vorliegende Dissertation beschäftigt sich mit dem Umgang mit fehlenden Werten in hierarchischen Daten. Besonderes Augenmerk wird dabei auf die multiple Imputation hierarchischer Daten gelegt. Obwohl sich bereits zahlreiche Arbeiten mit der Behandlung fehlender Werte und

MI beschäftigt haben, bestehen weiterhin offene Fragen und Herausforderungen bezüglich der Anwendung von MI in hierarchischen Daten. Die vorliegende Arbeit liefert eine umfassende Diskussion dieses Themas und setzt sich in fünf Publikationen mit der Nutzung von MI zur Behandlung fehlender Werte in hierarchischen Daten auseinander. In dieser Folge widmete sich die Dissertation auch der Analyse imputierter Daten mit einem Fokus auf Mehrparametertests (Modellvergleiche u.a.). Darüber hinaus stellt die Dissertation ein Softwarepaket vor, welches das Ziel verfolgt, eine einfache Anwendung von MI in hierarchischen Daten sowie eine automatisierte Analyse imputierter Daten in der Forschungspraxis zu ermöglichen. Im Folgenden fasse ich die verschiedenen Teile der Dissertation kurz zusammen.

Fehlende Werte und multiple Imputation

Ein angemessener Umgang mit fehlenden Werten ist für die Ziehung statistischer Inferenzen unerlässlich, vor allem wenn fehlende Werte in systematischer Form auftreten (z.B. Allison, 2001; Enders, 2010). Nach Rubin (1976) können die hypothetischen, vollständigen Daten $\mathbf{Y}$ unterteilt werden in einen beobachteten Teil $\mathbf{Y}_{obs}$ und einen unbeobachteten Teil $\mathbf{Y}_{mis}$. Ein vollkommen zufälliger Ausfall ("missing completely at random", MCAR) liegt vor, wenn die Wahrscheinlichkeit eines Ausfalls unabhängig ist von beobachteten und unbeobachteten Daten, das heißt $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$. Ein bedingt zufälliger Ausfall ("missing at random", MAR) liegt vor, wenn die Wahrscheinlichkeit eines Ausfalls von den beobachteten Daten abhängt, jedoch nach deren Kontrolle von den unbeobachteten Daten unabhängig ist, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs})$. Bei einem nichtzufälligen Ausfall ("missing not at random", MNAR) steht die Wahrscheinlichkeit eines Ausfalls jedoch außerdem mit den unbeobachteten Daten in Beziehung, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. In solchen Fällen ist es nur unter starken Annahmen möglich überhaupt statistische Inferenzen zu ziehen (Carpenter & Kenward, 2013; Little & Rubin, 2002). In der Literatur zu fehlenden Werten ist vielfach gezeigt worden, dass traditionelle Verfahren zum Umgang mit fehlenden Werten (z.B. listenweiser Fallausschluss) nur unter MCAR zu allgemein unverzerrten Schlussfolgerungen führen. Demgegenüber erlauben es moderne Verfahren wie MI und ML die beobachteten Daten vollständig zu nutzen und ermöglichen somit (a) unverzerrte Schlussfolgerungen auch unter MAR und (b) eine höheren Teststärke nachfolgender statistischer Analysen (z.B. Collins

et al., 2001).

Die multiple Imputation (Rubin, 1987) ist ein Verfahren zum Umgang mit fehlenden Werten, in dem plausible Ersetzungen für die fehlenden Werte anhand der beobachteten Daten und eines statistischen Modells erzeugt werden. Die so vervollständigten Datensätze können dann mit konventionellen Verfahren analysiert und deren Ergebnisse in einem einzelnen Ergebnis zusammengefasst werden (Little & Rubin, 2002; Rubin, 1987). In der Praxis können verschiedene Implementationen von MI unterschieden werden, in denen Imputationen entweder anhand eines multivariaten Modells für alle Variablen simultan ("joint modeling", JM) oder schrittweise für jede Variable nacheinander erzeugt werden ("fully conditional specification", FCS). Beispielsweise können kontinuierliche Variablen entweder anhand der multivariaten Normalverteilung (Schafer, 1997) oder mithilfe mehrerer (univariater) Regressionsmodelle imputiert werden (van Buuren et al., 2006; für eine vergleichende Diskussion siehe auch Carpenter & Kenward, 2013). Darüber hinaus werden in der Praxis häufig modellbasierte Verfahren wie etwa ML (auch "full information" ML, FIML) oder Bayesianische Verfahren verwendet, die es erlauben ein Analysemodell direkt anhand der unvollständigen Daten zu schätzen (z.B. Enders, 2010; Little & Rubin, 2002). Trotz der umfangreichen Literatur zu fehlenden Werten und MI, ist aktuell nur wenig darüber bekannt, wie fehlende Werte in hierarchischen Daten behandelt werden und die besonderen Eigenschaften und Anforderungen hierarchischer Daten dabei berücksichtigt werden können (siehe auch Enders et al., 2016; Hox et al., 2016; van Buuren, 2011).

Multiple Imputation hierarchischer Daten

Eine angemessene Anwendung von MI erfordert im Allgemeinen, dass das gewählte Imputationsmodell die Struktur der Daten und die vom Analysemodell implizierten Zusammenhänge mit einbezieht (z.B. Meng, 1994; Schafer, 2003). In hierarchischen Daten muss deshalb die hierarchische Datenstruktur bei der Imputation berücksichtigt werden (siehe auch Andridge, 2011; Drechsler, 2015; Enders et al., 2016; Lüdtke et al., 2017; Taljaard et al., 2008). Die Struktur der Daten ist dabei in zweierlei Weise von Bedeutung. Erstens können Variablen in hierarchischen Daten auf verschiedenen Ebenen erfasst werden, zum Beispiel direkt auf Ebene 1 (z.B. Personen) oder 2 (z.B. Gruppen), und fehlende Werte können auf allen Ebenen auftreten.

Zweitens können die für die Analyse hierarchischer Daten bestimmten Mehrebenenmodelle teilweise komplexe Zusammenhänge zwischen den Variablen auf Ebene 1 und 2 enthalten, zum Beispiel zufällige Effekte von Variablen auf Ebene 1 oder "cross-level"-Interaktionseffekte (CLIs; für eine allgemeine Diskussion siehe Raudenbush & Bryk, 2002; Snijders & Bosker, 2012b). Darüber hinaus können Variablen in hierarchischen Daten in der Regel in zwei Anteile zerlegt werden, die ausschließlich zwischen Gruppen beziehungsweise innerhalb von Gruppen variieren (z.B. Kreft et al., 1995). In diesem Zusammenhang zeigten Lüdtke et al. (2008), dass eine solche Zerlegung sowohl im Sinne *manifester* und *latenter* Gruppenmittelwerte möglich ist, wobei die Bedeutung der Konstrukte auf Ebene 2 und die Schätzung von Zusammenhängen (z.B. Kontexteffekte) von der Art der Zerlegung abhängen kann (siehe auch Asparouhov & Muthén, 2006; Croon & van Veldhoven, 2007; Grilli & Rampichini, 2011).

Trotz einer umfangreichen Literatur bestehen zahlreiche offene Fragen zum Umgang mit fehlenden Werten in hierarchischen Daten. Beispielsweise sind die genauen Auswirkungen der Nutzung manifester und latenter Gruppenmittelwerte unter MI nicht bekannt. Dies ist besonders bedeutsam, da (a) die Nutzung manifester und latenter Gruppenmittelwerte mit unterschiedlichen Schätzungen der Effekte auf Ebene 2 einhergeht und (b) die verschiedenen Implementation von MI sich in der Nutzung der Gruppenmittelwerte unterscheiden. Dies wirft ebenfalls die Frage auf, wie fehlende Werte auf Ebene 2 behandelt und Informationen auf Ebene 1 dabei berücksichtigt werden sollten. Darüber hinaus ist aktuell unklar, wie genau fehlende Werte im Rahmen von Random-Coefficients-Modellen mit oder ohne CLIs behandelt werden können. Die vorliegende Dissertation widmet sich diesen Fragen in insgesamt drei Publikationen. Artikel 1 beinhaltet eine Einführung in die Behandlung fehlender Werte in hierarchischen Daten mit besonderem Fokus auf Random-Intercept-Modelle und betrachtete neben verschiedenen Implementationen von MI (JM und FCS) auch FIML. Artikel 2 erweiterte diese Betrachtungen und berücksichtigte zusätzlich Random-Coefficients-Modelle mit und ohne CLI sowie kategoriale Daten und fehlende Werte auf Ebene 2. Artikel 3 widmete sich schließlich explizit der Imputation fehlender Werte auf Ebene 2, wobei im Besonderen auf die Nutzung manifester und latenter Gruppenmittelwerte eingegangen wurde. Im Folgenden fasse ich den theoretischen Hintergrund und die Befunde dieser Arbeiten kurz zusammen.

Random-Intercept-Modelle. Für die multiple Imputation hierarchischer Daten stehen erneut mehrere Implementationen zur Verfügung. Im Rahmen des "joint modeling"-Ansatzes (JM) können fehlende Werte in hierarchischen Daten auf der Grundlage eines multivariaten Mehrebenenmodells imputiert werden (Goldstein et al., 2009; Schafer & Yucel, 2002). Für fehlende Werte auf Ebene 1 und 2 kann dieses Modell in einer leicht vereinfachten Notation wie folgt ausgedrückt werden:

$$\begin{aligned} \mathbf{y}_{1ij} &= \boldsymbol{\mu}_1 + \mathbf{u}_{1j} + \mathbf{e}_{ij} \quad \text{(Ebene 1)} \\ \mathbf{y}_{2j} &= \boldsymbol{\mu}_2 + \mathbf{u}_{2j}\,, \qquad \text{(Ebene 2)} \end{aligned} \tag{1}$$

wobei $\mathbf{y}_{1ij}$ die Variablen auf Ebene 1 bezeichnet und $\mathbf{y}_{2j}$ die Variablen auf Ebene 2. Die zufälligen Achsenabschnitte der Variablen auf Ebene 1 und die Residuen der Variablen auf Ebene 2, $\mathbf{u}_j = (\mathbf{u}_{1j}, \mathbf{u}_{2j})$, sind multivariat normalverteilt mit Kovarianzmatrix $\boldsymbol{\Psi}$. Die Residuen der Variablen auf Ebene 1, $\mathbf{e}_{1ij}$, sind multivariat normalverteilt mit Kovarianzmatrix $\boldsymbol{\Sigma}$. Durch die Kovarianzstruktur auf Ebene 1 und 2 ($\boldsymbol{\Sigma}$ und $\boldsymbol{\Psi}$) berücksichtigt JM folglich die Zusammenhänge aller Variablen auf beiden Ebenen. Bei näherer Betrachtung wird zudem deutlich, dass die Zusammenhänge auf Ebene 2 durch *latente* Gruppenmittelwerte (d.h. zufällige Effekte) repräsentiert werden (siehe auch Kreft & de Leeuw, 1998; Lüdtke et al., 2008).

Alternativ können fehlende Werte in hierarchischen Daten erneut mithilfe der "fully conditional specification" (FCS) imputiert werden (siehe auch van Buuren, 2011; Yucel et al., 2007). Für fehlende Werte auf Ebene 1 können hierfür univariate Mehrebenenmodelle verwendet werden. Für die p-te Variable mit fehlenden Werten auf Ebene 1 gilt damit:

$$y_{1ijp} = \mathbf{y}_{ij(-p)}\boldsymbol{\beta}_{1p} + u_{1jp} + e_{ijp}\,, \tag{2}$$

wobei $\mathbf{y}_{ij(-p)}$ alle anderen Variablen auf Ebene 1 und 2 bezeichnet sowie die zwischen Gruppen variierenden Anteile (z.B. Gruppenmittelwerte) der Variablen auf Ebene 1. Für die q-te Variable mit fehlenden Werten auf Ebene 2 kann hingegen ein Regressionsmodell verwendet werden:

$$y_{2jq} = \mathbf{y}_{j(-q)}\boldsymbol{\beta}_{2q} + u_{2jq}\,, \tag{3}$$

wobei $\mathbf{y}_{j(-q)}$ alle anderen Variablen auf Ebene 2 bezeichnet sowie die zwischen Gruppen variierenden Anteile (z.B. Gruppenmittelwerte) der Variablen auf Ebene 1. Der FCS-Ansatz berücksichtigt folglich ebenfalls die Zusammenhänge zwischen Variablen auf Ebene 1 und 2,

da die Variablen (und deren Gruppenmittelwerte) als Prädiktoren in die jeweils anderen Imputationsmodelle eingehen. Durch die sequentielle Natur von FCS müssen die Gruppenmittelwerte aktualisiert werden, wenn die zugrundeliegenden Variablen auf Ebene 1 imputiert worden sind. Obwohl FCS prinzipiell sowohl manifeste als auch latente Gruppenmittelwerte nutzen könnte, verwenden aktuelle Implementationen ausschließlich *manifeste* Gruppenmittelwerte.

Vergleich der verschiedenen Verfahren. In Artikel 1 wurden verschiedene Verfahren zur Imputation fehlender Werte (JM und FCS) sowie FIML im Rahmen des Random-Intercept-Modells betrachtet, wobei zusätzlich zwischen Analysemodellen unterschieden wurde, in denen die Schätzung der Effekte auf Ebene 2 anhand manifester oder latenter Gruppenmittelwerte erfolgte. Solange das Analysemodell latente Gruppenmittelwerte verwendete, lieferten alle Verfahren unverzerrte Ergebnisse. Verwendete das Analysemodell jedoch manifeste Gruppenmittelwerte, lieferten nur JM und FCS unverzerrte Ergebnisse, wohingegen FIML (in der statistischen Software M*plus*; L. K. Muthén & Muthén, 2012) zu stark verzerrten Schätzungen der Regressionskoeffizienten auf Ebene 2 führte. Dies kann dadurch erklärt werden, dass FIML zur Behandlung fehlender Werte in unabhängigen Variablen latente Gruppenmittelwerte adaptiert, wodurch sich eine ungewollte Änderung des Analysemodells ergibt (siehe Lüdtke et al., 2008). Dies kann verhindert werden, indem manifeste Gruppenmittelwerte vorab berechnet werden. Diese Prozedur führt ebenfalls zu einer leichten Verzerrung bei systematisch fehlenden Werten (z.B. MAR); die Verzerrung fällt jedoch deutlich geringer aus. In Artikel 2 wurden neben dem Random-Intercept-Modell auch Anwendungen in Random-Coefficients-Modellen mit und ohne CLI sowie mit kategorialen Variablen und fehlenden Werten auf Ebene 2 untersucht. Die Befunde zum Random-Intercept-Modell waren im wesentlichen identisch zu jenen in Artikel 1. Darüber hinaus zeigten die Ergebnisse aus den zusätzlichen Bedingungen mit fehlenden Werten in kategorialen Variablen und auf Ebene 2, dass MI (sowohl JM als auch FCS) in diesem Fällen zu ebenfalls unverzerrten Ergebnissen führt. Allerdings konnten diese Ergebnisse noch nicht die Frage beantworten, ob auch eine *formale* Äquivalenz zwischen JM und FCS besteht beziehungsweise zwischen der Nutzung von manifesten und latenten Gruppenmittelwerten. In Artikel 2 wurden weiterhin Probleme gängiger Imputationsverfahren in Anwendungen des Random-Coefficients-Modells identifiziert. Diese Fragen betrachte ich im Folgenden gesondert.

Fehlende Werte auf Ebene 2. In Artikel 3 wurden verschiedene Ansätze zur Imputation fehlender auf Ebene 2 untersucht. Bislang haben nur wenige Studien fehlende Werte auf Ebene 2 überhaupt berücksichtigt (Enders et al., 2016; van Buuren, 2011; siehe auch Cheung, 2007; Gibson & Olejnik, 2003). Insbesondere stellt sich die Frage, wie genau die zwischen Gruppen variierenden Anteile der Variablen auf Ebene 1 in das Imputationsmodell für fehlende Werte auf Ebene 2 mit einfließen sollten (d.h. als manifeste oder latente Gruppenmittelwerte). Frühere Arbeiten zur Beziehung zwischen manifesten und latenten Gruppenmittelwerte zeigten für den Fall balancierter Stichproben (d.h. mit Gruppen identischer Größe), dass manifeste und latente Gruppenmittelwert zu identischen Ergebnissen führen (z.B. Carpenter & Kenward, 2013). Kürzlich argumentierten Resche-Rigon und White Resche-Rigon and White (in press) jedoch in Bezug auf fehlende Werte auf Ebene 1, dass im Falle unbalancierter Daten die bedingte Verteilung der Variablen nicht nur von den manifesten Gruppenmittelwerten sondern auch von der Gruppengröße abhängt. Resche-Rigon und White empfahlen daher die Verwendung eines Imputationsmodells mit heteroskedastischen Fehlertermen auf Ebene 1. Artikel 3 erweiterte diese Argumentation in zweierlei Hinsicht. Erstens wurde anhand mathematischer Herleitungen gezeigt, dass die Verwendung manifester Gruppenmittelwerte im Falle unbalancierter Daten zu (negativ) verzerrten Schätzungen der Kovarianzen und Regressionskoeffizienten auf Ebene 2 führen kann. Obwohl das Ausmaß der Verzerrung in der Regel gering ausfällt, wurde gezeigt, dass die Verzerrungen umso größer ausfällt, je kleiner die Gruppen insgesamt sind und je stärker sie in ihrer Größe variieren. Zweitens wurde eine alternative Spezifikation von FCS präsentiert, die mithilfe der "plausible values"-Technik (Mislevy, 1991) eine direkte Berücksichtigung der latenten Gruppenmittelwerte in FCS erlaubt (für eine ähnliche Anwendung siehe Yang & Seltzer, 2016). In zwei Simulationsstudien wurde gezeigt, dass dieser Ansatz selbst in stark unbalancierten Stichproben zu unverzerrten Schätzungen von Kovarianzen und Regressionskoeffizienten auf Ebene 2 führt.

Random-Coefficients-Modelle und CLIs. Während die Behandlung fehlender Werte im Kontext von Random-Intercept-Modellen relativ unproblematisch sowohl mit JM als auch mit FCS möglich ist, zeigten die Simulationsstudien in Artikel 2, dass die aktuell verfügbaren Verfahren zur Imputation hierarchischer Daten im Rahmen von Random-Coefficients-Modellen zu

teils verzerrten Parameterschätzungen führen können. Während diese Verfahren im Falle fehlender Werte in der abhängigen Variable noch zu unverzerrten Ergebnissen führten, führte die Anwendung von MI (sowohl JM als auch FCS) im Falle fehlender Werte in den unabhängigen Variablen zu teils verzerrten Ergebnissen. Dies war insbesondere der Fall für Schätzungen der Slopevarianz, die unter FCS nur durch das "Umdrehen" des Imputationsmodells adressiert werden konnte (Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016a), sowie für Schätzungen der CLI, die unter FCS mithilfe der "passiven Imputation" mit einbezogen wurde. Aus statistischer Sicht sind diese Befunde nicht überraschend, da in mehreren Arbeiten gezeigt wurde, dass Analysemodelle mit nicht-linearen Effekten (z.B. der CLI) zu komplexen bedingten Verteilungen der Variablen führen, die mithilfe konventioneller Imputationsverfahren nicht adäquat berücksichtigt werden können (S. Kim et al., 2015; siehe auch Seaman et al., 2012; von Hippel, 2009; Zhang & Wang, 2016). In diesem Zusammenhang wurden kürzlich alternative Verfahren vorgeschlagen, die das Analysemodell bei der Behandlung (bzw. Imputation) der fehlenden Werte explizit mit berücksichtigen (Bartlett et al., 2015; Goldstein et al., 2014; siehe auch Stubbendick & Ibrahim, 2003). Für die Anwendung in Mehrebenenanalysen sind diese Verfahren bislang noch nicht in konventioneller Software verfügbar und erfordern deshalb die Nutzung allgemeiner Software für Bayesianische Analysen (z.B. Erler et al., 2016). In Artikel 2 wurde die Schätzung des Random-Coefficientes-Modells anhand modellbasierter, Bayesianischer Verfahren in einer Simulation näher betrachtet und im Rahmen der Dissertation noch weiter ausgebaut. Die Ergebnisse dieser Simulationsstudien legen nahe, dass modellbasierte, Bayesianische Verfahren zu unverzerrten Parameterschätzungen in Random-Coefficients-Modellen mit und ohne CLI führen können. Dies war auch der Fall, wenn diese Verfahren verwendet wurden, um Imputation für die fehlenden Werte zu erzeugen, die anschließend mit konventionellen Methoden ausgewertet wurden (siehe auch Quartagno, 2016).

Die Analyse multipel imputierter Daten

Die vorliegende Dissertation beschäftigte sich ebenfalls mit der Analyse multipel imputierter Datensätze. Besonderes Augenmerk wurde dabei auf Mehrparametertests gelegt, bei denen statistische Hypothesen anhand mehrerer Parameter getestet werden. Beispiel hierfür sind etwa

der Omnibus-Test in der Varianzanalyse (ANOVA), bei dem mehrere Mittelwertsunterschiede simultan gegen Null getestet werden, sowie Modellvergleiche im Allgemeinen (z.B. der Wald- oder der Likelihood-Ratio-Test, LRT). Während relativ klare Empfehlungen darüber existieren, wie einzelne Parameterschätzungen aus multipel imputierten Daten getestet werden können, sind die Empfehlungen für Mehrparametertests weit weniger klar (Enders, 2010; Little & Rubin, 2002; Schafer, 1997; van Buuren, 2012). In der Literatur zur Analyse multipel imputierter Daten werden verschiedene Verfahren für Mehrparametertests diskutiert. Das erste der betrachteten Verfahren ist die D_1-Statistik (Li, Raghunathan, & Rubin, 1991). Diese wird direkt auf Grundlage der Parameterschätzungen und deren Kovarianzmatrizen aus den multipel imputierten Daten berechnet und überträgt somit den klassischen Wald-Test auf multipel imputierte Daten (siehe auch Schafer, 1997). Für Anwendungen, in denen die Kovarianzmatrix des Schätzers nicht zur Verfügung stehen, schlugen außerdem Li, Meng, et al. (1991) die D_2-Statistik vor, die lediglich anhand der χ^2-Statistiken (oder p-Werte) aus den multipel imputierten Daten berechnet wird. Das dritte Verfahren, die D_3-Statistik bietet schließlich die Möglichkeit die Teststatistik des LRTs für den Vergleich zweier Modelle anhand multipel imputierten Daten zu berechnen (Meng & Rubin, 1992).

Obwohl relativ wenig über die Leistung dieser Verfahren bekannt ist (siehe auch Enders, 2010; Schafer, 1997), besagen aktuelle Empfehlungen zur Anwendung von Mehrparametertests, dass D_1 und D_3 gegenüber D_2 unbedingt zu empfehlen seien, da D_2 nur eine geringe Teststärke aufweisen würde und sowohl zu konservativen als übermäßig liberalen Schlussfolgerungen führen könne. Aus diesem Grunde evaluierte Artikel 4 diese Verfahren zunächst im Rahmen der ein- und mehrfaktoriellen ANOVA. Darüber hinaus wurden im Rahmen der Dissertation weitere Simulationsstudien durchgeführt, die diese Verfahren im Rahmen von Mehrebenenmodellen miteinander verglichen, wobei sowohl Tests mehrerer (fester) Regressionskoeffizienten sowie Tests von Varianzkomponenten (d.h. auf zufällige Effekte) betrachtet wurden. Insgesamt wiesen die Simulationsergebnisse auf ein komplexeres Befundmuster hin, das nur teilweise den Empfehlungen in der Literatur entspricht. Im Einklang mit der Literatur wurde in Artikel 4 gezeigt, dass D_1 in der Regel das genaueste Verfahren zur Durchführung von Mehrparametertests darstellt. Allerdings legen die Befunde auch nahe, dass D_2 zu vergleichbaren Ergebnissen wie D_1

und D_3 führt, wenn die Anzahl der Imputationen angemessen hoch und der Datenausfall nicht zu extrem ausfällt. Dies umfasst relativ viele realistische Datenkonstellationen, sodass D_2 in vielen Anwendungen in der psychologischen Forschung als Alternative zu D_1 oder D_3 gesehen werden könnte. Dies ist insbesondere von Bedeutung, da D_2 besonders einfach anzuwenden ist, besonders im Vergleich mit D_3. Im Rahmen von Mehrparametertests in Mehrebenenmodellen konnten diese Ergebnisse im wesentlichen bestätigt werden. Die Ergebnisse für den Test der Varianzkomponenten, für den D_1 nicht verwendet werden kann, legten sogar nahe, dass die Ergebnisse von D_2 hierfür insgesamt verlässlicher sein könnten als die von D_3.

Das R-Paket `mitml`

Obwohl die multiple Imputation zu den am häufigsten empfohlenen Verfahren zum Umgang mit fehlenden Werten gehört, wird MI in der psychologischen Forschung nur relativ selten verwendet, vor allem in Anwendungen mit hierarchischen Daten (Diaz-Ordaz et al., 2014; Jelicic et al., 2009; Nicholson et al., 2017; Peugh & Enders, 2004). In der vorliegenden Dissertation vertrete ich den Standpunkt, dass die geringe Verbreitung von MI unter anderem dadurch bedingt ist, dass (a) die verfügbaren Imputationsverfahren häufig technisch sehr anspruchsvoll und teilweise umfassende Kenntnisse der Programmierung oder Statistik erfordern und (b) die unter MI nötigen Analyseverfahren für komplexere Fragestellungen (z.B. Mehrparametertests) nur spärlich in statistischer Software implementiert sind. Aus diesem Grund wurde im Rahmen der vorliegenden Dissertation das R-Paket `mitml` entwickelt, welches das Ziel verfolgt eine einfachere Anwendung von MI in hierarchischen Daten zu ermöglichen und weiterhin eine Anzahl größtenteils automatischer Werkzeuge für die Analyse multipel imputierter Daten zur Verfügung stellt. In diesem Sinne beinhaltete Artikel 5 der vorliegenden Dissertation eine umfangreiche Einführung in die multiple Imputation hierarchischer Daten.

Zur Imputation fehlender Werte baut das R-Paket `mitml` auf bestehenden Implementationen von JM auf (Quartagno & Carpenter, 2016a; Schafer & Yucel, 2002). Um die Spezifikation der Imputationsmodelle zu vereinfachen, verwendet `mitml` eine Formelsprache, die die Spezifikation des JM in Form einer Modellgleichung ermöglicht (siehe Gleichung 1); alternativ können Imputationsmodelle in Analogie zum R-Paket `mice` (van Buuren & Groothuis-Oudshoorn, 2011)

anhand numerischer Codes definiert werden. Insgesamt ermöglicht `mitml` auf diese Weise mit nur wenigen Zeilen Code die Spezifikation selbst komplexer Imputationsmodelle für kategoriale und kontinuierliche Variablen sowie für fehlende Werte auf Ebene 1 und 2. In diesem Zusammenhang werden auch Imputationsmodelle für heteroskedastische Varianzstrukturen auf Ebene 1 unterstützt (Quartagno & Carpenter, 2016b; Yucel, 2011). Darüber hinaus ermöglicht `mitml` (a) die Diagnose des Konvergenzverhaltens der Imputationsalgorithmen anhand statistischer Kennwerte (z.B. Gelman & Rubin, 1992) und diagnostischer Grafiken, (b) eine einfache Transformation und Analyse der imputierten Daten sowie (c) die Durchführung einfacher und komplexerer Hypothesentests anhand größtenteils automatischer Funktionen. Dies beinhaltet die verschiedenen Verfahren für Mehrparametertests (D_1, D_2, D_3), die in teils allgemein anwendbarer Form (D_1 und D_2) und teils für spezifische Modellklassen implementiert wurden (erforderlich für D_3). Neben einfachen Hypothesentests (Rubin, 1987) ermöglicht dies die Durchführung von Modellvergleichen sowie Tests komplexerer Kontrasthypothesen (anhand der "delta-Methode"; siehe auch Casella & Berger, 2002; Fox, 2008) mit multipel imputieren Datensätzen.

In diesem Zusammenhang nutze Artikel 5 der vorliegenden Dissertation viele Aspekte dieses Pakets für eine umfassende Einführung in die Imputation hierarchischer Daten anhand eines Datensatzes aus der deutschen Teilstudie der "Progress of International Reading Literacy Study" (PIRLS). Neben der Berücksichtigung verschiedener Analysemodelle wurde hierbei besonderes Augenmerk auf die korrekte Spezifikation des Imputationsmodells gelegt. Darüber hinaus besprach der Artikel in umfangreicher Form die Konvergenzdiagnostik im Rahmen von MI sowie die Analyse imputierter Daten anhand realistischer Beispielfragestellungen. Diese Einführung richtete sich vor allem an praktisch arbeitende Wissenschaftler und Wissenschaftlerinnen und verfolgte unter anderem das Ziel eine nützliche und dabei leicht verständliche Einführung in die Behandlung fehlender Werte in hierarchischen Daten zu bieten und somit zur weiteren Verbreitung und Nutzung der multiplen Imputation beizutragen.

Appendix

Article 1: Missing data in multilevel research

Grund, S., Lüdtke, O., & Robitzsch, A. (in press). Missing data in multilevel research. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook for multilevel theory, measurement, and analysis*. Washington, DC: American Psychological Association.

Multilevel research is often faced with missing data. Over the past years, powerful methods such as multiple imputation (MI) and maximum likelihood estimation (ML) have become available for the treatment of incomplete data. In this chapter, we provide a general introduction to the problem of missing data, and we discuss the theory and application of these methods as well as their individual strengths and weaknesses. We offer guidance on how ML and MI may be used for an effective treatment of missing values in multilevel research and what role the multilevel structure may play in the treatment of incomplete data. Finally, we provide results from a computer simulation study as well as an empirical example that illustrates the use of these methods in multilevel analyses.

Multilevel data are often incomplete, for example, when participants refuse to answer some items in a questionnaire or they drop out of a study with several measurement occasions. Even though there is a consensus that current state-of-the-art procedures for statistical analyses with missing data should be preferred (e.g., Allison, 2001; Enders, 2010; Little & Rubin, 2002; Newman, 2014; Schafer & Graham, 2002), simpler methods such as listwise deletion (LD)

prevail and are still widely applied in research practice (Jelicic et al., 2009; Nicholson et al., 2017; Peugh & Enders, 2004). This is problematic because these methods can distort parameter estimates and statistical inference. In this chapter, we provide a general introduction to the problem of missing data in multilevel research, and we present two principled methods for handling incomplete data: multiple imputation (MI) and maximum likelihood estimation (ML). We discuss how these procedures may be used to address missing data in multilevel research, and we consider their commonalities as well as their individual strengths and weaknesses. A brief computer simulation study is used to illustrate the statistical behavior of the parameter estimates obtained from these methods. Finally, we illustrate their application in a data analysis example and provide the syntax files and computer code needed to reproduce our results.

Example: Job satisfaction and leadership style

To provide an illustration of the ideas presented here, we adopt a running example in which we examine the relationship between job satisfaction and several work-related variables. For the purpose of this chapter, we regard the multilevel structure as cross-sectional, for example, with employees at Level 1 nested within work groups at Level 2. The example is based on the data in Klein et al. (2000). It features a sample of 750 employees from 50 work groups with measures of job satisfaction (*SAT*), negative leadership style (*LS*), workload (*WL*), and cohesion (*COH*). We slightly altered the data set by (a) transforming workload into a categorical variable (high vs. low) and (b) treating cohesion as a *global* variable that was directly assessed at level 2 (e.g., a supervisor rating). We investigated the relationship of employees' job satisfaction with negative leadership style, workload, and cohesion using a multilevel random intercept model (Snijders & Bosker, 2012b). In the hierarchical notation of Raudenbush and Bryk (2002), the Level-1 equation of the model reads

$$SAT_{ij} = \beta_{0j} + \beta_{1j}(LS_{ij} - \overline{LS}_{\bullet j}) + \beta_{2j} WL_{ij} + r_{ij} \tag{1}$$

with Level-2 equations

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\overline{LS}_{\bullet j} + \gamma_{02}COH_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20}\ .\end{aligned} \tag{2}$$

Here, SAT_{ij} denotes the job satisfaction of an employee i in group j. The ratings on leadership style were subjected to group-mean centering, where LS_{ij} denotes the employees' individual ratings on leadership style, and $\overline{LS}_{\bullet j}$ denotes the average rating in group j. Finally, WL_{ij} denotes employees' workload, and COH_j denotes a work group's cohesion (e.g., a supervisor rating). The random intercept, u_{0j}, and the residuals, r_{ij}, were each assumed to follow a normal distribution with mean zero and variances τ_0^2 and σ^2. In the remainder of this chapter, we will express this model in a combined notation (e.g., Snijders & Bosker, 2012b),

$$SAT_{ij} = \gamma_{00} + \gamma_{10}(LS_{ij} - \overline{LS}_{\bullet j}) + \gamma_{01}\overline{LS}_{\bullet j} + \gamma_{20}WL_{ij} + \gamma_{02}COH_j + u_{0j} + r_{ij}\ . \tag{3}$$

In this chapter, we focus on multilevel models in which only the intercept varies across groups. Longitudinal research designs as well as multilevel models with additional random effects (e.g., random slopes) are considered in the Discussion section.

Missing data in multilevel research

It is well known that simpler methods of dealing with missing data (e.g., LD) can severely compromise statistical decision making (e.g., Enders, 2010; Little & Rubin, 2002). For example, when analyses are based only on the complete cases, then parameter estimates can be biased (i.e., the estimates may systematically differ from the "true" values that hold in the population) when data are missing in a systematic manner (e.g., see Schafer & Graham, 2002). However, even when data are missing in an unsystematic manner, inferences based on LD are often inefficient (i.e., low statistical power) due to the reduction in sample size and because potentially useful information about the missing data is being ignored (e.g., Newman, 2014). Therefore, the common goal of the "principled" methods for handling missing data is to (a) provide unbiased estimates for the statistical parameters of interest, (b) acknowledge the uncertainty that is due to missing data, and (c) make full use of the data in order to limit the loss of efficiency. However, before we devote ourselves to these methods, it is useful to first establish a formal framework for

discussing the missing data problems and the challenges that may arise in multilevel research. In the following section, we discuss (a) possible mechanisms that may have led to missing data and (b) different patterns of missing data that may occur in multilevel data.

Missing data mechanisms

Rubin (1976) considered three broad classes of missing data mechanisms. We assume that there is a hypothetical complete data set, $\mathbf{Y}$, which can be decomposed into an observed part, $\mathbf{Y}_{obs}$, and an unobserved part, $\mathbf{Y}_{mis}$, where an indicator matrix, $\mathbf{R}$, denotes which elements are observed and which ones are missing. Rubin defined data to be missing at random (MAR) if the probability of observing data, $P(\mathbf{R})$, is independent of the missing data given the observed data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs})$. In other words, under MAR there remains no link between the chance of observing data and the data themselves (i.e., they occur at random) once the observed data are taken into account. A special case of this scenario occurs if the probability of missing data is even completely independent of the data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$, which is referred to as missing *completely* at random (MCAR). By contrast, if the probability of missing data is related to the unobserved data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, it is more difficult to infer from incomplete data, and strong assumptions must be made about the missing data mechanism (see Carpenter & Kenward, 2013; Enders, 2011). This is referred to as missing *not* at random (MNAR).

The meaning of these mechanisms can be subtle, and they are best explained in an example (see also Enders, 2010). Consider the simple scenario illustrated in Figure 1, where negative leadership style is associated with lower job satisfaction, and ratings on leadership style are missing (R_{LS}) as a function of job satisfaction, say, because employees with low job satisfaction were less willing to answer questions about their supervisors (single-headed arrows). In this scenario, larger values of leadership style would be more likely to be missing (double-headed arrow), rendering statements about this variable misleading as long as they do not take the missing data mechanism into account (left panel). For example, the estimated mean of leadership style may be well below the "true" mean because larger values have a higher chance to be missing. However, with job satisfaction taken into account, these ties are broken (right panel): Given the

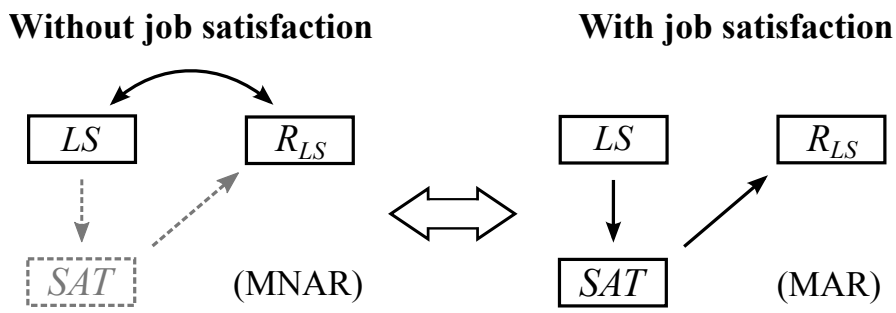

Figure 1: Example for systematic data loss and the effects of ignoring possible causes of missing data. *LS* = leadership style; *SAT* = job satisfaction; R_{LS} = indicator for missing values in leadership style.

values of job satisfaction, the scores of leadership style are now MAR, allowing us to estimate the *conditional* mean of leadership style given job satisfaction (e.g., using linear regression) and to make statements about the overall mean on that basis (see also Carpenter & Kenward, 2013).

The notion of missing data mechanisms allows us to identify conditions under which a missing data treatment may yield more or less accurate results in some model of interest. For example, listwise deletion (LD) generally provides unbiased estimates for a model of interest only under MCAR (see also Newman, 2014). In addition, LD may provide unbiased results in some very specific scenarios in which data are MAR or MNAR (e.g., Galati & Seaton, 2016; Little, 1992). However, since the assertion of specific missing data mechanisms requires making untestable assumptions, LD should be avoided in favor of procedures that make full use of the data and which are applicable under a more general set of assumptions (e.g., ML and MI; see also Schafer & Graham, 2002). Both ML and MI provide unbiased results under MAR. In such a case, the exact mechanism need not be known and may even be different from individual to individual as long as the observed data are sufficient to "break the link" between the unobserved data and the probability that they are missing (Carpenter & Kenward, 2013). To make this assumption more plausible, it is often recommended to include auxiliary variables in the missing data treatment which are not part of the model of interest but which are related to the probability of missing data or the variables with missing data themselves (Collins et al., 2001; Enders, 2008; Graham, 2003). Including such variables is also beneficial if they are related to the variables of interest because they provide information about missing values and may improve statistical power (Collins et al., 2001).

Table 1: Hypothetical Example for a Pattern of Missing Data in a Multilevel Sample

Case	Group	SAT_{ij}	LS_{ij}	WL_{ij}	COH_j	$\overline{LS}_{\bullet j}$
1	1	2.3	?	high	3.8	?
2	1	1.7	?	low	3.8	?
3	1	1.7	?	high	3.8	?
4	2	1.8	2.3	low	?	2.2
5	2	1.4	2.1	high	?	2.2
6	2	?	?	?	?	2.2
7	3	3.4	1.2	low	2.7	1.4
8	3	2.8	1.8	?	2.7	1.4
9	3	3.1	1.2	low	2.7	1.4
10	?	2.1	2.3	high	?	?

Note. Missing observations are indicated by question marks.

Patterns of missing data

For the treatment of missing data, it can also be useful to distinguish different *patterns* of missing data. Such a distinction may help to identify problems with the data and navigate choices regarding the missing data treatment. On the basis of Newman (2014), we distinguish three basic patterns: *item*, *construct*, and *unit* nonresponse. Item nonresponse denotes cases in which participants fail to answer a single item on a questionnaire (e.g., an item concerning payment in a questionnaire for job satisfaction). By contrast, construct and unit nonresponse, respectively, denote cases in which all items pertaining to a certain construct or even the full questionnaire for a participant may be missing (e.g., because a participant was absent on the day the company conducted a survey). In the present chapter, we focus on item nonresponse, though construct missing data can often be addressed using similar methods (see also Gottschall et al., 2012). Unit nonresponse can be more complicated to deal with and is often addressed by employing survey weights (e.g., Särndal, Swensson, & Wretman, 2003).

In multilevel research, item, construct, and unit nonresponse may occur at different levels of the sample (see also van Buuren, 2011). On the basis of Kozlowski and Klein (2000), we may again distinguish three different patterns: missing data at Level 1, in *shared* variables at Level 2, and in *global* variables at Level 2. Missing data at Level 1 refer to the lowest level of the sample (e.g., missing data for employees). Global variables refer to variables that are directly

assessed at Level 2 (e.g., missing data in supervisor rating), whereas shared variables denote variables that are assessed at Level 1 and then aggregated at Level 2 (e.g., a group average based on incomplete data collected from employees). Because missing data in both at Level 1 and in shared variables at Level 2 originate at Level 1, they can usually be addressed by the same methods. Missing data in global variables sometimes require additional considerations but can be treated using similar tools. Additional patterns of missing data are possible, for example, incomplete data about group membership, but these will not be our focus in the present chapter (for a discussion, see Goldstein, 2011; Hill & Goldstein, 1998).

For example, consider Table 1. In the first group of employees, only a single response to the workload variable is missing (Level 1, item missing). In the second group, the ratings on leadership style are missing for all employees (Level l, item missing), and the group mean is missing as a result (shared Level 2, item missing). In the third group, one employee did not respond to any item (Level 1, unit missing). In that group, the group mean might be calculated from the observed values, but it will be subject to uncertainty and possible bias because the underlying items are incomplete (shared Level 2, item missing). In addition, the cohesion score is missing for all employees in that group (global Level 2, item missing). Finally, the last employee could not be assigned to a group with sufficient certainty.

Methods for handling missing data

In the following section, we consider two general procedures that are currently regarded as principled methods for handling missing data (e.g., Schafer & Graham, 2002). First, we consider multiple imputation (MI). We elaborate on different approaches to multilevel MI, and we discuss potential challenges when specifying imputation models for multilevel data. As a second procedure, we consider the estimation of multilevel models by maximum likelihood (ML). Finally, we provide a comparison of the two procedures from a practical point of view.

Multiple imputation (MI)

The basic idea of MI is to replace missing values with an "informed guess" obtained from the observed data and a statistical model (the imputation model). A schematic representation of

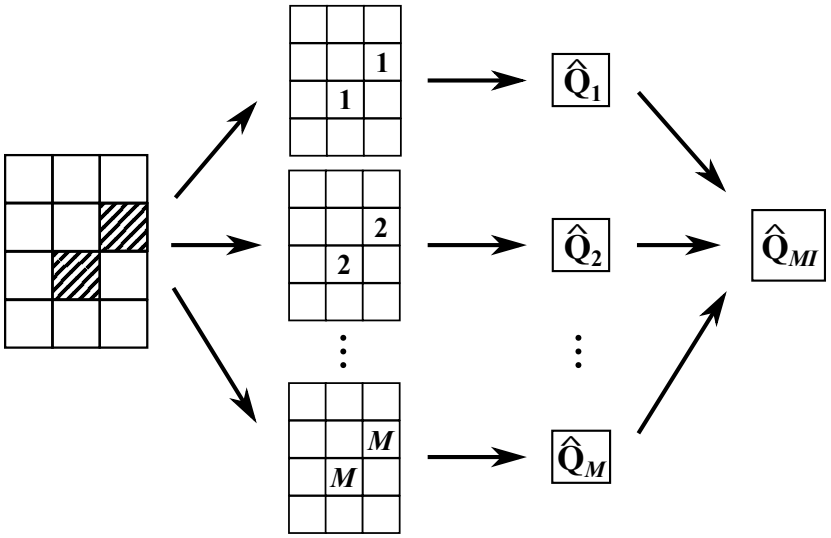

Figure 2: Schematic representation of multiple imputation (MI) and the analysis of multiply imputed data sets. $\hat{Q}$ = estimator of the parameter of interest.

this process is displayed in Figure 2. Multiple imputation generates several (M) replacements for the missing data by drawing from a predictive distribution of the missing data, given the observed data and the parameters of the imputation model. The M data sets are then analyzed separately, yielding M sets of parameter estimates (i.e., $\hat{Q}_1, \ldots, \hat{Q}_M$), and these are combined into a set of final parameter estimates (i.e., $\hat{Q}_{MI}$) and inferences using the rules outlined by Rubin (1987).

When performing MI, the imputation model must be chosen in such a way that it "matches" the model of interest, that is, it must be specified in such a way that it preserves the relationships among variables and the relevant features of the analysis model (Meng, 1994; Schafer, 2003). For example, if the model of interest is a regression model with an interaction effect, then the imputation model must also include the interaction; otherwise, it will be more difficult to detect the interaction effect in subsequent analyses (Enders, Baraldi, & Cham, 2014). In multilevel research, it is important that the imputation model incorporates the multilevel structure of the data. In the following, we review different strategies for accommodating the multilevel structure during MI, including ad hoc strategies on the basis of single-level MI. We consider two broad approaches to MI: joint modeling and the fully conditional specification of MI. In the joint modeling approach, a single statistical model is specified for all incomplete variables simultaneously. In the fully conditional specification, each variable is imputed in turn using a

sequence of models (for a discussion, see Carpenter & Kenward, 2013). Finally, we discuss strategies for analyzing multiply imputed data sets and pooling their results.

Strategies based on single-level MI. Perhaps the simplest approach to multilevel MI is to ignore the multilevel structure of the data and employ single-level MI. Using this strategy, the multilevel structure is disregarded altogether. Not surprisingly, it has been shown that single-level MI may lead to biased estimates in subsequent multilevel analyses (Black et al., 2011; Enders et al., 2016; Taljaard et al., 2008). Lüdtke et al. (2017) demonstrated that single-level MI tends to underestimate the intraclass correlation (ICC; also known as ICC(1)) of variables with missing data and may either under- or overestimate within- and between-group effects in multilevel random intercept models. Figure 3 shows the expected bias in the ICC of a variable Y relative to its true value (i.e., in percent) and for different numbers of individuals per group (n), different values of the ICC of Y and an auxiliary variable X, and different amounts of missing data (25%, 50%). As can be seen, single-level MI tends to underestimate the true ICC. For example, in the scenario with $n = 5$ individuals per group and 25% missing data, single-level MI is expected to yield an estimate of only .062 when the true ICC is .100 and of only .191 when the true ICC is .300. In either case, the true ICC is underestimated by approximately 37%.

To remedy this situation, it has been suggested that the multilevel structure be represented by a number dummy indicator variables (i.e., the DI approach; e.g., Graham, 2009). This strategy effectively estimates a separate group mean for each group by estimating the imputation model conditional on group membership, thus incorporating group differences during MI (see also Enders et al., 2016). For example, the differences in job satisfaction between the 50 work groups in our running example may be represented in a regression model by the intercept and an additional 49 dummy variables (with one group selected as a reference group). The performance of this strategy depends on the situation in which it is applied. As demonstrated by Drechsler (2015), the DI approach tends to overestimate the ICC of variables with missing data but yields approximately unbiased estimates of the regression coefficients in a multilevel analysis model when missing data are restricted to the dependent variable (see also Andridge, 2011). However, because the DI approach exaggerates the variance between groups, it provides only a biased

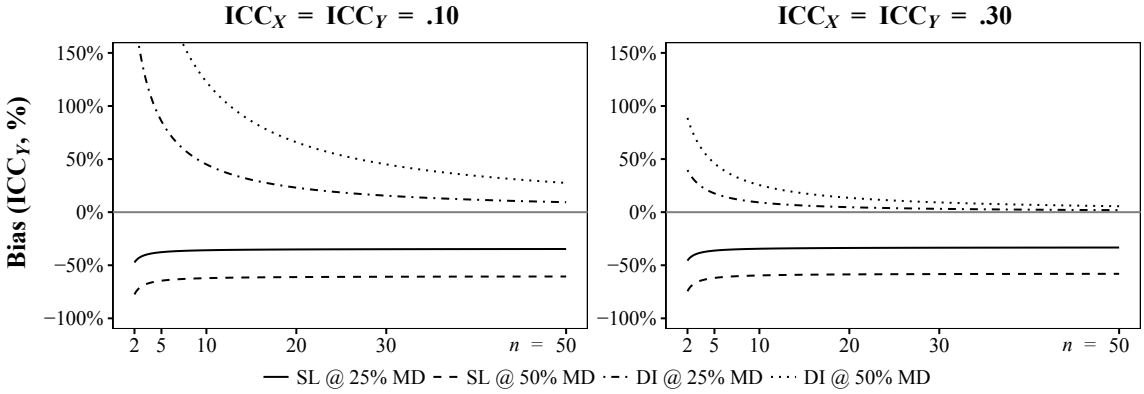

Figure 3: Expected bias for the estimator of the ICC of a variable of interest (Y) under single-level MI (SL) and the dummy-indicator approach (DI). It is assumed that all groups contain the same number of individuals (n) and the same proportion of missing data (MD) in Y. ICC$_Y$ = intraclass correlation of the variable of interest; ICC$_X$ = intraclass correlation of an auxiliary variable.

estimate of the between-group effect if missing values occur in explanatory variables (Lüdtke et al., 2017). As shown in Figure 3, the DI approach tends to overestimate the true ICC. The bias is particularly strong when the true ICC is small and there are only few individuals per group. For example, with $n = 5$ individuals per group and 25% missing data, the DI strategy is expected to yield an estimate of around .186 when the true ICC is .100 and of around .353 when the true ICC is .300. This corresponds to an overestimation of the ICC by approximately 86% and 18%, respectively.

Joint modeling. To accommodate the nested structure of multilevel data, it has been recommended that MI be performed using mixed-effects models (e.g., Enders et al., 2016; Lüdtke et al., 2017; Yucel, 2008). In the joint modeling approach to multilevel MI, a single model is specified for all variables with and without missing data, and imputations are generated from this model for all variables simultaneously.[1] The joint model may be regarded as a multivariate extension of univariate multilevel models, that is, it addresses multiple dependent variables simultaneously. The model reads

$$\begin{aligned} \mathbf{y}_{1ij} &= \boldsymbol{\gamma}_1 + \mathbf{u}_{1j} + \mathbf{r}_{1ij} \quad \text{(Level 1)} \\ \mathbf{y}_{2j} &= \boldsymbol{\gamma}_2 + \mathbf{u}_{2j}\,, \qquad \text{(Level 2)} \end{aligned} \tag{4}$$

[1]The joint model can be expressed in a more general way, which allows including fully observed variables as predictor variables on the right-hand side of the model. However, in the present chapter, we consider only the "empty" specification of the model because it is easy to specify and widely applicable in the context of multilevel random intercept models (for a discussion, see Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016b).

where $\mathbf{y}_{1ij}$ denotes a vector of responses for individual i in group j with fixed intercepts $\boldsymbol{\gamma}_1$, random intercepts $\mathbf{u}_{1j}$, and residuals $\mathbf{r}_{ij}$, and $\mathbf{y}_{2j}$ denotes a vector of responses for group j (i.e., global variables) with fixed intercepts $\boldsymbol{\gamma}_2$ and residuals $\mathbf{u}_{2j}$. The random effects and residuals at Level 2 ($\mathbf{u}_{1j}, \mathbf{u}_{2j}$), are assumed to jointly follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$. The residuals at Level 2 follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. The joint model was originally developed by Schafer and Yucel (2002) to treat missing data at Level 1 and has since been extended to address missing data in categorical variables and variables at Level 2 (Asparouhov & Muthén, 2010b; Carpenter & Kenward, 2013; Goldstein et al., 2009).

To illustrate how the joint model accommodates the multilevel structure, consider our running example and the illustration in Figure 4. The model of interest is a random intercept model which includes variables assessed at Level 1 and 2 as well as relations between job satisfaction and leadership style both within and between groups (Equation 3). The joint model includes all variables as dependent variables in a multivariate random intercept model (Figure 4). For each variable at Level 1, the model includes a random intercept $\mathbf{u}_{1j} = (u_{SAT,j}, u_{LS,j}, u_{WL,j})$, representing the components of these variables that vary between groups, and a residual term $\mathbf{r}_{1ij} = (r_{SAT,ij}, r_{LS,ij}, r_{WL,ij})$, representing the differences within groups. For cohesion, which was assessed directly at Level 2, the model includes a residual term $\mathbf{u}_{2j} = (u_{COH,j})$. The critical point in this model is that it assumes that the random effects and residuals at Level 2 (i.e., global and shared variables) may be correlated ($\boldsymbol{\Psi}$) and that residuals at Level 1 may be correlated as well ($\boldsymbol{\Sigma}$). This illustrates that the joint model indeed "matches" the multilevel structure because it allows differentiating (a) the within- and between-group components that can be present in variables at Level 1 and (b) the relations between variables within and between groups. The joint model or variants thereof are implemented in the packages `pan` (Schafer & Zhao, 2014) and `jomo` (Quartagno & Carpenter, 2016a) for the statistical software R as well as in the standalone software packages SAS (Mistler, 2013b), M*plus* (Asparouhov & Muthén, 2010b), and `REALCOM` (Carpenter, Goldstein, & Kenward, 2011).

Fully conditional specification. As an alternative to the joint model, the joint distribution of the variables with missing data may be approximated by imputing one variable at a time

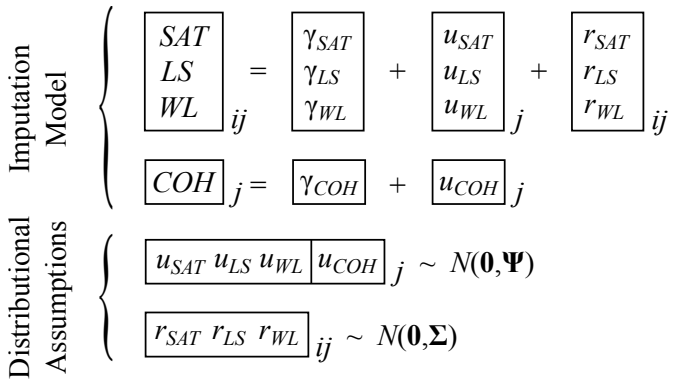

Figure 4: Schematic representation of the joint imputation model and its distributional assumptions in the running example. *SAT* = job satisfaction; *LS* = leadership style; *WL* = workload; *COH* = cohesion.

using a sequence of univariate models. To address multivariate patterns of missing data, the procedure iterates back and forth between variables with missing data, conditioning on the other variables in the data set (or a subset of them). This approach is referred to as the fully conditional specification of MI (FCS; van Buuren et al., 2006). Specifically, for a set of variables at Level 1 and 2, a sequence of conditional imputation models may be specified as follows

$$\begin{aligned} y_{1ijp} &= \mathbf{y}_{ij(-p)}\boldsymbol{\gamma}_p + u_{jp} + r_{ijp} \quad \text{(Level 1)} \\ y_{2jq} &= \mathbf{y}_{j(-q)}\boldsymbol{\gamma}_q + u_{jq}\,, \qquad\quad \text{(Level 2)} \end{aligned} \tag{5}$$

where y_{1ijp} is the p-th variable with missing data at Level 1, and $\mathbf{y}_{ij(-p)}$ is a set of predictors for that variable which may include any variable other than y_{1ijp}. Similarly, y_{2jq} is the q-th variable with missing data at Level 2 (i.e., a global variable), and $\mathbf{y}_{j(-q)}$ is a set of predictor variables which may include any other variable at Level 2 (i.e., global variables) as well as the group means of any variable at Level 1 and (between-group components). The random intercepts u_{jp} as well as the residuals r_{ijp} and u_{jq} in each model are each assumed to follow independent normal distributions (see also van Buuren, 2011). To address multiple variables with missing data, the FCS algorithm arranges them in a sequence and visits one variable at a time, generating imputations from the imputation model assigned to each variable. Once a variable has been completed in this manner, it may be used as a predictor in any of the other imputation models. Once all variables have been visited, the sequence is repeated, and new imputations are generated until the algorithm converges, yielding the first of multiple imputations.

The sequential nature of the FCS algorithm requires some re-thinking. In contrast to the

joint model, the FCS algorithm allows that different predictors may be selected for each target variable, and—conversely—that all target variables may act as predictors in any other target's imputation model. Moreover, in order to preserve the relationships between variables, it is in fact *required* that the imputation model for each target variables conditions on the other variables. To incorporate relationships between variables at Level 2, the group means of variables at Level 1 must be calculated and included as predictors. In addition, the group means must be updated once new imputations for the underlying variables have been obtained; this process of updating the group means is known as *passive* imputation (e.g., Royston, 2005).

To illustrate multilevel FCS, consider our running example and the illustration in Figure 5. Missing data in job satisfaction, leadership style, and workload can be imputed using separate multilevel models, where the latter incorporates a model appropriate for binary categorical data (e.g., a logistic model). Cohesion may be imputed using a regression model at Level 2. In order to preserve the relationship between the variables within and between groups, all variables are included as predictor variables in the other variables' imputation models. In addition, the group means are updated once new imputations have been generated for the underlying variables (i.e., passive imputation). The FCS and similar approaches for multilevel data are implemented in the package `mice` (van Buuren & Groothuis-Oudshoorn, 2011) for the statistical software R as well as in the standalone software packages M*plus* (Asparouhov & Muthén, 2010b) and Blimp (Keller & Enders, 2016).

Incomplete categorical variables. There are several options for treating missing values in categorical and ordinal variables. The first option is to treat categorical variables as continuous for the purpose of MI and to round the resulting values to comply with the original categories in that variable. For example, imputations for ordinal data may be rounded using 0.5, 1.5, etc. as thresholds; for binary data, adaptive rounding may be used, which adjusts this threshold according to the mean of the imputed values (see Carpenter & Kenward, 2013). Adaptive rounding has been shown to perform well for binary missing data (Bernaards, Belin, & Schafer, 2007), but also MI *without* rounding appears to work well for binary and (some) ordinal variables (Schafer, 1997; W. Wu et al., 2015). Finally, it is possible to impute categorical and ordinal variables using a latent variable approach. In this approach, imputations are generated for a set

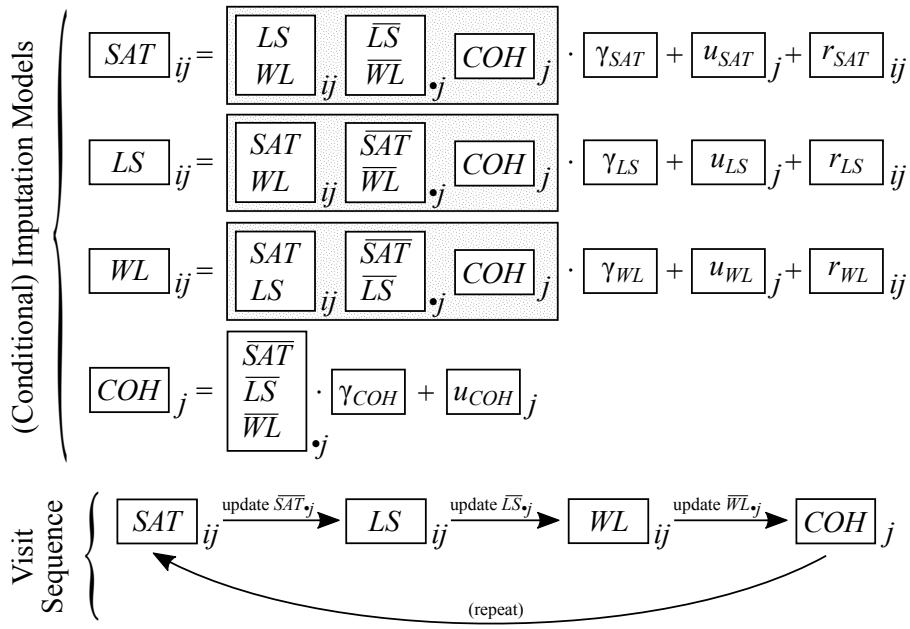

Figure 5: Schematic representation of the sampling steps in the fully conditional specification (FCS) of multilevel MI in the running example. *SAT* = job satisfaction; *LS* = leadership style; *WL* = workload; *COH* = cohesion.

of underlying latent variables that represent the relative probability of being assigned to a given category. Based on the latent scores, the assignment to a category can then be simulated using an appropriate link function (e.g., a probit link for latent normal variables; see Carpenter & Kenward, 2013). For a variable with C categories, this approach assumes $C - 1$ latent variables that represent the possible contrasts between categories (see also Carpenter & Kenward, 2013; Goldstein et al., 2009). For binary variables, this is equivalent to generating imputations from a generalized linear mixed-effects model (e.g., a logistic or probit model). These procedures, too, appear to work well for both binary and polytomous data (Demirtas, 2009; W. Wu et al., 2015; see also Enders et al., 2016).

Analyzing multiply imputed data. The idea underlying MI is to generate plausible replacements for each missing value, thus transforming a data set with "missing data" to a data set with "complete data." This process is repeated M times (hence the qualifier "multiple"), yielding M completed versions of the original data (see Figure 2). Once the set of M data sets has been obtained, the model of interest must be fit separately to each data set, yielding M estimates of some parameter of interest, say $\hat{Q}_m$ (e.g., regression coefficients; $m = 1, \ldots, M$), and M estimates of the sampling variance of that estimate, $\hat{V}_m$ (e.g., squared standard errors). According

to Rubin (1987), the combined point estimate is the average of the individual estimates,

$$\hat{Q}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m \, . \tag{6}$$

The combined estimate of the sampling variance of the estimator incorporates two different sources of uncertainty,

$$\hat{V}_{MI} = \widehat{W} + \left(1 + \frac{1}{M}\right) \hat{B} \, , \tag{7}$$

where $\widehat{W}$ denotes the sampling variance *within* imputations, that is, the average of the individual variance estimates,

$$\widehat{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_m \, , \tag{8}$$

and B denotes the sampling variance *between* imputations, that is, the variance of the point estimates across data sets,

$$\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{Q}_m - \hat{Q}_{MI})^2 \, . \tag{9}$$

Using the combined point and variance estimates, $\hat{Q}_{MI}$ and $\hat{V}_{MI}$, standard hypothesis tests can be carried out on the basis of a Student's t distribution with ν degrees of freedom. Rubin (1987) recommended calculating the degrees of freedom as follows

$$\nu = (M-1) \left[1 + \frac{1}{\text{RIV}}\right]^2 \, , \tag{10}$$

where the expression

$$\text{RIV} = \frac{\widehat{W}}{(1 + \frac{1}{M})\hat{B}} \tag{11}$$

denotes the relative increase in the sampling variance of the estimator that is due to missing data (see also Barnard & Rubin, 1999). In addition, several alternative formulas have been proposed for more complex hypotheses that may involve several parameters simultaneously, for example, when testing the overall effect of categorical explanatory variables or when testing for random slopes using a likelihood-ratio test (see Appendix A; see also Reiter & Raghunathan, 2007).

The general idea of Rubin's rules is to approximate the sampling distribution of $\hat{Q}$ that would be obtained with infinite M but based on only a small number of imputations. Naturally, the larger this number is chosen the better the approximation becomes, which raises the question of "how many are needed?" Traditionally, $M = 5$ imputations have been recommended (Rubin,

1987), but more can be necessary when the amount of missing data increases or the model of interest becomes more complex (Bodner, 2008; Graham, Olchowski, & Gilreath, 2007). This is especially important because most software packages for multilevel MI generate $M = 5$ imputations by default. In our experience, $M = 20$ imputations are usually sufficient for estimating and testing the parameters in most applications of multilevel models. However, when large portions of the data are missing (say above 50%) or complex hypotheses are tested that involve multiple parameters, we recommend generating 50-100 imputed data sets (see also Bodner, 2008; Raghunathan, 2015).

Maximum likelihood (ML)

The general principle of maximum likelihood estimation (ML) is to choose the values of the parameters in a statistical model in such a way that the likelihood of the data becomes maximal. When the data contain missing values, it is often possible to estimate the model directly using only the observed data. This procedure is often referred to as *direct* or *full information* ML. Using ML, the likelihood is evaluated on a case-by-case basis, that is, cases with incomplete records contribute to the likelihood only to the extent to which they have data (Little & Rubin, 2002). The ML estimates of the parameters in a model of interest are consistent when the data are MAR or MCAR, that is, missing data occur in an unsystematic fashion when the variables in the model are taken into account (Little & Rubin, 2002).

The main principle by which ML "deals" with missing data is that it imposes distributional assumptions on incomplete variables. For this reason, common multilevel software packages often handle missing values only in the dependent variable of the model (e.g., HLM, SAS), where such assumptions are already in place, but cases with missing values in explanatory variables are discarded because no distributional assumptions have been made for them. To circumvent this restriction, it has been suggested to adopt the framework of *structural equation modeling* (SEM), which allows introducing distributional assumptions for all variables by defining them as endogenous (i.e., dependent) variables in a single analysis model (e.g., Allison, 2012; Enders, 2010). For example, in the statistical software M*plus*, this is achieved by including the variances and covariances of the explanatory variables in the modeling statement. Using this strategy

it is often possible to prevent the software from discarding these cases and to apply the ML principle to both the dependent and explanatory variables in a model of interest. Furthermore, this perspective allows for including auxiliary variables that may improve the plausibility of the MAR assumption and the accuracy of estimates under ML (Enders, 2008; Graham, 2003). Software that supports ML for multilevel models from the perspective of SEM includes the standalone software packages M*plus* (L. K. Muthén & Muthén, 2012), Latent GOLD (Vermunt & Magidson, 2013), `gllamm` (Rabe-Hesketh, Skrondal, & Pickles, 2004), and xxM (Mehta, 2013).

As an alternative to direct ML, estimates for the parameters in a multilevel model may be obtained from a two-stage procedure by first estimating a covariance matrix within and between groups based on the observed data; in the second stage, the parameters of interest are derived from the variances and covariances estimated in the first stage (Yuan & Bentler, 2000). Conceptually, the two-stage ML is similar to the perspective taken in SEM. We do not consider this approach further, but using two-stage ML may offer advantages when working with nonnormal variables and because auxiliary variables are easily incorporated in the estimation procedure (Savalei & Bentler, 2009; Yuan, Yang-Wallentin, & Bentler, 2012).

Comparison of ML and MI

From a theoretical point of view, ML and MI are not vastly different, and both can be expected to yield similar results when they operate under similar assumptions (Schafer & Graham, 2002). However, from a practical point of view, the differences may be substantial. Fitting models using ML is often easy, provided that a software package can be found that supports estimating the model of interest. Furthermore, because ML does not separate the treatment of missing data from the analysis, the missing data model is always consistent with the analysis model, that is, both models are always based on the same set of assumptions Allison, 2012. However, integrating the treatment of missing data and the estimation of the analysis model into a single step also has disadvantages. First, the distributional assumptions needed for the treatment of missing data now also enter the analysis model even though they may not have been an original part of it. Second, auxiliary variables must be incorporated directly into the model of interest,

thus making the analysis model more complex (Graham, 2003). In applications with few, well-behaved variables, this is usually no problem; but in practice, it can become problematic, for example, when the inclusion of auxiliary variables leads to a mix of continuous and categorical variables at both Level 1 and 2. Such models are difficult for the user to specify, and a given software package may not even fully support it, forcing the user to alter the model or make decisions he or she would not have made otherwise.

Conducting MI, on the other hand, is more complicated at first glance. First, an imputation model must be chosen that is consistent with the model of interest. Then, the user must specify the number of imputations and for how many iterations the sampling procedure should run. Finally, he or she must ensure that the algorithm has converged before any analyses can be carried out (see also Allison, 2012). Once the imputations have been generated, the user must fit the analysis model to each of the imputed data sets and combine their results into a final set of parameter estimates and inferences. Especially for unexperienced users, performing MI can be a daunting task. On the other hand, modern procedures for multilevel MI are powerful and very flexible in accommodating a variety of models. In addition, many software packages for multilevel MI automatize at least some of these steps. Finally, the separation between the treatment of missing data and the analysis phase makes it straightforward to handle a variety of variables and to include auxiliary variables without altering the model of interest.

Simulation

Next, we report the results from a computer simulation study. This study was intended to illustrate the general performance of ML and MI in a controlled setting. We conducted this study with two models of interest in mind. The first model of interest (Model 1) was the model from our running example

$$SAT_{ij} = \gamma_{00} + \gamma_{10}(LS_{ij} - \overline{LS}_{\bullet j}) + \gamma_{01}\overline{LS}_{\bullet j} + \gamma_{20}WL_{ij} + \gamma_{02}COH_j + u_{0j} + r_{ij} \; . \quad \text{(3 revisited)}$$

This represents the standard formulation of multilevel models, in which the observed group means represent the shared perception of leadership style among members of the same group. The second model of interest (Model 2) is also known as the “multilevel latent covariate model”

(Lüdtke et al., 2008) and differs from the first model in that it uses the true, unobserved group means or between-group components to represent the shared perception of individuals in each group. The model reads

$$SAT_{ij} = \gamma_{00} + \gamma_{10} LS_{W,ij} + \gamma_{01} LS_{B,j} + \gamma_{20} WL_{ij} + \gamma_{02} COH_j + u_{0j} + r_{ij} \,, \tag{12}$$

where $LS_{W,ij}$ and $LS_{B,j}$ denote the within- and between-group components of leadership style (Asparouhov & Muthén, 2006; Lüdtke et al., 2008). Formulating the model in terms of the true within- and between-group components can be beneficial because it corrects for the fact that the group mean is calculated from a finite number of observations and thus provides only an unreliable measure of the true between-group component (see Croon & van Veldhoven, 2007; Raudenbush & Bryk, 2002). In the organizational literature, the reliability of the group mean is also known as the ICC(2), and it expresses the extent to which differences between the observed group means reflect true differences between groups (see also Bliese, 2000; LeBreton & Senter, 2008). It is a matter of debate in the multilevel literature which formulation of the model of interest is more appropriate. For example, it may be argued that the formulation in Model 2 is appropriate if the shared perception among individuals is of primary interest (e.g., ratings of team climate, leadership effectiveness), whereas Model 1 may be appropriate if the variation within groups is itself of interest or if the observed group mean is simply regarded as a summary measure (e.g., gender ratio, socioeconomic status; for further discussion, see Lüdtke et al., 2008). However, the main motivation for including these two approaches to modeling between-group effects in the present chapter was that their distinction is important for the treatment of missing data under ML (see below).

In the simulation study, the samples were generated from either Model 1 (the "standard" model) or Model 2 (the "latent covariate" model) in order to allow a comparison between conditions in which one of the two models is the "true" model. The parameters of the simulation were loosely based on the data from Klein et al. (2000). The samples consisted of $G = 50$ groups of size $n = 10$. All variables were standardized across groups with mean zero and unit total variance. For the ratings on leadership style and job satisfaction, we assumed an ICC of .10 and .20, respectively. In addition, we assumed that negative leadership style was

Table 2: Mean Estimates (and Coverage Rates for the 95% Confidence Interval) for the Two Models of Interest for MI and ML

		Model 1					Model 2		
	True	JM	FCS	ML1	ML2	True	JM	FCS	ML1
γ_{00}	0.000	0.003 (95.0)	0.002 (95.0)	0.004 (96.1)	0.011 (94.1)	0.000	0.001 (94.8)	0.000 (94.3)	0.001 (95.1)
γ_{10}	−0.200	−0.202 (94.7)	−0.200 (94.7)	−0.203 (94.9)	−0.200 (94.7)	−0.200	−0.201 (93.8)	−0.200 (94.0)	−0.200 (94.3)
γ_{01}	−0.700	−0.648 (94.6)	−0.660 (94.5)	−1.215 (91.8)	−0.633 (90.5)	−0.700	−0.708 (95.4)	−0.714 (95.1)	−0.803 (96.1)
γ_{20}	−0.300	−0.301 (94.8)	−0.300 (95.0)	−0.302 (94.9)	−0.303 (94.8)	−0.300	−0.298 (94.9)	−0.297 (94.9)	−0.299 (94.9)
γ_{02}	0.100	0.102 (94.3)	0.102 (93.8)	0.067 (95.5)	0.105 (92.7)	0.100	0.098 (95.0)	0.099 (94.4)	0.095 (95.4)
τ_0^2	0.083	0.085 (95.4)	0.082 (94.0)	0.035 (91.9)	0.081 (91.5)	0.088	0.079 (95.0)	0.078 (93.9)	0.069 (92.3)
σ^2	0.751	0.747 (94.1)	0.747 (94.1)	0.746 (94.1)	0.749 (94.3)	0.790	0.786 (94.6)	0.786 (94.5)	0.786 (94.8)

Note. JM = joint modeling of MI; FCS = fully conditional specification of MI; ML1 = maximum likelihood with true within- and between-group components for leadership style; ML2 = maximum likelihood with group means for leadership style calculated from the observed data; γ_{00} = intercept; γ_{10} = within-group effect of leadership style; γ_{01} = between-group effect of leadership style; γ_{20} = effect of workload; γ_{02} = effect of cohesion; τ_0^2 = intercept variance; σ^2 = residual variance.

correlated with cohesion at the group level ($r = -.15$). For the two workload categories (high vs. low), we generated a standard normal variable with an ICC of .20, and we dichotomized that variable using 0.38 as a breaking point, resulting in 35% and 65% of individuals with high and low workload, respectively. For simplicity, we assumed that workload was uncorrelated with the other explanatory variables. Finally, we assumed the following fixed effects in the data generating model: $\gamma_{00} = 0$ (intercept), $\gamma_{10} = -.20$ and $\gamma_{01} = -.70$ (leadership style), $\gamma_{20} = -.30$ (workload), and $\gamma_{02} = .10$ (cohesion). The variance components τ_0^2 and σ^2 then followed. We induced missing values in cohesion completely at random (5%), and in leadership style (15%) and workload (10%) based on job satisfaction (lower job satisfaction corresponded to higher chance of missing data). Finally, we induced missing values in job satisfaction completely at random (10%).

Using this procedure, we generated 5,000 data sets from both Model 1 and 2. In each data set, we carried out MI using both joint modeling (using `jomo`; Quartagno & Carpenter, 2016a)

and FCS (using `mice`; van Buuren & Groothuis-Oudshoorn, 2011) in the statistical software R. Afterwards, we fitted the respective model of interest using M*plus* 7 (L. K. Muthén & Muthén, 2012). To estimate the model using ML, we also used M*plus*, and we addressed missing data in explanatory variables by specifying distributional assumptions for these variables. In the context of Model 2, applying ML is relatively easy because M*plus* already imposes the necessary distributional assumptions when decomposing leadership style into its within- and between-group components. The distributional assumptions for the remaining variables can be added by defining them as endogenous variables at Level 1 and 2, respectively.[2] On the other hand, in the context of Model 1, missing data in explanatory variables pose a greater challenge when estimating the model using ML. We consider two strategies for this case, neither of which are completely satisfying. In the first strategy (ML1), distributional assumptions are specified as before by defining explanatory variables as endogenous variables at Level 1 or 2, respectively. However, this strategy unintentionally adopts the within- and between-group decomposition for leadership style (as in Model 2), thus correcting between-group effects that did not require correction. As a second option (ML2), the group means of leadership style may be calculated beforehand from the observed data, and distributional assumptions may be imposed only on the within-group deviations of leadership style. In this specification, the group means are consistent with the analysis model, but the between-group effects of leadership style may be biased if values are missing in a systematic manner (similar to LD).

In Table 2, we included the mean estimates of the three procedures for the two models of interests as well as the coverage of the 95% confidence interval. Ideally, the mean estimates should be close to the true values in the data-generating model, and the coverage rates should be close to 95%. In the context of Model 2, both MI and ML yielded parameter estimates that were very close to the true values, and coverage rates were close to the nominal value of 95%. However, the between-group effect of leadership style (γ_{01}) was slightly too large under ML, which may be attributed to the small sample size at Level 2 (Lüdtke et al., 2008).

[2]Using ML, it was also not straightforward to accommodate both (a) the multilevel structure of the variables and (b) the fact that workload is categorical. Therefore, we treated workload as a continuous variable. While this may be acceptable for a dichotomous variable with similar frequencies in both categories, it will lead to problems when explanatory variables have multiple categories or some categories occur much more frequently than others.

In the context of Model 1, the parameter estimates obtained from MI were again close to the true values, but the between-group effect of leadership style (γ_{01}) was slightly underestimated. Under ML, specifying leadership style as an endogenous variable (ML1), thus adopting the within- and between group decomposition, led to severe bias in the between-group regression coefficients. By contrast, when the group means were calculated beforehand from the observed data (ML2), thus treating only the within-group deviations as endogenous, the group-level effect of leadership style (γ_{01}) was only slightly underestimated. The coverage rates were relatively close to the nominal value of 95% for most parameters but tended to be slightly smaller under ML, especially when the group means were calculated from the observed data (ML2).

In conclusion, both ML and MI provided accurate results when their assumptions were met and when these assumptions were consistent with the model of interest. These requirements were more easily fulfilled in the context of Model 2, in which case both MI and ML yielded reasonable parameter estimates. However, in the context of Model 1, the results were more diverse. Under ML, following the usual advice to treat explanatory variables as endogenous may lead to an unwanted "shift" in the analysis model, which severely distorted parameter estimates. When the group means were calculated beforehand, we observed only little bias. However, this approach slightly overestimated the precision of the parameter estimates because it ignored the fact that group means were calculated from incomplete records. Under MI, estimates were accurate and confidence intervals showed good coverage properties, providing the most reasonable approximation to the true parameters overall.

Example application

In the following section, we apply the missing data methods to our running example. The running example is based on the data in Klein et al. (2000) and essentially mimics the conditions in our simulation study except that the example data set contained *un*standardized variables instead. Missing values were induced in the data set in the same way as in the simulation study. As a result, 21.9% of the employees had missing values on at least one variable; these were distributed across job satisfaction (9.2%), leadership style (12.3%), workload (11.5%), and

Table 3: Estimates for the Parameters in the Model of Interest Obtained from ML and MI in the Running Example

	M*plus* (ML2)		`jomo` (MI)			
Parameter	Est.	SE	Est.	SE	RIV	FMI
Intercept (γ_{00})	0.291^{*}	0.136	$0.257^{\dagger}$	0.140	0.167	0.143
Level 1						
Leadership style (γ_{10})	-0.526^{***}	0.091	-0.532^{***}	0.092	0.341	0.255
Workload (γ_{20})	-0.863^{***}	0.197	-0.842^{***}	0.195	0.259	0.206
Level 2						
Leadership style (γ_{01})	-1.491^{***}	0.319	-1.566^{***}	0.349	0.237	0.192
Cohesion (γ_{02})	0.237^{**}	0.088	0.243^{**}	0.091	0.075	0.070
Level 2 residual variance (τ_0^2)	0.268^{*}	0.128	0.286	—	—	—
Level 1 residual variance (σ^2)	4.940^{***}	0.283	4.962	—	—	—

Note. ML2 = maximum likelihood with group means for leadership style calculated from the observed data; SE = standard error; RIV = relative increase in variance; FMI = fraction of missing information.
$^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (two-tailed)

cohesion (4.0%). The data set is included in the R package `mitml` (Grund, Robitzsch, & Lüdtke, 2016). The model of interest was the "standard" multilevel model in Equation 3 (Model 1). We applied MI using the joint model implemented in the `jomo` package in R, and we estimated the model of interest using the `lme4` package (Bates et al., 2016). To assist with the analyses, we used the `mitml` package, which provides a wrapper function for the `jomo` package as well as tools for analyzing multiply imputed data sets (see also Grund, Lüdtke, & Robitzsch, 2016b). For ML estimation, we used M*plus*, where we calculated the group means of leadership style from the observed records (as in ML1) and adopted the within- and between-group decomposition for the remaining variables (as in ML2). The computer code and the M*plus* syntax file are provided in Appendix B.

To set up the imputation model using `jomo` and `mitml`, two formulas had to be specified which denoted the imputation model for variables at Level 1 and 2, respectively (see Equation 4 and Figure 4). In accordance with the "empty" specification of the model, all variables are treated as target variables, and no predictor variables are specified except a "one" for the intercept. We generated $M = 100$ imputations in this manner. The number of iterations for the algorithm was chosen in such a way that convergence could be established by inspecting convergence criteria (e.g., Gelman & Rubin, 1992) and diagnostic plots for the parameters of

the imputation model (see also Grund, Lüdtke, & Robitzsch, 2016b; Schafer & Olsen, 1998). After running MI, the model of interest was fitted to each of the imputed data sets using `lme4`, and the parameter estimates were pooled using Rubin's rules in order to obtain a final set of parameter estimates and inferences. The results obtained from ML and MI are presented in Table 3. The two analyses suggested that negative leadership style had a relatively strong impact on employees' job satisfaction when controlling for employees' workload and the work group's cohesion. Under MI, for any one-unit change in the leadership style ratings within groups (Level 1), the expected change in job satisfaction was $-.532$ ($p < .001$). Between groups, a one-unit change in the shared perception of leadership style ratings (Level 2) was associated with an expected change in job satisfaction of -1.566 ($p < .001$). Furthermore, there was a negative effect of high (vs. low) workload (-0.842, $p < .001$) on job satisfaction and a positive effect of cohesion (0.243, $p = 0.007$). The results obtained from ML were virtually identical. Perhaps the largest difference between the two procedures was the standard error for the between-group effect of leadership style, which might reflect the slightly too narrow confidence intervals under ML observed in the simulation study.

In addition, we also investigated whether the within-group effect of leadership style *varies* across groups, that is, whether there is significant variance in the slope of leadership style. To this end, we fitted an alternative model that contains a random slope for within-group effect of leadership style. The alternative model was compared with the model of interest using the D_3 statistic (Meng & Rubin, 1992), which can be interpreted as a pooled LRT for multiply imputed data sets (Appendix A). The D_3 statistic suggested that there is not enough evidence to conclude that the effect of leadership style truly varies across groups, $F(2, 3707.9) = 2.621$ ($p = .071$). Therefore, the alternative model was rejected in favor of the model of interest.[3] Furthermore, we were interested in whether the effect of leadership style was larger between than within groups. For this purpose, we used the D_1 statistic which allowed us to test the difference between the two coefficients against zero using a linear constraint (Appendix A; see

[3]Note that, because the imputation model did not include random slopes, it did not "match" the alternative model. For that reason, the hypothesis test is not completely trustworthy and is included here only for the purpose of illustration.

also Kreft et al., 1995). The D_1 statistic suggested that the two parameters were significantly different from one another, $F(1, 2471.3) = 8.253$ ($p = .004$), that is, the between-group effect (-1.57) was significantly larger than the within-group effect (-0.53).

Discussion

In this chapter we provided an introduction to multilevel modeling with missing data. In particular, we looked at two principled methods for handling missing data: multiple imputation (MI) and estimation by maximum likelihood (ML). The general idea of ML and MI is not vastly different, and both procedures may be regarded as state-of-the-art procedures for handling missing data (Schafer & Graham, 2002). Differences between the two methods are most often of a practical nature. Although both procedures tend to give the same answers if they are based on similar assumptions, carrying out a given task is often easier with one procedure as compared with the other. For example, ML is very easy to incorporate in one's regular workflow because the missing data treatment is performed during the estimation of the model of interest (see also Allison, 2012). On the other hand, addressing missing values and including auxiliary variables may prove to be challenging depending on where the missing data occur and how complex the model becomes once all factors are taken into account, for example, if categorical variables contain missing data or between group effects are represented by observed group means. By contrast, MI allows for very flexible modeling of different types of variables, and including auxiliary variables is straightforward. On the other hand, performing MI and analyzing multiple data sets can be challenging, especially for less experienced users or if nonstandard analyses and hypothesis tests are required. That being said, although we clearly see ML as the easier-to-use alternative (see Allison, 2012; Enders, 2010), we tend to favor MI due to its flexibility and because it separates the imputation from the analysis phase (see Carpenter & Kenward, 2013; Schafer & Graham, 2002; see also Grund, Lüdtke, & Robitzsch, in press-b).

As in every introduction to these or similar procedures, it is not possible to consider all possible research scenarios with the attention they deserve. In this chapter, we restricted our discussion to cross-sectional multilevel models with a single level of clustering, that is, indi-

viduals nested within some higher-level collective. In principle, the procedures discussed here generalize naturally to models with further levels of clustering, for example, three-level models (Goldstein, 2011; Keller, 2015; Yucel, 2008), models with cross-classified random effects (Goldstein, 2011; Hill & Goldstein, 1998), or models with multiple memberships (Goldstein, 2011; Yucel et al., 2008). However, these procedures are not widely available in standard software, and more research is needed to evaluate their performance in realistic research scenarios.

Another topic that we did not discuss explicitly is the treatment of missing data in longitudinal research designs (e.g., repeated measurements, diary studies, experience sampling, ecological momentary assessment). This topic is particularly interesting, however, because multilevel models are frequently used for analyzing longitudinal data. Fortunately, many of the ideas presented here can also be applied to longitudinal data (see also Black et al., 2013; Newman, 2003). For example, assume that a researcher is interested in estimating a growth curve model with missing data in the dependent variable that should be treated using MI. It is then useful to distinguish studies in which the longitudinal design is balanced or unbalanced with respect to time, that is, whether all participants were measured at the same or a different set of time points (see W. Wu, West, & Taylor, 2009). If all participants were measured on the same set of time points, then the longitudinal data structure can be expressed in a wide data format, and single-level MI may be used for treating the missing values in the dependent variable (for a two-stage ML procedure, see Yuan et al., 2012). However, if participants were measured at potentially different or unbalanced time points, then procedures based on mixed-effects models for multilevel MI may be more appropriate (see Equation 4). However, even though the model by Schafer and Yucel (2002) was developed explicitly with applications to longitudinal data in mind, the model lacks flexibility to incorporate some covariance structures at Level 1 that are commonly used in longitudinal analysis models (see Pinheiro & Bates, 2000). Similar problems may be observed when estimating growth curve models using ML because it is difficult to establish a homogeneous covariance structure for this type of data (W. Wu et al., 2009).

Even though there has been a substantial interest in missing data methods for multilevel data in recent years, some questions still provide challenges for the future. One such example is the

treatment of missing data in multilevel models with random slopes or in models with nonlinear and interaction effects. For example, it has been shown that current implementations of MI are not perfectly suited for handling missing data in explanatory variables in multilevel models with random slopes (e.g., Enders et al., 2016; Gottfredson et al., 2017; Grund, Lüdtke, & Robitzsch, 2016a; see also von Hippel, 2009). Similar problems may occur under ML but have yet to be discussed more thoroughly in the applied missing data literature (however, see Enders et al., 2014). In order to make sure that imputations are consistent with the model of interest, it has been argued that the substantive analysis model should be taken into account during MI (Bartlett et al., 2015; Carpenter & Kenward, 2013). Several authors have proposed procedures that incorporate these ideas using rejection sampling or a Metropolis-Hastings algorithm for multilevel MI, but these procedures are not yet available in standard software (Erler et al., 2016; Goldstein et al., 2014; L. Wu, 2010). Similar procedures have been proposed in the context of ML, where the likelihood function in a multilevel model can be factored into separate components referring to the model of interest and additional models for explanatory variables with missing data (Ibrahim et al., 2001; Stubbendick & Ibrahim, 2003).

To sum up, missing data are an ever-present problem in research practice. We believe that both ML and MI provide powerful tools for the treatment of missing data in multilevel research. The two procedures both come with their own strengths and weaknesses, and one may be preferred over the other for a specific missing data problem. At the end of the day, however, they are more similar than they are different, and both offer a substantial improvement over approaches such as LD in terms of generality, theoretical foundation, accuracy of parameter estimates, and statistical power. In the present chapter, we provided an introduction to these methods, and we offered guidance on how to apply them in multilevel research. The treatment of missing data is not without its challenges, and there remain many open (and interesting) questions for the future. However, we believe that these methods are a valuable addition to the researcher's toolbox which, if applied correctly, can improve the quality of the conclusions we draw from our data and that of our research altogether. We hope that this chapter will promote the adoption of MI and ML and encourage researchers to use these procedures in their own research projects.

Appendix A: Multiparameter hypothesis tests in MI

In research practice, statistical hypotheses often involve multiple parameters simultaneously (e.g., linear constraints, comparisons of nested models). In complete-data analyses, these are often performed using the Wald test or likelihood-ratio tests (LRT). For pooling a series of Wald tests based on a series of parameter vectors, $\hat{\mathbf{Q}}_m$, and covariance matrices, $\hat{\mathbf{V}}_m$, Li, Raghunathan, and Rubin (1991) proposed using the test statistic

$$D_1 = \frac{(\hat{\mathbf{Q}}_{MI} - \mathbf{Q}_0)^T \hat{\mathbf{W}}^{-1} (\hat{\mathbf{Q}}_{MI} - \mathbf{Q}_0)}{K(1 + \text{ARIV}_1)}, \tag{13}$$

where $\hat{\mathbf{Q}}_{MI}$ and $\hat{\mathbf{W}}$ are the average estimates of the parameter vector and its covariance matrix (see Equations 6 and 8), $\mathbf{Q}_0$ contains the hypothesized values of the parameters under the null hypothesis, and ARIV_1 is an estimate of the average relative increase in variance (ARIV) due to nonresponse across parameters (see Enders, 2010). The D_1 statistic can be used in a similar manner as Rubin's rules (1987), that is, it can be used for testing a set of parameters (or a linear transformation thereof) that have an approximately normal sampling distribution (e.g., regression coefficients).

It is sometimes difficult to calculate D_1, for example, because estimates of the covariance matrix are unavailable. As an alternative, Li, Meng, et al. (1991) proposed pooling a set of Wald-like test statistics, D_m, as follows

$$D_2 = \frac{\bar{D}K^{-1} + (M + 1)(M - 1)^{-1}\text{ARIV}_2}{1 + \text{ARIV}_2}, \tag{14}$$

where $\bar{D}$ is the average of the D_m, and ARIV_2 is an alternative estimate of the ARIV. The D_2 statistic can be used for any quantity that follows a χ^2-distribution, for example, a Wald test of a set of regression coefficients (or a linear transformation thereof) or an LRT comparing two nested models (see also Snijders & Bosker, 2012b).

As a third option, Meng and Rubin (1992) have proposed a test statistic for pooling a series of LRTs as follows

$$D_3 = \frac{\tilde{L}}{K(1 + \text{ARIV}_3)}, \tag{15}$$

where the ARIV_3 is another estimate of the average relative increase in variance, which includes (a) the average LRT statistic evaluated at the *actual* parameter estimates, and (b) the average

LRT statistic evaluated at the *average* parameter estimates for the two models ($\tilde{L}$). This test statistic can be used in the same manner as the LRT, for example, for comparing two nested statistical models (see above).

In general, D_1 and D_3 tend to be the more reliable procedures and should be used when possible. However, because software implementations of D_1 and D_3 are sometimes not available, D_2 may be an interesting alternative given its ease of application. Even though D_2 was optimized to work with a small number of imputations ($M = 3$), results from D_2 tend to be much more robust when more imputations (say, $M \geq 20$) are used (Grund, Lüdtke, & Robitzsch, 2016c; Licht, 2010). Care should be taken when large portions of the data are missing (say, more than 50%) because D_2 and (to a lesser extent) D_3 tend to be less robust in these cases.

Appendix B: Computer code for the example application

Printed below is the computer code used for multilevel MI in the data analysis example.

```
# *** Description of the 'leadership' data set:
#
# GRPID:   indicator for work groups
# JOBSAT:  job satisfaction (Leve 1)
# NEGLEAD: ratings on negative leadership style (Level 1)
# WLOAD:   workload (Level 1, "low" vs. "high")
# COHES:   group cohesion (Level 2)

# Multiple imputation is performed with an "empty" joint model using jomo. The
# model of interest is fit using lme4, and the mitml package is used for pooling
# tests and parameters.

library(lme4)
library(mitml)

# set up random number generator
set.seed(1234)

# load data
data(leadership)

# *** Imputation phase:
#

# set up "empty" model
fml <- list( NEGLEAD + JOBSAT + WLOAD ~ 1 + (1|GRPID) ,   # Level 1 model
             COHES ~ 1 )                                  # Level 2 model

# impute
imp <- jomoImpute(leadership, formula=fml, n.burn=5000, n.iter=500, m=100)
```

```
# assess convergence
summary(imp)              # convergence criteria ("Rhat")
plot(imp)                 # diagnostic plots

# create list of completed data sets
implist <- mitmlComplete(imp, print="all")

# *** Analysis phase:
#

# apply group mean centering
implist <- within(implist,{
  G.NEGLEAD <- clusterMeans(NEGLEAD,GRPID)
  I.NEGLEAD <- NEGLEAD - G.NEGLEAD
})

# fit model of interest and pool parameter estimates
fit <- with(implist, lmer(JOBSAT ~ I.NEGLEAD + G.NEGLEAD + WLOAD + COHES + (1|GRPID)))
testEstimates(fit, var.comp=TRUE)

# test for random slope of leadership style (using D3)
fit2 <- with(implist, lmer(JOBSAT ~ I.NEGLEAD + G.NEGLEAD + WLOAD + COHES +
                      (1+I.NEGLEAD|GRPID)))
anova(fit, fit2)

# test for contextual effect of leadership style (using D1)
context <- "G.NEGLEAD - I.NEGLEAD"
testConstraints(fit, constraint=context)
```

Printed below is the M*plus* syntax that was used for ML estimation of the model of interest.

```
DATA:
file = leadership.dat;

VARIABLE:
names = GRPID JOBSAT COHES NEGLEAD WLOAD;
usevariables = JOBSAT COHES NEGLEAD WLOAD NEGLEADM;
within = NEGLEAD;
between = COHES NEGLEADM;
cluster = GRPID;
missing = all (-99);

DEFINE:
NEGLEADM = cluster_mean (NEGLEAD);    ! calculate group means from the observed data
center NEGLEAD (groupmean);           ! group mean centering

ANALYSIS:
type = twolevel;
estimator = ml;

MODEL:
%within%
JOBSAT on NEGLEAD
  WLOAD (1);                ! restrict effect of workload to be equal at both levels
NEGLEAD with WLOAD;         ! explanatory variables as endogenous, allow covariances

%between%
JOBSAT on NEGLEADM COHES
```

```
  WLOAD (1);                ! restrict effect of workload to be equal at both levels
NEGLEADM with COHES;        ! explanatory variables as endogenous, allow covariances
NEGLEADM with WLOAD;
COHES with WLOAD;
```

Article 2: Multiple imputation of missing data for multilevel models: Simulations and recommendations

Grund, S., Lüdtke, O., & Robitzsch, A. (in press). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*. doi:10.1177/1094428117703686

Multiple imputation (MI) is one of the principled methods for dealing with missing data. In addition, multilevel models have become a standard tool for analyzing the nested data structures that result when lower level units (e.g., employees) are nested within higher level collectives (e.g., work groups). When applying MI to multilevel data, it is important that the imputation model takes the multilevel structure into account. In the present paper, based on theoretical arguments and computer simulations, we provide guidance using MI in the context of several classes of multilevel models, including models with random intercepts, random slopes, cross-level interactions (CLIs), and missing data in categorical and group-level variables. Our findings suggest that, oftentimes, several approaches to MI provide an effective treatment of missing data in multilevel research. Yet we also note that the current implementations of MI still have room for improvement when handling missing data in explanatory variables in models with random slopes and CLIs. We identify areas for future research and provide recommendations for research practice along with a number of step-by-step examples for the statistical software R.

Multilevel models have become one of the standard tools for analyzing clustered empirical data. Such data are often found in organizational psychology, for example, when employees are nested within work groups or enterprises, or in longitudinal studies when measurement occasions are nested within persons. In addition, empirical data are often incomplete, for example, when some participants fail to 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 deletion (LD; Allison, 2001; Enders, 2010; Little & Rubin, 2002; Newman, 2014; Schafer & Graham, 2002). One central requirement of MI is that the imputation model must be at least as general as the model of interest in order to preserve its key features. In multilevel data, it is important that the imputation model takes the multilevel structure into account (e.g., Andridge, 2011; Drechsler, 2015). However, depending on the research question, the multilevel structure may manifest itself in the analysis model in a number of ways (e.g., random intercepts and slopes, relations between variables within and between groups), leading to a multitude of possible multilevel analysis models, each directed at different research questions (e.g., Aguinis & Culpepper, 2015; Snijders & Bosker, 2012b).

The motivation behind the present paper is twofold. First, we offer simulation results regarding the performance of MI when the substantive analysis model belongs to one of several types of multilevel models. Second, we provide an introduction to and recommendations for MI of multilevel data directed toward readers who are not yet familiar with the often technical literature on MI. Our article is divided into four sections. In the first, we focus on the multilevel random intercept model and discuss imputation procedures that are suitable for application in such models. In the second section, we focus on the random coefficients model and the specific challenges that arise when working with random slopes and cross-level interactions (CLIs). In the third section, we briefly discuss missing data in categorical and group-level variables. In each section, we present results from simulation studies in which we used different MI procedures as well as full-information maximum likelihood (FIML). Finally, in the last section, we provide recommendations for how to handle missing data for different types of multilevel models. We conclude with a discussion of our findings and possible topics for future research.

Missing data and multiple imputation

The basic idea of MI is to replace missing values by forming an "informed guess" that is based on the observed data and a statistical model (the imputation model). Multiple imputation generates several (M) replacements for the missing data by drawing repeatedly from the posterior predictive distribution of the missing data, given the observed data and the parameters of the imputation model. The M data sets completed in this manner are then analyzed separately, yielding M sets of parameter estimates. To obtain final estimates and inferences, these results are pooled using the rules described in Rubin (1987; see also Enders, 2010).

The use of MI in most (but not all) implementations is predicated on the assumption that the data are "missing at random" (MAR). The definition of MAR, according to Rubin (1976), assumes that a hypothetical complete data set can be divided into observed and unobserved parts, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where an indicator matrix $\mathbf{R}$ denotes which data are missing or observed. According to Rubin, data are MAR if the probability of observing data, $P(\mathbf{R})$, is independent of the unobserved data given the observed data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{obs})$. In other words, under

MAR, there is no link between the chance of observing a value and the value itself, given the data that one has observed. A special case of this occurs when $P(\mathbf{R})$ is completely independent of the data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$. This is referred to as missing *completely* at random (MCAR). If the MAR assumption is violated, that is, data are missing *not* at random (MNAR), the application of MI requires strong assumptions about the missing data mechanism. Such applications are relatively rare and are most often used as sensitivity analyses (see Carpenter & Kenward, 2013). In the present paper, we focus on applications of MI that operate under MAR.

Two aspects make MI a particularly attractive method for dealing with missing data. First, MI recognizes the uncertainty that is due to missing data by generating multiple (as opposed to single) replacements for each missing value, and by drawing the parameters of the imputation model from Bayesian posterior distributions, given the currently imputed data and a set of prior beliefs. Second, because the imputation phase is separated from the analysis phase, MI is able to make full use of the data by including variables in the imputation model that are either predictive of missingness, thus improving the plausibility that MAR holds, or related to the variables of interest, thus improving the power of its predictions (Collins et al., 2001).

Multiple imputation for multilevel models

A crucial point in the application of MI to multilevel data is that the imputation model not only includes all relevant variables, but also that it "matches" the model of interest (i.e., the substantive analysis model; see Meng, 1994; Schafer, 2003). In other words, the imputation model must capture the relevant aspects of the analysis model, making the imputation model at least as general as (or more general than) the analysis model. If the imputation model is more restrictive than the analysis model, then imputations are generated under a simplified set of assumptions, and the results of subsequent analyses may be misleading. For example, consider the case in which the model of interest is a multilevel random intercept model (Snijders & Bosker, 2012b) in which an individual-level outcome Y is regressed on an individual-level explanatory variable X

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + u_{0j} + e_{ij}\,, \tag{1}$$

where $\bar{X}_{\bullet j}$ denotes the group mean of X in group j, $(X_{ij} - \bar{X}_{\bullet j})$ denotes the individual deviation in X for a person i in group j, and γ_{10} and γ_{01} denote the regression coefficients of X within and between groups, respectively (see Hofmann & Gavin, 1998; Kreft et al., 1995). The intercepts, u_{0j}, and the residuals, e_{ij}, are assumed to follow independent normal distributions with mean zero and with variances τ_0^2 and σ^2, respectively.

Two aspects of the model in Equation 1 are worth noting, and both must be accommodated during MI in order for subsequent analyses to yield proper results. First, the model accounts for the clustered structure of the data by including random effects for each group (Snijders & Bosker, 2012b). Therefore, the imputation model must also take the clustered structure into account. Failing to do so, for example, by using single-level MI, might lead to biased parameter estimates and might distort statistical decision making (e.g., Andridge, 2011; Enders et al., 2016; Lüdtke et al., 2017; Taljaard et al., 2008). Second, the model differentiates between the effects of X at the individual and the group level (i.e., for $(X_{ij} - \bar{X}_{\bullet j})$ and $\bar{X}_{\bullet j}$). If the imputation model does not allow these effects to be different, then the parameters will be "conflated" during MI, and estimates obtained in subsequent analyses may be biased (see Enders et al., 2016; Lüdtke et al., 2017; Preacher et al., 2010). In other words, ignoring the existence of separate effects for $(X_{ij} - \bar{X}_{\bullet j})$ and $\bar{X}_{\bullet j}$ in the imputation model will make it more difficult to find them in subsequent analyses. In the following section, we discuss several MI procedures that can be used to accommodate the multilevel random intercept model.

Joint modeling and the fully conditional specification of MI

The procedures available for multilevel MI can be roughly divided into two broad paradigms: the joint modeling approach (JM) and the fully conditional specification of MI (FCS). Both approaches offer the necessary tools for dealing with multilevel missing data. Here, we consider the JM approach implemented in the `pan` package (Schafer & Yucel, 2002) and the FCS approach known as "multiple imputation by chained equations" implemented in the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011) in the statistical software R (R Core Team, 2016).

Joint modeling (JM). In the JM approach, a single model is specified for all variables with missing data, and imputations are simultaneously generated from this model for all variables

with missing data. For individual-level variables, the joint model reads

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j + \mathbf{e}_{ij} \,, \tag{2}$$

where $\mathbf{Y}_{ij}$ contains a number of individual-level target variables with arbitrary patterns of missing data, $\mathbf{X}_{ij}$ contains fully observed predictor variables with associated fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}_{ij}$ contains fully observed predictor variables with associated random effects $\mathbf{u}_j$, and $\mathbf{e}_{ij}$ denotes the residuals at the individual level. The random effects, $\mathbf{u}_j$, are assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$. The residuals, $\mathbf{e}_{ij}$, follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

The design matrices, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$, on the right-hand side of the model equation may contain any number of variables as long as they are fully observed (see Schafer & Yucel, 2002). If the model of interest is a multilevel random intercept model, it is possible to include all variables (both partially and fully observed) as target variables on the left-hand side of the model equation, whereas the right-hand side includes only the intercept (i.e., $\mathbf{X}_{ij} = \mathbf{Z}_{ij} = 1$). For example, consider the random intercept model in Equation 1 and assume that X and/or Y are partially missing. Treating both X and Y as target variables, the JM becomes

$$\left[X_{ij}, Y_{ij}\right]^T = \left[\beta_{0(x)}, \beta_{0(y)}\right]^T + \left[u_{j(x)}, u_{j(y)}\right]^T + \left[e_{ij(x)}, e_{ij(y)}\right]^T \,, \tag{3}$$

where the random effects $\left[u_{j(x)}, u_{j(y)}\right]^T$ and the residuals $\left[e_{ij(x)}, e_{ij(y)}\right]^T$ follow independent multivariate normal distributions with mean zero and covariance matrices $\boldsymbol{\Psi} = \begin{bmatrix} \psi_x^2 & \psi_{xy} \\ \psi_{xy} & \psi_y^2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$. In this specification, the joint model decomposes the variables into separate within- and between-group components represented by $\left[e_{ij(x)}, e_{ij(y)}\right]^T$ and $\left[u_{j(x)}, u_{j(y)}\right]^T$, thus allowing for different relations (i.e., covariances) between X and Y to be estimated at the individual and the group level (Lüdtke et al., 2017; see also Grund, Lüdtke, & Robitzsch, 2016b).[1] Similar models are also implemented in the statistical software M*plus* (L. K. Muthén & Muthén, 2012; see also Enders et al., 2016) and in the R package `jomo` (Quartagno & Carpenter, 2016a).

[1] It is possible to include fully observed variables on the right-hand side of the JM (i.e., in $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$). This strategy is conceptually close to the FCS approach described in the next section. Hence, for the JM, we restricted our attention to the multivariate model in which all variables are treated as target variables (see also Enders et al., 2016).

The FCS approach. In contrast to the JM approach, the FCS approach imputes missing data separately for each variable with missing data, conditioning on some or all of the other variables in the data set. To address multivariate patterns of missing data, the FCS algorithm iterates back and forth between different target variables. Again, consider the analysis model in Equation 1. For missing data in X and Y, an appropriate FCS approach may generate imputations on the basis of the following two univariate models

$$\begin{aligned} X_{ij} &= \beta_{0(x)} + \beta_{1(x)}(Y_{ij} - \bar{Y}_{\bullet j}) + \beta_{2(x)}\bar{Y}_{\bullet j} + u_{j(x)} + e_{ij(x)} \\ Y_{ij} &= \beta_{0(y)} + \beta_{1(y)}(X_{ij} - \bar{X}_{\bullet j}) + \beta_{2(y)}\bar{X}_{\bullet j} + u_{j(y)} + e_{ij(y)} \; . \end{aligned} \tag{4}$$

The FCS approach iterates between these equations, generating imputations for each missing variable in turn. If both variables are affected by missing data, then the group means are updated at each iteration of the sampling algorithm on the basis of the most recent imputations for X and Y (passive imputation; see below). Similar to JM, unsystematic differences between groups in X and Y are captured by the inclusion of random effects, $u_{j(x)}$ and $u_{j(y)}$. In contrast to JM, however, the FCS approach uses the observed group means, $\bar{Y}_{\bullet j}$ and $\bar{X}_{\bullet j}$, to represent the different relations between X and Y at the individual and the group level.[2] For missing Y, there is no difference between the imputation and the analysis model. For missing X, the imputation model has similar implications as in the JM approach, but it relies not only on random effects but also on the observed group means to represent the relation between X and Y at the group level. In applications with more than two variables, the general approach remains the same: For each additional variable with missing data, an additional equation must be specified, each conditioning on the other variables and their respective group means.

Summary. Summing up, there are two points worth noting. First, both the JM and the FCS approach allow for different relations between variables to be estimated at the individual and the group level. Second, the two approaches differ in the way in which they accomplish this task. In the JM, the group level is represented by random effects, whereas the FCS approach relies on the observed group means. However, even though the general approach is different, it has been

[2]It is possible to ignore the group means entirely in the FCS approach, for example, by imputing missing data in X by assuming only an overall effect of Y. However, this strategy "conflates" the individual- and group-level effects and may introduce bias into the parameter estimates. Here, we consider only the more general model that includes the group means in the set of predictor variables (see also Enders et al., 2016).

argued that the two approaches imply similar covariance structures at the individual and the group level and can be used interchangeably (e.g., Carpenter & Kenward, 2013, p. 220; Lüdtke et al., 2017; Mistler, 2015; however, see also Resche-Rigon & White, in press). Therefore, we expected the two procedures to yield approximately the same, unbiased parameter estimates, making both suitable for MI in quite general applications of the multilevel random intercept model.

Model-based treatment using FIML

As an alternative to MI, it is often possible to use model-based procedures such as FIML to treat missing data (for an introduction, see Enders, 2010). FIML is often considered to be very user-friendly because missing data are handled directly during the estimation of the analysis model without requiring any additional steps to be taken by the user (e.g., Allison, 2012; Graham, 2009). Currently, the most popular and versatile implementation of FIML for multilevel models is available in the statistical software M*plus* (L. K. Muthén & Muthén, 2012). FIML estimates the parameters of the analysis model directly from the incomplete data set by maximizing the observed-data likelihood. As a result, the use of FIML to treat missing data is closely tied to the analysis model (Schafer & Graham, 2002). In the traditional multilevel model (e.g., Equation 1), the observed-data likelihood includes only the dependent variable in the analysis (e.g., Y), and distributional assumptions are imposed only on that variable. For that reason, FIML initially deals with missing data only in the dependent variable, whereas cases with missing data in explanatory variables are often discarded (see also Hox et al., 2016). To treat missing data in explanatory variables (e.g., X), the model must be extended in such a way that the likelihood function will incorporate all variables with missing data, thus imposing additional distributional assumptions on the data. In M*plus*, this is typically achieved by specifying a set of latent variables for the explanatory variables with missing data (for an illustration, see Enders, 2010).

Although it may not be immediately obvious, this strategy can have negative side-effects in multilevel modeling because of the way in which M*plus* estimates multilevel models with latent variables. For example, consider the model in Equation 1 with missing values in X and Y. To

estimate this model, M*plus* uses a decomposition approach similar to the JM, in which the two variables are decomposed into (latent) individual- and group-level components, each of which is assumed to follow a multivariate normal distribution (Rabe-Hesketh, Skrondal, & Zheng, 2012). However, in doing so, M*plus* adopts a different analysis model in which the group-level effects of X on Y are represented by latent variables instead of observed group means (i.e., $X_{\bullet j}$) as they are in the analysis model (for a discussion, see Lüdtke et al., 2008). As a result, parameter estimates may change substantially, both in meaning and in value (Grund, Lüdtke, & Robitzsch, in press-a). To avoid this shift in the analysis model, the user may calculate the group means beforehand from the observed data and specify a latent variable only for the within-group component of the explanatory variables (i.e., $X_{ij} - \bar{X}_{\bullet j}$). This strategy tends to reduce bias in group-level effects and will be preferred for the remainder of this article (see also Grund et al., in press-a). In addition, because the individual- and group-level components are assumed to follow a multivariate normal distribution, only linear relations are allowed between variables with missing data, and handling missing data in categorical variables may be challenging. The M*plus* syntax files needed to perform FIML estimation for the models presented here are given in the supplemental online materials.

Study 1: Random intercept models

Next, we present findings from a computer simulation study in which we compared the performance of different MI procedures in the context of multilevel random intercept models. In addition to the JM and the FCS approach, we also investigated single-level MI, which ignores the multilevel structure altogether, LD, and FIML as discussed above. The main question was which procedures would preserve the relevant features of the substantive analysis model. Here, we provide only a brief sketch of the study's design. For interested readers, we provide further details in Appendix A.

The substantive analysis model was the random intercept model in Equation 1, and the data were generated from this model. The parameters of the data-generating model were chosen in such a way that they would imply a given value for the intraclass correlations (ICCs) of X and Y. Missing data were generated on Y or X in either a random fashion (MCAR) or conditional

Table 1: Simulation Conditions for the Data-Generating Model and the Generation of Missing Values

	Study 1	Study 2	Study 3a	Study 3b
Data conditions				
No. of individuals	5, 10	5, 10	5, 10	5, 10
No. of groups	50, 100, 200, 500	50, 100, 200, 500	50, 100, 200, 500	50, 100, 200, 500
ICC of X and Y	.10, .20, .50	.10, .20, .50	.10, .20, .50	.10, .20, .50
ICC of D			.10, .20, .50	
Correlation XW		.20		.20
Correlation XD			.20	
Model parameters				
Effect of $(X_{ij} - \bar{X}_{\bullet j})$	.20	.50	.50	0
Effect of $\bar{X}_{\bullet j}$	.20, .50	0	.50	.20
Effect of W_j		.35		.20
Effect of D_{ij}			.20	
CLI $(X_{ij} - \bar{X}_{\bullet j})W_j$		.0, .20		
GLI $\bar{X}_{\bullet j}W_j$		0		
Total slope variance		.10		
Int.-slope covariance		0		
Missing values				
Pattern/mechanism	$Y \sim X$, $X \sim Y$	$Y \sim X$, $X \sim Y$	$D \sim Y$	$W \sim Y$
Effect	MCAR, MAR	MCAR, MAR	MCAR, MAR	MCAR, MAR
Proportion	25%	25%	25%	25%
No. of conditions	192	192	48	48

Note. The residual intercept and slope variance were determined by the remaining simulation parameters and by setting a target value for the ICC of Y and the total slope variance ($\gamma_{11}^2 + \tau_1^2$). CLI = cross-level interaction; GLI = group-level interaction.

on the other variable (MAR). A summary of the simulation conditions is provided in Table 1. We varied the number of groups (k = 50, 100, 200, 500), the number of individuals within each group (n = 5, 10), the ICCs of X and Y ($\rho_{I,X} = \rho_{I,Y}$ = .10, .20, .50), the effect of $\bar{X}_{\bullet j}$ (γ_{01} = .20, .50), and the missing data mechanism. The effect of $(X_{ij} - \bar{X}_{\bullet j})$ was held constant at .20 (γ_{10}), thus providing conditions in which the effects at the individual and the group level were equal or different in the population model. Taken together, these conditions mimic typical applications of multilevel models in cross-sectional and longitudinal organizational research (e.g., smaller and larger ICCs, smaller and larger numbers of observations per unit or group). In addition, they provide information about the small- and large-sample properties of each procedure and about conditions that are interesting from a methodological point of view (e.g.,

with or without contextual effects). Each condition was replicated 1,000 times. We applied the following procedures to each data set:

1. LD
2. single-level FCS, ignoring the multilevel structure (FCS-SL)
3. multilevel FCS with separate within- and between-group effects (Equation 4; FCS-ML)
4. multilevel JM (Equation 2; JM)
5. FIML

The parameters of interest were the ICC of Y, estimated from an empty model, and the regression coefficients within and between groups, γ_{10} and γ_{01}, from the substantive analysis model. For each condition, each procedure, and each parameter, we calculated the bias, the RMSE, and the coverage of the 95% confidence interval to evaluate performance. The bias is defined as the difference between an estimator's average value and its true value. The RMSE is the square root of the average squared difference between average estimates and true values, combining information about bias and efficiency of parameter estimates. The coverage of the 95% confidence interval denotes the relative frequency with which the 95% confidence interval covers the true value. The properties of an estimator may be considered suboptimal if the bias exceeds 10%, the RMSE is large in comparison with other procedures, or the coverage rate is below 90% (or very close to 100%).

Results. Our findings are summarized in Table 2 and Figure 1. The complete collection of results for the parameters of interest is provided in the supplemental online materials. Consistent with our expectations, single-level MI (FCS-SL) reduced the ICC of Y when Y was partially missing. In such a case, the between-group regression coefficient was biased downwards, and the within-group regression coefficient was biased upwards to different extents as determined by the true magnitude of the ICCs of X and Y. With missing values in X, FCS-SL either over- or understated the true size of the regression coefficients, depending on the ICCs. As shown in Figure 1, this bias did not decrease in larger samples. The results from the two appropriate MI procedures (FCS-ML and JM) were similar to one another. Both procedures had a slight tendency to overestimate the ICC of Y and to underestimate the between-group coefficient in

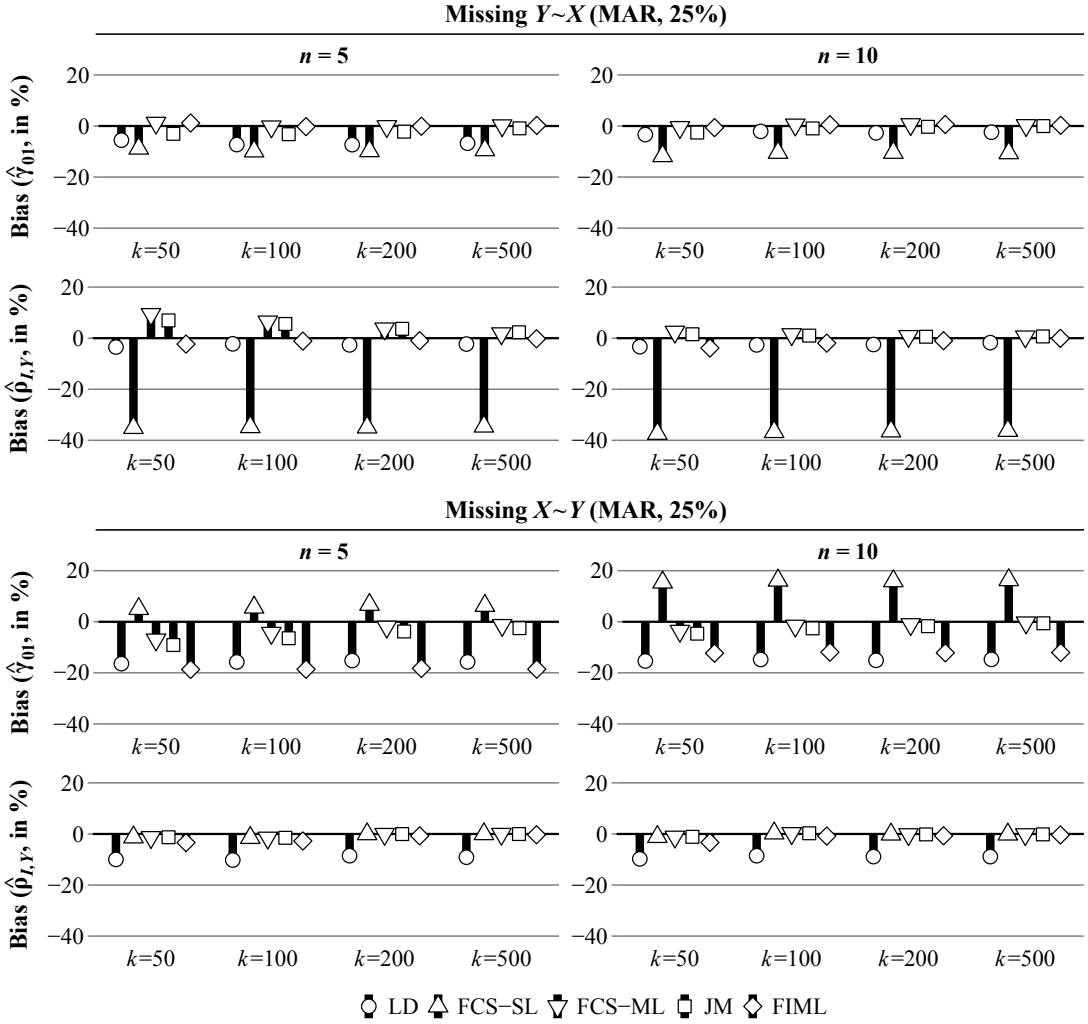

Figure 1: Estimated bias for the between-group regression coefficient of X (γ_{01}) and the ICC of Y ($\rho_{I,Y}$) in Study 1 for different numbers of individuals (n) and groups (k), moderate ICCs ($\rho_{I,X} = \rho_{I,Y} = .20$), and different missing data mechanisms (MAR; $Y \sim X$ and $X \sim Y$). LD = listwise deletion; FCS-SL = single-level FCS; FCS-ML = multilevel FCS; JM = multilevel JM; FIML = full-information maximum likelihood.

smaller samples (i.e., $k = 100$ or lower, with $n = 5$) with low ICCs ($\rho_{I,X} = \rho_{I,Y} = .10$). However, this bias was seldom substantial and decreased as the sample size increased (see Figure 1). FIML produced unbiased estimates of the regression coefficients with missing Y but biased estimates of the between-group regression coefficient (γ_{01}) with missing X. Finally, LD led to substantially biased estimates of all parameters of interest when data were MAR, especially when values were missing in X. In conditions in which the within- and between-group coefficients were equal ($\gamma_{01} = .20$), the results were essentially the same, and both JM and FCS-ML provided approximately unbiased estimates of the parameters of interest.

The coverage of the 95% confidence interval was acceptable in all conditions for FCS-ML

Table 2: Bias (in %), RMSE, and Coverage of the 95% Confidence Interval for the ICC of Y and the Within- and Between-Group Regression Coefficients in Study 1 (Small Groups, $n = 5$)

	LD			FCS-SL			FCS-ML			JM			FIML		
	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.
Missing $Y \sim X$ (MAR, 25%)															
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .10$														
$\hat{\gamma}_{10}$	−1.4	0.06	94.2	9.6	0.06	94.0	−1.3	0.06	93.7	1.6	0.06	95.1	−1.4	0.06	94.5
$\hat{\gamma}_{01}$	−9.1	0.11	92.6	−10.0	0.10	95.0	1.1	0.10	95.7	−4.5	0.10	97.1	1.1	0.10	95.4
$\hat{\rho}_{I,Y}$	−2.0	0.05	—	−30.7	0.04	—	28.6	0.05	—	24.7	0.05	—	12.6	0.04	97.5
$k = 500$															
$\hat{\gamma}_{10}$	0.0	0.03	95.8	10.4	0.03	90.0	−0.1	0.03	96.1	2.6	0.03	95.6	−0.1	0.03	96.0
$\hat{\gamma}_{01}$	−10.0	0.07	77.4	−10.9	0.07	76.3	−0.1	0.05	95.4	−3.7	0.05	94.4	−0.0	0.05	95.3
$\hat{\rho}_{I,Y}$	−4.3	0.02	—	−32.1	0.04	—	13.2	0.02	—	13.7	0.02	—	7.1	0.02	97.9
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .50$														
$\hat{\gamma}_{10}$	1.3	0.06	94.5	24.0	0.08	91.1	1.7	0.06	94.2	1.2	0.06	93.8	1.5	0.06	94.2
$\hat{\gamma}_{01}$	−2.7	0.09	94.6	−6.4	0.09	91.9	−0.5	0.09	95.1	−0.8	0.09	94.9	−0.5	0.09	94.8
$\hat{\rho}_{I,Y}$	−1.5	0.05	—	−35.0	0.18	—	−0.8	0.05	—	−0.8	0.05	—	−1.3	0.05	95.8
$k = 500$															
$\hat{\gamma}_{10}$	−0.2	0.03	95.5	22.3	0.05	72.4	−0.2	0.03	95.3	−0.3	0.03	96.0	−0.2	0.03	96.0
$\hat{\gamma}_{01}$	−2.6	0.04	92.9	−6.2	0.05	83.0	−0.4	0.04	93.5	−0.3	0.04	93.5	−0.3	0.04	93.7
$\hat{\rho}_{I,Y}$	−1.3	0.03	—	−34.7	0.17	—	−0.5	0.02	—	−0.4	0.02	—	−0.5	0.02	94.2
Missing $X \sim Y$ (MAR, 25%)															
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .10$														
$\hat{\gamma}_{10}$	−6.9	0.06	94.3	5.9	0.06	96.1	1.5	0.06	94.6	4.2	0.06	94.6	1.2	0.06	95.1
$\hat{\gamma}_{01}$	−17.3	0.12	81.7	−1.7	0.09	98.5	−8.1	0.10	95.7	−11.5	0.10	94.1	−23.2	0.14	68.3
$\hat{\rho}_{I,Y}$	−11.9	0.05	—	−0.8	0.04	—	−0.8	0.04	—	−0.8	0.04	—	4.6	0.04	97.3
$k = 500$															
$\hat{\gamma}_{10}$	−6.8	0.03	91.6	6.6	0.03	92.7	1.1	0.03	94.1	3.7	0.03	93.9	1.6	0.03	94.2
$\hat{\gamma}_{01}$	−18.0	0.10	36.0	−1.8	0.04	97.7	−4.1	0.04	93.9	−7.5	0.05	89.7	−24.1	0.13	8.4
$\hat{\rho}_{I,Y}$	−11.8	0.03	—	0.5	0.02	—	0.5	0.02	—	0.5	0.02	—	3.8	0.02	96.8
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .50$														
$\hat{\gamma}_{10}$	−4.7	0.06	94.3	−9.6	0.05	96.7	−1.2	0.06	94.8	−1.1	0.06	94.5	0.1	0.06	94.1
$\hat{\gamma}_{01}$	−10.5	0.10	90.5	20.9	0.14	83.2	−0.9	0.09	95.6	−1.0	0.09	96.0	−9.2	0.09	90.9
$\hat{\rho}_{I,Y}$	−4.8	0.06	—	−0.9	0.05	—	−0.9	0.05	—	−0.9	0.05	—	−1.5	0.05	93.8
$k = 500$															
$\hat{\gamma}_{10}$	−4.0	0.03	93.9	−9.0	0.03	91.8	−0.1	0.03	95.3	−0.2	0.03	95.2	0.5	0.03	94.8
$\hat{\gamma}_{01}$	−9.8	0.06	74.8	21.4	0.12	36.3	−0.4	0.04	94.4	−0.4	0.04	95.2	−8.7	0.06	77.5
$\hat{\rho}_{I,Y}$	−4.1	0.03	—	−0.2	0.02	—	−0.2	0.02	—	−0.2	0.02	—	−0.3	0.02	94.6

Note. $\hat{\gamma}_{10}$ = within-group regression coefficient; $\hat{\gamma}_{01}$ = between-group regression coefficient; $\hat{\rho}_{I,Y}$ = ICC of Y (estimated from an empty model); LD = listwise deletion; FCS-SL = single-level FCS; FCS-ML = multilevel FCS; JM = multilevel JM; FIML = full-information maximum likelihood.

and in all but the most extreme conditions under JM. However, owing to persistent bias, the coverage rates under FCS-SL frequently dropped below 90% in larger samples (i.e., above $k = 200$, $n = 10$ or $k = 500$, $n = 5$). Coverage rates for FIML were acceptable with missing Y, but dropped below 90% with missing X; those for LD were acceptable under MCAR but

unacceptable under MAR. Finally, the RMSE tended to be lowest under FCS-ML and JM as well as under FIML with missing Y. By contrast, the RMSE for the parameters of interest tended to be larger under FCS-SL as well as under LD and FIML with missing X, indicating that these procedures were altogether less accurate and efficient. For example, the average RMSE for the between-group regression coefficient was 3.3% larger under LD with missing Y as compared with JM; with missing X, this difference increased to 9.7%. Note also that the small-sample bias under JM and FCS-ML (e.g., for the ICC of Y) did not increase the RMSE, indicating that these procedures remained accurate and efficient overall, even in smaller samples. All in all, our results suggest that the JM and the FCS-ML approach are equally appropriate in the context of the multilevel random intercept model.

Random slopes and cross-level interactions

Beyond the scope of random intercept models, those engaged in organizational research often seek to understand how the effects of various quantities differ across higher level organizational units. Multilevel models with random slopes allow (a) individual-level effects to vary across groups and (b) for the inclusion of group-level explanatory variables to explain that variability (i.e., CLIs). Recently, Aguinis and Culpepper (2015) stated that random slopes and CLIs were "at the heart of [...] any theory that considers outcomes to be a result of combined influences emanating from different levels of analysis," adding that "the extent to which we understand the presence of cross-level interactions is an indication of theoretical progress" (p. 156).

Consider a multilevel random coefficients model (Snijders & Bosker, 2012b) in which an individual-level outcome Y is regressed on an individual-level variable X and a group-level variable W. In addition to the random intercept, we allow the individual-level slope to vary across groups, and we include a CLI to account for some of that variation

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + \gamma_{02}W_j + \gamma_{11}W_j(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{03}W_j\bar{X}_{\bullet j} + \\ & u_{0j} + u_{1j}(X_{ij} - \bar{X}_{\bullet j}) + e_{ij}\,, \end{aligned} \tag{5}$$

where u_{1j} denotes the random slope associated with $(X_{ij} - \bar{X}_{\bullet j})$ in group j, γ_{11} denotes the CLI, and γ_{03} denotes the group-level interaction of $\bar{X}_{\bullet j}$ and W_j. The random effects $(u_{0j}, u_{1j})^T$ are

assumed to follow a multivariate normal distribution with mean zero and covariance matrix **T**.

Two aspects of the model in Equation 5 are worth noting. First, the slope of the regression of Y on X is assumed to vary across groups. Incorporating the variability in the slope in the imputation model is particularly important if the slope variance itself is of interest, because ignoring the slope variance may lead one to underestimate it in subsequent analyses. Second, the CLI denotes the degree to which the effect of $(X_{ij} - \bar{X}_{\bullet j})$ changes as a function of W_j. Thus, if estimating the CLI is of interest, then the imputation model should allow for the individual-level effect of X to interact with W (similarly for the interaction at the group level).

Accommodating random slopes and CLIs

In contrast to applications in the multilevel random intercept model, performing MI is not straightforward when the model of interest includes random slopes and CLIs, particularly when the explanatory variables contain missing data (e.g., Enders et al., 2016; Gottfredson et al., 2017; Grund, Lüdtke, & Robitzsch, 2016a). For example, consider the model of interest in Equation 5. In order for an imputation model to be consistent with this model of interest, it has to acknowledge the fact that the relation between X and Y is assumed to vary both systematically as a function of W (i.e., due to the interaction effects) and unsystematically (i.e., due to random slopes). However, the presence of such terms implies a complex joint distribution for the dependent and explanatory variables which is difficult to emulate in conventional software for multilevel MI (e.g., S. Kim et al., 2015). More advanced methods for accommodating the model of interest when generating imputations are currently being developed, but these are not yet available in standard software for multilevel MI (for further details, see the Discussion section). For this reason, we focus on the procedures that are available in standard software for multilevel MI, which often provide options for accommodating random slopes and CLIs, albeit to different (and arguably imperfect) extents.

Joint modeling (JM). As mentioned earlier, modeling the joint distribution of the dependent and explanatory variables in a general manner is not a straightforward endeavor if the model of interest includes random slopes or interaction effects. For this reason, we used `pan` to implement the JM in a manner that is similar to what we presented above. We assume that X and Y are

treated as target variables (i.e., on the left-hand side), whereas W is assumed to be completely observed and written on the right-hand side of the model. Thus, the imputation model becomes

$$\left[X_{ij}, Y_{ij}\right]^T = \left[\beta_{0(x)}, \beta_{0(y)}\right]^T + W_j \left[\beta_{w(x)}, \beta_{w(y)}\right]^T + \left[u_{j(x)}, u_{j(y)}\right]^T + \left[e_{ij(x)}, e_{ij(y)}\right]^T , \quad (6)$$

where $\left[\beta_{w(x)}, \beta_{w(y)}\right]^T$ denotes the vector of regression coefficients from regressing X and Y on W, and the remaining notation is as before. In this specification, the joint model includes possible relations among the three variables at the group level as well as relations between X and Y at the individual and the group level. However, the joint model includes only a random intercept for each target variable, whereas the slope variance and the interaction effects in the analysis model are completely ignored. In general, the JM approach may still provide reasonable estimates of the regression coefficients when the substantive analysis model contains random slopes because the inclusion of random slopes does not change the expected value of the estimates for the regression coefficients. However, when the substantive model also includes interaction effects, the integrity of its estimates may be compromised.

The FCS approach. To address missing data in multilevel models with random slopes, it has been recommended that researchers specify conditional models that include varying slopes between pairs of variables (Enders et al., 2016). In addition, product terms involving W can be introduced to accommodate the CLI. If both random slopes and product terms are included, the two conditional models become

$$\begin{aligned} X_{ij} &= \beta_{0(x)} + \beta_{1(x)}(Y_{ij} - \bar{Y}_{\bullet j}) + \beta_{2(x)}\bar{Y}_{\bullet j} + \beta_{w(x)}W_j + \beta_{1wy(x)}W_j(Y_{ij} - \bar{Y}_{\bullet j}) + \beta_{2wy(x)}W_j\bar{Y}_{\bullet j} + \\ &\quad u_{0j(x)} + u_{1j(x)}(Y_{ij} - \bar{Y}_{\bullet j}) + e_{ij(x)} \\ Y_{ij} &= \beta_{0(y)} + \beta_{1(y)}(X_{ij} - \bar{X}_{\bullet j}) + \beta_{2(y)}\bar{X}_{\bullet j} + \beta_{w(y)}W_j + \beta_{1xw(y)}W_j(X_{ij} - \bar{X}_{\bullet j}) + \beta_{2xw(y)}W_j\bar{X}_{\bullet j} + \\ &\quad u_{0j(y)} + u_{1j(y)}(X_{ij} - \bar{X}_{\bullet j}) + e_{ij(y)} , \end{aligned} \quad (7)$$

where $u_{0j(\cdot)}$ and $u_{1j(\cdot)}$ denote the random intercepts and slopes in the conditional models, and the coefficients $\beta_{1wy(x)}$, $\beta_{1xw(y)}$, $\beta_{2wy(x)}$, and $\beta_{2xw(y)}$ denote the interaction effects by which the within- and between-group relations of X and Y change as a function of W.

There are two aspects worth noting. The first is related to the way in which random slopes are handled in the conditional models. The imputation model for missing values in Y is identical to the analysis model. Thus, imputing Y should be straightforward. However, previous research

has shown that missing values in the explanatory variable X pose a much greater challenge because "reversing" the random slope model may produce biased estimates of the regression coefficients and the slope variance in the analysis model (Gottfredson et al., 2017; Grund, Lüdtke, & Robitzsch, 2016a; see also Enders et al., 2016). This is not entirely surprising because the analysis model (Equation 5) and the imputation model for missing X (Equation 7, first line) make different statements about the varying relation between X and Y. In other words, although replacing $u_{1j}(X_{ij} - \bar{X}_{\bullet j})$ in the analysis model by $u_{1j(x)}(Y_{ij} - \bar{Y}_{\bullet j})$ in the imputation model may serve as a "proxy" for the relation of interest, the two statements are not equivalent.

The second aspect is related to the presence of nonlinear effects (i.e., interaction effects) in the conditional models. At each iteration of the FCS algorithm, the product terms $W_j(X_{ij} - \bar{X}_{\bullet j})$ and $W_j(Y_{ij} - \bar{Y}_{\bullet j})$ must be "updated" to incorporate the most recent imputations of X and Y. The simplest strategy for updating the products terms is to recalculate them after X and Y have been imputed. This is commonly referred to as "passive imputation" (Royston, 2004; van Buuren, 2012). As an alternative, product terms may be regarded as "just another variable" (von Hippel, 2009). This strategy replaces the passive imputation step with an imputation model for each product term (e.g., a regression model). However, both strategies have been shown to yield biased parameter estimates (e.g., Seaman et al., 2012; Vink & van Buuren, 2013) because they do not correctly reflect the complex joint distribution of the dependent and explanatory variables in the model when the model of interest includes interaction effects (S. Kim et al., 2015). In the present study, we used passive imputation because (a) it is easy to use and readily available in standard software and (b) implementing "just another variable" is not straightforward with group-mean-centered data.[3]

FIML. Similar to MI, analyzing the incomplete data with FIML can be difficult if the model of interest includes random slopes and CLIs. As before, we focus on FIML estimation in the statistical software M*plus*. If missing data occur only on Y, estimating the model of interest in M*plus* is straightforward because the observed-data likelihood can be evaluated directly on

[3]Under FCS, fitting the imputation models requires that some instances of the product terms are observed. However, with missing data at the individual level, the group means are no longer known, rendering the product terms unobserved as a result. In other words, it is not clear how "just another variable" might differentiate interaction effects at the individual level, the group level, and across levels without knowing the group means.

the basis of the incomplete data. However, if missing values occur on X, it is currently not possible to include X in the analysis model in M*plus* without dropping cases with missing X from the analysis (for a discussion, see also Shin & Raudenbush, 2010).

Summary. In contrast to applications in the multilevel random intercept model, missing data pose a greater challenge when the model of interest includes random slopes. Multilevel MI can be expected to provide proper results when only the dependent variable Y contains missing data. However, if the explanatory variable with a random slope, X, contains missing data, conducting MI is not straightforward. Specifically, the "reversed" imputation model for missing X contains only a proxy for the relation of interest, and accommodating product terms (i.e., CLIs and group-level interactions) is still an open area of research (see the Discussion section). As a result, neither JM nor FCS was expected to provide perfect results.

Study 2: Random slope models

In this section, we present findings from a simulation study in which we compared different MI procedures in the context of multilevel models with random slopes and CLIs. The model of interest was the random coefficients model presented in Equation 5, and the data were also generated from this model (see Appendix A). The parameters of the data-generating model were chosen in such a way as to imply a given value for the ICCs of X and Y and a given "total" variance for the random slope (i.e., $Var(\beta_{1j}) = Var(\gamma_{11}W_j + u_{1j}) = \gamma_{11}^2 + \tau_1^2$). Missing data were generated as before (MCAR and MAR on either X or Y). A summary is presented in Table 1. We varied the number of groups (k = 50, 100, 200, 500), the number of individuals (n = 5, 10), the size of the CLI (γ_{11} = 0, .20), and the missing data mechanism. The effect of $(X_{ij} - \bar{X}_{\bullet j})$ was held constant at .50 (γ_{10}) and the effect of W_j at .35 (γ_{02}); the remaining effects were set to zero. Each condition was replicated 1,000 times. We applied the following procedures to each data set:

1. LD
2. single-level FCS, ignoring the multilevel structure and product terms (FCS-SL)

3. multilevel FCS, ignoring random slopes but including passive imputation of product terms (FCS-CLI/no RS)
4. multilevel FCS including random slopes and passive imputation of product terms (Equation 7; FCS-CLI/RS)
5. FIML

The FIML estimation was conducted in M*plus* as described above. Because FIML could not be used to estimate the model in conditions with missing X, we included FIML only for conditions with missing Y. The parameters of interest were the within-group regression coefficient of X (γ_{10}), the effect of W (γ_{02}), the CLI (γ_{11}), and the slope variance (τ_1^2). For each condition, each procedure, and each parameter, we calculated the bias, the RMSE, and the coverage rate of the 95% confidence interval as before.

Results. Our main findings are summarized in Table 3 and Figure 2. In presenting our results, we focus on the MI procedures because FIML could be applied only in conditions with missing Y, and the estimates were approximately unbiased in these conditions. For the remaining procedures, the difference between cases with missing data in Y and X was substantial, and sample size continued to play an important role. When only the dependent variable Y was incomplete, FCS-CLI/RS provided approximately unbiased estimates for the parameters of interest, with bias present for the slope variance (τ_1^2) in smaller samples but tending toward zero as the samples grew larger (Figure 2). The bias in smaller samples was quite large for the slope variance, especially when the samples consisted of smaller groups ($n = 5$). With larger groups ($n = 10$), the bias was reduced by approximately half (Figure 2).[4] When the random slope was ignored (FCS-CLI/no RS), we obtained almost identical estimates for the regression coefficients, but the slope variance was underestimated regardless of sample size. When both the interaction effects and the random slopes were ignored (JM), the estimates of the CLI were biased as well. Moreover, when the imputation model ignored the multilevel structure

[4]The bias for the slope variance in small samples may be explained by the Bayesian prior distribution employed during MI. In our study, we used the standard "least-informative" inverse-Wishart prior distributions, which imply relatively large values for the variance components as compared with the true size of the slope variance (here, $\tau_1^2 = .10$ if $\gamma_{11} = 0$, and $\tau_1^2 = .06$ if $\gamma_{11} = .20$). Such problems are well known in the Bayesian literature and can often be mitigated by choosing the prior distribution on the basisof the data (McNeish, 2016).

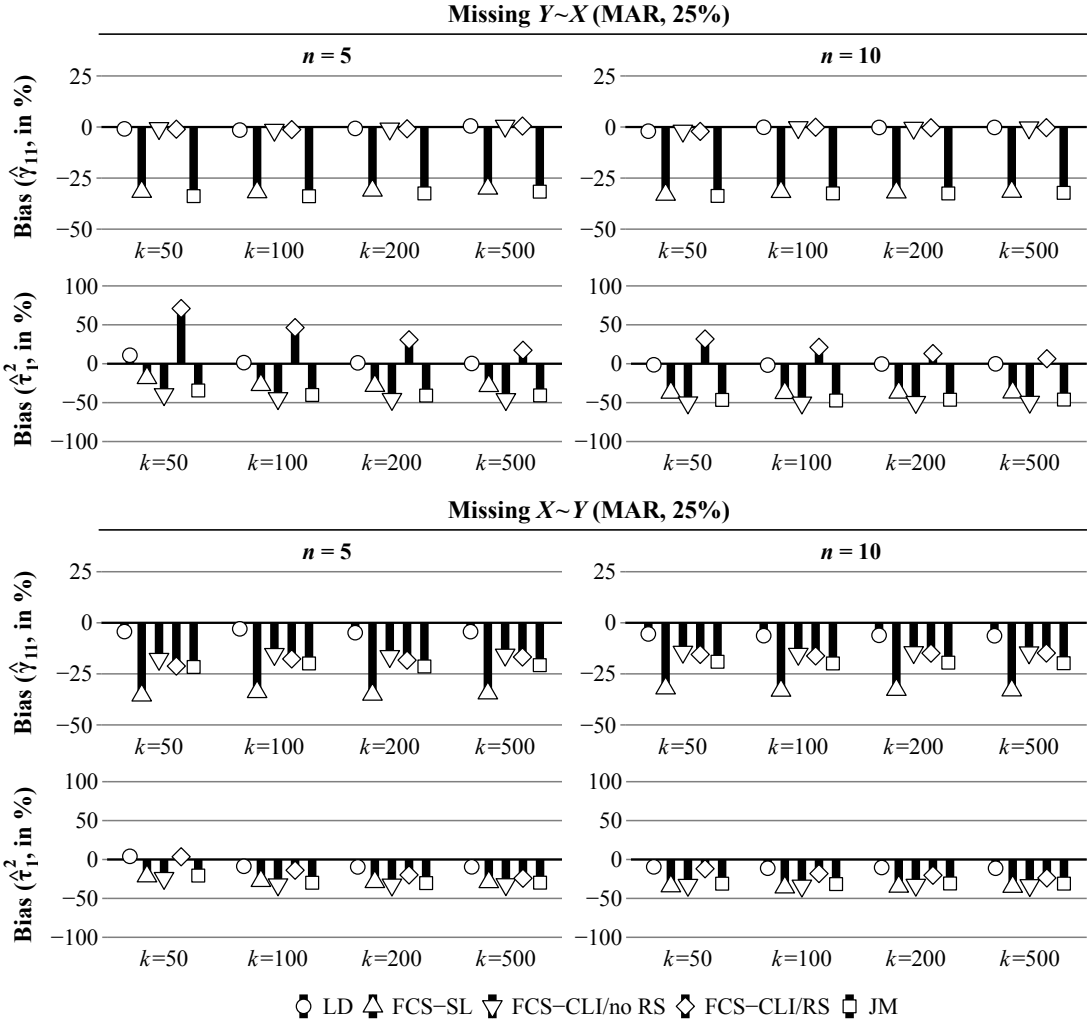

Figure 2: Estimated bias for the CLI (γ_{11}) and the slope variance (τ_1^2) in Study 2 for different numbers of individuals (n) and groups (k), and different missing data mechanisms (MAR; $Y \sim X$ and $X \sim Y$). LD = listwise deletion; FCS-SL = single-level FCS; FCS-CLI/no RS = multilevel FCS including only product terms; FCS = multilevel FCS including product terms and random slopes; JM = multilevel JM.

altogether (FCS-SL), all regression coefficients were biased independent of sample size. In conditions with no CLI ($\gamma_{11} = 0$), the performance of FCS-CLI/RS was the same, but the bias in the slope variance was greatly reduced (see Footnote 4). In these cases, both FCS-CLI/no RS and JM also provided approximately unbiased estimates of the regression coefficients (see also Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016a). Finally, LD provided approximately unbiased estimates of the slope variance, but the estimates of the regression coefficient for W (γ_{02}) were biased under MAR regardless of sample size. The results for the coverage of the 95% confidence interval and the RMSE were in line with the bias. However, the coverage was slightly too low under FCS-CLI/no RS and JM, illustrating that the confidence intervals were

Table 3: Bias (in %), RMSE, and Coverage of the 95% Confidence Interval for the Within-Group Regression Coefficient of X, the Between-Group Regression Coefficient of W, and the CLI of X with W in Study 2 (Small Groups, $n = 5$)

	LD			FCS-SL			FCS-CLI/no RS			FCS-CLI/RS			JM		
	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.
Missing $Y \sim X$ (MAR, 25%)															
$k = 100$							$\rho_{I,X} = \rho_{I,Y} = .10$								
$\hat{\gamma}_{10}$	0.0	0.06	94.7	−7.2	0.07	91.0	0.2	0.06	92.4	0.0	0.06	95.7	−2.2	0.06	92.8
$\hat{\gamma}_{02}$	−10.3	0.06	89.5	−11.1	0.06	87.8	−0.4	0.05	95.5	−0.2	0.05	94.0	−9.6	0.06	91.2
$\hat{\gamma}_{11}$	0.1	0.06	94.5	−30.7	0.08	83.4	0.2	0.06	92.4	0.3	0.06	96.1	−32.5	0.08	80.1
$k = 500$															
$\hat{\gamma}_{10}$	−0.1	0.03	95.4	−7.6	0.05	68.3	−0.2	0.03	93.1	−0.1	0.02	95.9	−2.4	0.03	90.8
$\hat{\gamma}_{02}$	−9.8	0.04	70.0	−10.6	0.04	57.9	−0.0	0.02	95.5	0.1	0.02	95.6	−9.0	0.04	71.7
$\hat{\gamma}_{11}$	−0.5	0.03	95.2	−30.9	0.06	24.2	−0.5	0.02	94.1	−0.5	0.02	96.0	−32.2	0.07	16.5
$k = 100$							$\rho_{I,X} = \rho_{I,Y} = .50$								
$\hat{\gamma}_{10}$	0.2	0.06	94.0	−15.6	0.10	83.5	0.2	0.06	92.3	0.1	0.06	95.2	−1.3	0.06	92.1
$\hat{\gamma}_{02}$	−6.4	0.08	95.0	−6.7	0.08	90.5	0.3	0.08	95.1	0.4	0.08	94.5	−4.9	0.08	95.1
$\hat{\gamma}_{11}$	1.1	0.06	94.7	−31.0	0.08	92.8	1.0	0.06	93.6	1.1	0.06	96.1	−31.9	0.08	83.9
$k = 500$															
$\hat{\gamma}_{10}$	0.2	0.03	95.4	−15.5	0.08	27.7	0.2	0.03	93.9	0.2	0.03	95.7	−1.3	0.03	92.0
$\hat{\gamma}_{02}$	−6.2	0.04	90.7	−6.2	0.04	83.9	0.4	0.03	94.7	0.4	0.03	94.7	−4.7	0.04	91.5
$\hat{\gamma}_{11}$	−0.8	0.03	94.9	−31.4	0.07	44.2	−0.8	0.03	93.1	−0.7	0.03	95.6	−33.1	0.07	17.1
Missing $X \sim Y$ (MAR, 25%)															
$k = 100$							$\rho_{I,X} = \rho_{I,Y} = .10$								
$\hat{\gamma}_{10}$	−4.9	0.06	91.4	−12.6	0.08	80.7	−2.9	0.06	92.8	−5.9	0.06	91.0	−2.7	0.06	93.4
$\hat{\gamma}_{02}$	−14.9	0.07	81.6	−1.7	0.04	95.2	0.1	0.04	95.5	0.0	0.04	95.2	−0.2	0.04	95.2
$\hat{\gamma}_{11}$	−4.2	0.06	93.8	−28.4	0.07	87.8	−14.8	0.05	94.3	−17.3	0.06	95.2	−19.9	0.06	93.2
$k = 500$															
$\hat{\gamma}_{10}$	−5.1	0.04	82.4	−12.5	0.07	23.7	−3.2	0.03	87.9	−4.7	0.03	84.3	−2.8	0.03	89.9
$\hat{\gamma}_{02}$	−15.1	0.06	32.6	−2.0	0.02	92.6	−0.1	0.02	96.0	−0.1	0.02	95.8	−0.5	0.02	95.2
$\hat{\gamma}_{11}$	−4.9	0.03	94.0	−28.4	0.06	29.5	−15.3	0.04	77.2	−16.5	0.04	75.0	−20.2	0.04	62.3
$k = 100$							$\rho_{I,X} = \rho_{I,Y} = .50$								
$\hat{\gamma}_{10}$	−2.6	0.06	94.1	−37.9	0.19	4.3	−3.7	0.06	93.3	−6.6	0.06	92.8	−3.1	0.06	93.3
$\hat{\gamma}_{02}$	−10.7	0.08	89.9	−3.1	0.08	95.1	−0.6	0.07	95.3	−0.7	0.07	95.5	−0.4	0.07	95.2
$\hat{\gamma}_{11}$	−2.9	0.06	93.8	−54.5	0.12	46.5	−16.7	0.06	92.2	−19.5	0.06	92.4	−21.7	0.06	90.8
$k = 500$															
$\hat{\gamma}_{10}$	−3.2	0.03	89.5	−38.6	0.19	0.0	−3.4	0.03	87.6	−4.8	0.03	82.0	−2.9	0.03	88.8
$\hat{\gamma}_{02}$	−10.5	0.05	78.2	−2.5	0.03	93.7	−0.1	0.03	95.3	−0.1	0.03	95.5	−0.0	0.03	95.2
$\hat{\gamma}_{11}$	−2.6	0.03	94.4	−54.4	0.11	0.1	−15.6	0.04	76.9	−16.6	0.04	74.6	−20.9	0.05	60.4

Note. $\hat{\gamma}_{10}$ = within-group regression coefficient of X; $\hat{\gamma}_{11}$ = CLI; LD = listwise deletion; FCS-CLI/no RS = multilevel FCS including only product terms; FCS-CLI/RS = multilevel FCS including product terms and random slopes; JM = multilevel JM.

slightly too narrow when the slope variance was omitted from the imputation model.

When missing values occurred in the explanatory variable X, no procedure provided unbiased estimates of the CLI and the slope variance (see Table 3 and Figure 2). Even when the product terms and random slopes were included in the model (FCS-CLI/RS), multilevel MI

provided only biased estimates of the CLI and the slope variance. Ignoring the slope variance (FCS-CLI/no RS) led to slightly better estimates of the regression coefficients but increased the bias in the slope variance. Ignoring both the interaction effects and the random slopes (JM) led to further bias in the CLI but was otherwise comparable to FCS-CLI/no RS. On the other hand, single-level MI (FCS-SL) led to strongly biased estimates of both the main and interaction effects as well as the slope variance. It is interesting that LD provided the least biased estimates of the CLI and the slope variance in conditions with small groups even under MAR. On the other hand, LD introduced bias into the other estimates, particularly the main effect of W when the data were MAR. In conditions with no CLI ($\gamma_{11} = 0$), FCS-CLI/RS still showed a slight downward bias in the regression coefficient of $(X_{ij} - \bar{X}_{\bullet j})$ and the slope variance but yielded otherwise unbiased results. Ignoring the slope variance (FCS-CLI/no RS and JM) reduced the bias in the regression coefficients to essentially zero but increased bias in the slope variance. Results for LD were similar to conditions with CLI.

The coverage of the 95% confidence interval and the RMSE were closely related to the bias in the parameter estimates. For FCS-CLI/RS, the coverage was close to the nominal value of 95% for most parameters, but the coverage of the regression coefficient of $(X_{ij} - \bar{X}_{\bullet j})$ and the CLI dropped below 90% unless the sample was very small ($k = 50$, $n = 5$). As a result of reduced sample size, the coverage under LD was slightly higher but also fell below 90% as the sample size increased. Similar to before, the RMSE indicated a relative loss of efficiency under LD for estimates of the regression coefficients of W and to a lesser extent $(X_{ij} - \bar{X}_{\bullet j})$. For example, the average RMSE for regression coefficients of W were 13.8% larger under LD with missing Y as compared with FCS-CLI/RS and 38.7% larger with missing X. For the CLI, the RMSE was usually lowest under FCS-CLI/no RS and FCS-CLI/RS in smaller samples ($k \leq 100$) and under LD in larger samples ($k \geq 200$). However, these differences were very small: The average RMSE for the CLI under LD as compared with FCS-CLI/RS was approximately equal with missing Y (< 1%) and only 3.5% larger with missing X.

Taken together, our results indicate that the FCS approach provides reliable estimates for the parameters of interest when missing values are restricted to Y and still reasonable (though imperfect) estimates with missing values on X. Even though some parameter estimates obtained

from FCS were biased, they had better statistical properties overall than those of the competing methods. Ignoring the slope variance sometimes reduced bias in the regression coefficients but resulted in confidence intervals for these coefficients that were too narrow. LD provided the least biased estimates of the slope variance and the CLI but introduced bias in other parameters and tended to be slightly less efficient than MI.

Categorical and group-level missing data

In the previous two simulation studies, the models of interest were simplified in two ways: (a) the data were always continuous, thus not accounting for missing categorical data, and (b) data were missing only at the individual level, thus not accounting for missing data in group-level variables. Therefore, we conducted two smaller simulation studies that addressed these issues separately.

Study 3a: Missing categorical data

Turning back to the random intercept model, researchers are often interested in estimating the differences between groups of participants by including categorical variables in the model of interest, be it to control for group differences (e.g., due to gender, education, etc.) or to assess the effectiveness of interventions (e.g., treatment vs. control group). Especially in the former case, categorical variables may contain missing data. Here, we briefly discuss two procedures for multilevel MI—one using JM, one using FCS—that address missing data in multilevel categorical variables. We also discuss FIML estimation, and we evaluate their performance in a simulation study.

Here, the model of interest is a multilevel random intercept model with two explanatory variables at the individual level, one continuous and one binary

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{20}D_{ij} + u_{0j} + e_{ij}\,, \tag{8}$$

where D is a dummy-coded binary variable that takes on values for each individual i in group j. This model was also used to generate the data (see Appendix A). To ensure that D had a multilevel structure, we simulated a latent background variable D^* with a given value of its ICC.

Binary values were obtained by setting $D_{ij} = 1$ if $D^*_{ij} > 0$, and 0 otherwise, resulting in a 50% prevalence of either category. Missing data were induced in D as before (MCAR and MAR, based on Y). In addition, we varied the number of individuals ($n = 5, 10$) and the number of groups ($k = 50, 100, 200, 500$). The remaining parameters were held constant (see Table 1). For comparison, we included LD and single-level FCS as before.

Joint modeling (JM). Quite general procedures that use the JM approach are available for categorical data (e.g., Goldstein et al., 2009). These procedures have been implemented recently in the `jomo` package in R (Quartagno & Carpenter, 2016a), which allows continuous and categorical variables to be modeled simultaneously, where a categorical variable is represented by $c - 1$ underlying latent continuous variables (where c is the number of categories). For our model of interest involving three individual-level variables, the joint model reads

$$\left[X_{ij}, Y_{ij}, D^*_{ij}\right]^T = \left[\beta_{0(x)}, \beta_{0(y)}, \beta_{0(d*)}\right]^T + \left[u_{j(x)}, u_{j(y)}, u_{j(d*)}\right]^T + \left[e_{ij(x)}, e_{ij(y)}, e_{ij(d*)}\right]^T , \quad (9)$$

where $e_{ij(d*)}$ is constrained to have unit variance to identify the model. For missing data in D, the model is essentially a generalized linear mixed-effects model conditioning on X and Y (see Carpenter & Kenward, 2013; Goldstein et al., 2009). For dichotomous variables, equivalent procedures are available in the statistical software M*plus*. However, for categorical variables with multiple categories, the procedures in M*plus* differ from the approach taken in `jomo`.[5]

The FCS approach. Imputations for D may also be generated by directly conditioning on X and Y using FCS. Similar to the joint model, imputations may be generated from a generalized linear mixed-effects model (e.g., with a probit or logit link function)

$$\begin{aligned} D^*_{ij} = {} & \beta_{0(d*)} + \beta_{1(d*)}(X - \bar{X}_{\bullet j}) + \beta_{2(d*)}\bar{X}_{\bullet j} + \beta_{3(d*)}(Y_{ij} - \bar{Y}_{\bullet j}) + \beta_{4(d*)}\bar{Y}_{\bullet j} \\ & + u_{j(d*)} + e_{ij(d*)} , \end{aligned} \quad (10)$$

where $e_{ij(d*)}$ is constrained in a manner similar to what is done in the JM. Unfortunately, `mice` currently allows for MI of categorical variables only in single-level models. Procedures for multilevel data have been proposed by Snijders and Bosker (2012b) and Zinn (2013). The

[5] When using the imputation module in M*plus* to implement multilevel MI, categorical variables are treated as ordinal, and a single latent variable is used for each categorical variable regardless of the number of categories; variables with multiple categories are addressed by estimating $c - 1$ threshold parameters (Asparouhov & Muthén, 2010b).

Table 4: Bias (in %), RMSE, and Coverage of the 95% Confidence Interval for the Overall Regression Coefficients in Study 3a (Missing $D \sim Y$, MAR, 25%)

	LD			FCS-SL			FCS-ML			JM			FIML		
	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.
Missing $D \sim Y$ (MAR, 25%)															
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .10$														
γ_{10}	−6.0	0.05	90.2	−0.1	0.04	95.5	−0.1	0.04	95.3	−0.0	0.04	94.8	−0.1	0.04	95.0
γ_{20}	−6.1	0.09	95.4	−1.1	0.09	95.5	−0.3	0.09	95.0	−1.5	0.09	95.0	−0.7	0.09	95.0
$k = 500$															
γ_{10}	−5.9	0.04	69.3	−0.1	0.02	95.3	−0.1	0.02	94.9	−0.1	0.02	95.1	−0.1	0.02	94.8
γ_{20}	−5.9	0.04	94.5	−1.1	0.04	95.2	−0.3	0.04	95.8	−0.2	0.04	96.0	−0.7	0.04	95.5
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .50$														
γ_{10}	−4.2	0.05	93.2	0.8	0.04	95.3	0.0	0.04	95.0	−0.1	0.04	95.0	0.4	0.04	95.1
γ_{20}	−2.6	0.09	93.9	−17.2	0.08	93.1	−2.2	0.09	94.5	−0.4	0.09	94.4	−8.1	0.08	92.7
$k = 500$															
γ_{10}	−3.9	0.03	82.6	1.0	0.02	92.8	0.2	0.02	93.9	0.0	0.02	93.9	0.6	0.02	93.0
γ_{20}	−2.7	0.04	95.7	−17.4	0.05	84.4	−2.3	0.04	95.6	0.6	0.04	94.9	−8.6	0.04	91.2

Note. $\hat{\gamma}_{10}$ = overall regression coefficient of X; $\hat{\gamma}_{20}$ = overall regression coefficient of D; LD = listwise deletion; FCS-SL = single-level FCS; FCS-ML = multilevel FCS; JM = multilevel JM; FIML = full-information maximum likelihood.

procedure used here is essentially a combination of the two and is implemented in the R package `miceadds` (Robitzsch, Grund, & Henke, 2016).

FIML. As an alternative to MI, the model can also be estimated directly by applying FIML. However, because M*plus* assumes that the variables in the multilevel model are multivariate normal, it was not straightforward to include D as a multilevel categorical variable. Instead, we treated D as a multilevel continuous normal variable to estimate the model with FIML.

Results. Our main findings are summarized in Table 4. We restricted our reporting to the overall effects of X (γ_{10}) and D (γ_{20}) because we felt that they were the most important parameters for judging the performance of each method. Consistent with our expectations, both FCS-ML and JM provided approximately unbiased estimates of the two regression coefficients, whereas the other procedures each yielded biased estimates in some simulated conditions. The coverage was close to the nominal value of 95%, and the RMSE tended to be lowest under FCS-ML and JM. By contrast, LD yielded biased estimates of the two parameters. FCS-SL and FIML introduced bias in the regression coefficient of D (γ_{20}) in conditions with large ICCs, although the RMSE and coverage remained acceptable under FIML. We concluded that both multilevel JM and FCS are suitable for MI of multilevel categorical data. Note, however, that we

limited our attention to missing binary data. The FCS procedure can be extended to variables with multiple ordered or unordered categories.

Study 3b: Group-level missing data

The ideal case in which missing data occur only on the lowest level of multilevel data sets (i.e., on the level of individuals) rarely holds in practice. Moreover, data that are missing at the group level can be particularly cumbersome because they can force researchers to discard complete records at lower levels of the data. For example, consider a study in which employees were asked to rate the frequency of benevolent behavior engaged in by supervisors, and supervisors were asked the same question about their employees. If both variables were to be used as explanatory variables in some model of interest, missing data in supervisor ratings would lead one to discard employees' ratings as well, resulting in a severe loss of information. Surprisingly, the methodological literature has focused so far on ad hoc procedures, for example, separate imputation of individual- and group-level variables (Gibson & Olejnik, 2003) or "flat file" imputation using single-level MI (Cheung, 2007; for an overview, see Hox et al., 2016; van Buuren, 2011). However, recent advances in statistical software have greatly improved our ability to treat group-level missing data. Here, we briefly discuss two procedures—one using JM, one using FCS—that can be used to impute missing data at the group level.

Here, the model of interest was a multilevel random intercept model with explanatory variables at the individual (e.g., employee ratings) and the group level (e.g., supervisor ratings)

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + \gamma_{02}W_j + u_{0j} + e_{ij} \; . \tag{11}$$

Furthermore, we assumed that W was partially missing. The critical point in this model is that W is measured at the group level; that is, it does not vary across individuals in the same group. Thus, information located at the group level, including the information provided by individual-level variables, may be used to predict missing scores in W. For comparison, we also included LD and single-level FCS.

Joint modeling (JM). Computationally, the imputation of missing data at the group level is not much different from imputation at the individual level. Specifically, imputations for

group-level missing data can be obtained by conditioning on observed group-level variables and on the between-group components of individual-level variables by employing the same general paradigm that is already employed for multilevel MI (for details, see Carpenter & Kenward, 2013; Goldstein et al., 2009). Similar to before, the joint model for the three variables of interest can be written as

$$\left[X_{ij}, Y_{ij}, W_j\right]^T = \left[\beta_{0(x)}, \beta_{0(y)}, \beta_{0(w)}\right]^T + \left[u_{j(x)}, u_{j(y)}, u_{j(w)}\right]^T + \left[e_{ij(x)}, e_{ij(y)}, 0\right]^T \quad . \tag{12}$$

where $u_{j(w)}$ is the residual of W in group j. For missing values in W, imputations are generated in the present case by conditioning on $\left[u_{j(x)}, u_{j(y)}\right]$ at each iteration of the sampling algorithm (see Carpenter & Kenward, 2013), thus incorporating the group-level information supplied by X and Y in the prediction of missing W. The joint model can be implemented, for example, with the `jomo` package in R or in the statistical software M*plus*.

The FCS approach. Instead of conditioning on the random effects of individual-level variables as in the JM approach, group-level variables can be imputed by applying an FCS approach based on the observed group means of these variables. Specifically, for missing values in W, missing data may be imputed by using the linear regression

$$W_j = \beta_{0(w)} + \beta_{1(w)}\bar{X}_{\bullet j} + \beta_{2(w)}\bar{Y}_{\bullet j} + u_{j(w)} \,, \tag{13}$$

where $u_{j(w)}$ is the residual of W given $\bar{X}_{\bullet j}$ and $\bar{Y}_{\bullet j}$. If values are missing at both levels, then the FCS algorithm iterates back and forth between the individual- and group-level equations (Equations 4 and 13; see also Gelman & Hill, 2006; Yucel, 2008). As in the multilevel random intercept model, it can be argued that the FCS and the JM approach imply similar covariance structures that can be used interchangeably (see Study 1; Carpenter & Kenward, 2013).

FIML. Missing values in W can also be addressed using FIML. Because W is directly measured at the group-level, missing data in W can be addressed simply by specifying W as a latent variable in M*plus*.

Results. Our main findings are summarized in Table 5. We restricted our reporting to the group-level effects of X (γ_{01}) and W (γ_{02}) because we considered them to be the most important in this situation. FCS-ML, JM, and FIML all provided approximately unbiased estimates of the

Table 5: Bias (in %), RMSE, and Coverage of the 95% Confidence Interval for the Group-Level Regression Coefficients in Study 3b (Missing $W \sim Y$, MAR, 25%)

	LD			FCS-SL			FCS-ML			JM			FIML		
	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.	Bias	RMSE	Covg.
Missing $W \sim Y$ (MAR, 25%)															
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .10$														
γ_{01}	−0.8	0.10	94.7	6.3	0.09	94.4	1.3	0.09	94.4	2.9	0.09	95.0	0.7	0.09	94.2
γ_{02}	−3.3	0.05	95.3	−0.6	0.06	94.9	−1.4	0.05	95.0	−5.9	0.05	95.9	−0.3	0.05	95.5
$k = 500$															
γ_{01}	−3.0	0.05	95.4	5.8	0.04	94.1	−0.0	0.04	94.9	0.4	0.04	95.3	−0.1	0.04	94.8
γ_{02}	−3.2	0.03	92.7	−0.8	0.03	94.1	−0.5	0.03	93.5	−2.8	0.03	94.1	−0.3	0.03	93.2
$k = 100$	$\rho_{I,X} = \rho_{I,Y} = .50$														
γ_{01}	−7.1	0.11	95.1	−0.8	0.10	95.3	−0.8	0.10	95.2	−0.7	0.10	95.6	−1.1	0.10	95.1
γ_{02}	−7.9	0.08	95.3	6.8	0.09	92.3	−2.8	0.09	95.0	−2.6	0.09	95.0	−1.1	0.09	94.2
$k = 500$															
γ_{01}	−6.7	0.05	93.6	0.0	0.04	94.8	−0.1	0.04	95.3	−0.1	0.05	94.7	−0.2	0.04	94.9
γ_{02}	−6.9	0.04	93.7	8.3	0.05	89.8	−0.5	0.04	94.6	−0.4	0.04	94.9	−0.1	0.04	94.5

Note. $\hat{\gamma}_{01}$ = between-group regression coefficient of X; $\hat{\gamma}_{02}$ = between-group regression coefficient of W; LD = listwise deletion; FCS-SL = single-level FCS; FCS-ML = multilevel FCS; JM = multilevel JM; FIML = full-information maximum likelihood.

group-level effects in the model of interest. With a smaller number of groups and individuals within each group, FCS-ML and JM exhibited a small negative bias, which tended toward zero in larger samples. By contrast, LD and FCS-SL yielded only biased estimates of these parameters regardless of sample size. The coverage of the 95% confidence interval was close to the nominal value of 95% for FCS-ML, JM, and FIML, and the RMSE was lowest for these procedures. We concluded that multilevel JM and FCS as well as FIML are suitable methods for dealing with group-level missing data.

Recommendations for practice

There exist several approaches to the treatment of missing data in multilevel designs. As a result, researchers are faced with a multitude of options, several but not all of which may be suitable for a given task. In order to guide researchers in picking a suitable procedure, we provide a detailed list of recommendations in Table 6. This table covers different applications of multilevel models, including applications with random intercepts, random slopes, different variable types, and interaction effects. For each application, we distinguish between a general case with arbitrary patterns of missing data and a number of cases with missing data on specific variables

(e.g., categorical and group-level variables). For each case, we list the recommended and not-recommended procedures as well as the likely consequences of choosing the latter. Finally, we list statistical software that implements one or more of the recommended procedures, and we

Table 6: Recommended Missing Data Treatments and Software for Different Types of Multilevel Analysis Models

Model type (example)	Missing	Recommended	Not recommended	Current software (MI)
Random intercept model $Y_{ij} = \gamma_{00}+\gamma_{10}(X_{ij}-\bar{X}_{\bullet j})+\gamma_{01}\bar{X}_{\bullet j}+u_{0j}+e_{ij}$	any	• multilevel FCS ▷ passive imputation of group means • multilevel JM ▷ all variables specified as targets	• listwise deletion ▷ biased estimates, power loss • single-level MI ▷ biased estimates and SEs • FIML[a] ▷ biased estimates when using group-mean centering	R (`mice`, `pan`, `jomo`), M*plus*, Blimp, SAS (`MMI_IMPUTE`), MLwiN, `REALCOM` → see Example 1.1 (p. 53)
. . . with categorical variables (D_{ij}) $Y_{ij} = \gamma_{00} + \gamma_{10}D_{ij} + \gamma_{20}X_{ij} + u_{0j} + e_{ij}$	D	• multilevel FCS ▷ passive imputation of group means ▷ using logistic or probit models • multilevel JM ▷ all variables specified as targets ▷ using models for mixed data types	• listwise deletion ▷ biased estimates, power loss • single-level MI ▷ biased estimates and SEs • FIML[a] ▷ biased estimates under normality assumption	R (`mice`, `jomo`), M*plus*, Blimp, `REALCOM` → see Example 1.2 (p. 55)
. . . with variables at Level 2 (W_j) $Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + \gamma_{02}W_j + u_{0j} + e_{ij}$	W	• multilevel FCS ▷ including group means • multilevel JM ▷ all variables specified as targets ▷ using models for missing data at both levels • FIML	• listwise deletion ▷ biased estimates, power loss • single-level MI ▷ biased estimates and SEs	R (`mice`, `jomo`), M*plus*, Blimp, `REALCOM` → see Example 1.3 (p. 56)
. . . with interactions or nonlinear terms $Y_{ij} = \gamma_{00}+\gamma_{10}X_{ij}+\gamma_{20}Z_{ij}+\gamma_{30}X_{ij}Z_{ij}+u_{0j}+e_{ij}$	X, Z	• multilevel FCS ▷ passive imputation of group means and product terms	• listwise deletion ▷ biased estimates, power loss	R (`mice`) → see Example 1.4 (p. 57)

(continued)

Table 6: Recommended Missing Data Treatments and Software for Different Types of Multilevel Analysis Models (continued)

Model type (example)	Missing	Recommended	Not recommended	Current software (MI)
			• single-level MI ▷ biased estimates and SEs • multilevel JM ▷ biased estimates of interaction effects • FIML[a]	
Random slope model $Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + u_{0j} + u_{1j}(X_{ij} - \bar{X}_{\bullet j}) + e_{ij}$	any	• multilevel FCS ▷ passive imputation of group means ▷ including random slopes between pairs of variables	• listwise deletion ▷ biased estimates, power loss • single-level MI ▷ biased estimates and SEs • multilevel JM ▷ biased SEs • FIML[a]	R (`mice`), Blimp → see Example 2.1 (p. 59)
. . . with interactions or nonlinear terms $Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + \gamma_{02}W_j + \gamma_{11}W_j(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{03}W_j\bar{X}_{\bullet j} + u_{0j} + u_{1j}(X_{ij} - \bar{X}_{\bullet j}) + e_{ij}$	X, W	• multilevel FCS ▷ passive imputation of group means and product terms ▷ including random slopes between pairs of variables	• listwise deletion ▷ biased estimates, power loss • single-level MI ▷ biased estimates and SEs • multilevel JM ▷ biased estimates of interaction effects and SEs • FIML[a]	R (`mice`) → see Example 2.2 (p. 60)

[a] The present recommendations refer to FIML as it is currently implemented in the statistical software M*plus*.

provide reference to one of the step-by-step examples in Appendix B, which illustrate the use of multilevel MI for each application in the statistical software R (for a general introduction to multilevel MI, see Enders et al., 2016; Grund, Lüdtke, & Robitzsch, 2016b).

For applications in the multilevel random intercept model, multilevel MI—using either JM or FCS—provides an effective and general method for dealing with missing data. Procedures for multilevel JM, for example, are implemented in the software packages `pan` and `jomo` for the statistical software R as well as M*plus*, MLwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2015), `REALCOM` (Carpenter et al., 2011), and the SAS macro `MMI_IMPUTE` (Mistler, 2013a); multilevel FCS is implemented in the R package `mice` as well as M*plus* and Blimp (Keller & Enders, 2016). When treating missing data in categorical or group-level variables, researchers should choose implementations of multilevel MI that support these types of variables (e.g., `jomo`, M*plus* and `REALCOM` for multilevel JM; `mice`, M*plus*, and Blimp for multilevel FCS). FIML may be an option if missing data are restricted to the dependent variable in the analysis or if the analysis model includes latent instead of observed (i.e., manifest) group means to estimate group-level effects (Grund et al., in press-a; Lüdtke et al., 2008). By contrast, single-level MI should be avoided unless only a few cases contain missing data (e.g., less than 5%) and the ICC of the variables is relatively small (e.g., less than .10). Similarly, although LD provided reasonable estimates of model parameters (e.g., the CLI), we do not recommend that it be adopted in practice. This is because LD provides generally unbiased results only under MCAR, whereas its performance under MAR depends on the "strength" of missing data mechanism (i.e., the degree to which the data loss is systematic; see also Newman, 2014). This is problematic, because the missing data mechanism can never be ascertained from the data alone (e.g., Allison, 2001; Enders, 2010). For that reason, LD may provide an alternative if missing data are guaranteed to be MCAR, for example, in "planned missing data designs" (e.g., Graham, Taylor, Olchowski, & Cumsille, 2006). However, under more general conditions, we recommend against using LD. For applications involving random slopes or interaction effects, it is more difficult to provide general recommendations at the present time. Software for multilevel FCS may be used to treat missing data in such models if it supports the specification of random slope imputation models as well as passive imputation steps for the product terms (e.g., `mice`).

However, researchers should bear in mind that multilevel FCS with passive imputation is not a definite solution to the problem of missing data in such applications. Instead, model-based procedures may be considered in the future (for a brief exposition, see the Discussion section).

Apart from the procedure selected for the treatment of missing data, the performance of MI also depends on a few general factors. For example, researchers should try to include auxiliary variables in the imputation model, that is, variables that are related to either the occurrence of missing data or the variables with missing data themselves (e.g., Collins et al., 2001; Graham, 2009; Schafer & Graham, 2002). When more information can be included from auxiliary variables, then missing values can be inferred from the observed data with greater accuracy (for a discussion about the use of auxiliary variables under FIML, see Enders, 2008; Graham, 2003). In addition, the quality of estimates and inferences obtained from MI can often be improved by generating a larger number of imputations (Bodner, 2008; Graham et al., 2007). In our experience, generating 20 imputations is sufficient for most applications in which the primary goal is to estimate the model parameters, but as many as 100 or more imputations can be useful if the analyses involve testing more elaborate statistical hypotheses (Bodner, 2008; see also Grund, Lüdtke, & Robitzsch, 2016b).

Discussion

In the present article, we outlined several procedures for MI of multilevel missing data, each intended to accommodate typical research questions in organizational psychology and other areas in the social sciences. Through several smaller simulation studies, we tried to provide a broad overview of multilevel MI. We demonstrated that the current implementations of multilevel MI are able to accommodate quite general research questions and multilevel designs. For example, several procedures for multilevel MI, using either the JM or the FCS approach, were suitable in the broad context of random intercept models. In such a context, missing data can be treated fairly accurately and in a very general manner even when missing data occur at different levels of the sample or in categorical and continuous variables simultaneously.

However, we also pointed out applications in which the current implementations of multilevel

MI do not correctly accommodate the model of interest. Specifically, it is still challenging to implement multilevel MI for multilevel models with random slopes or interaction effects when the explanatory variables contain missing data (see also S. Kim et al., 2015). Even though multilevel FCS appears to be slightly more flexible than multilevel JM in accommodating the substantive model, both approaches ultimately contain limitations due to the ways in which they are currently implemented in statistical software. To alleviate this problem, it has been recommended that the substantive analysis model be taken into account when conducting MI, thus ensuring that imputations are generated in a manner consistent with the model of interest (Bartlett et al., 2015). With this procedure, Bartlett et al. (2015) demonstrated that the bias associated with nonlinear and interaction effects in single-level regression models can be greatly reduced (see also Goldstein et al., 2014). Unfortunately, this approach is currently not available in standard software for multilevel MI.

As an alternative to MI, multilevel models can be estimated directly from the incomplete data by applying model-based procedures such as FIML (e.g., in M*plus*). Even though current implementations of FIML are still quite general and easy to use, it can be challenging to estimate multilevel models with missing values in explanatory variables, for example, when the model of interest uses observed group means to incorporate group-level effects or it includes categorical variables, random slopes, or interaction effects (see also Shin & Raudenbush, 2010). The challenges of FIML are ultimately similar to those of MI, and similar proposals have been made with respect to how one might overcome these challenges. For example, Stubbendick and Ibrahim (2003) proposed a factorization approach to FIML estimation of multilevel models with missing data in explanatory variables (see also L. Wu, 2010). Unfortunately, this approach is also currently not available in standard software. As an alternative, the model-based treatment of missing data can be implemented in a Bayesian analysis approach (Erler et al., 2016; see also Goldstein et al., 2014; Zhang & Wang, 2016). However, the Bayesian approach requires specialized software for Bayesian analyses such as WinBUGS (Lunn et al., 2000) or JAGS (Plummer, 2016), and such software can be challenging to use in practice (e.g., syntax-based model specification, selection of priors and starting values). Our own experiences indicate that these procedures can provide unbiased estimates with good coverage properties even in

multilevel models with random slopes and CLIs. For interested readers, we provide an example of a model-based procedure in the supplemental online materials. This example includes a multilevel model with random slopes and cross- and group-level interactions with missing data in explanatory variables (i.e., the conditions simulated in Study 2). The model syntax for the JAGS software is provided. However, before they can be widely adopted, we recommend that these procedures be subjected to further research and implemented in standard software. Additional software packages that implement FIML for multilevel models are xxM (Mehta, 2013) and Latent GOLD (Vermunt & Magidson, 2013).

Despite limitations in complex multilevel analyses, multilevel MI provides a more reliable and efficient approach to the treatment of missing data in comparison with simpler methods (e.g., single-level MI). As an alternative, it has been suggested that the multilevel structure might be expressed by including dummy-indicator variables in a single-level imputation model (Drechsler, 2015). Although this strategy substantially increases the complexity of the imputation model when the model of interest includes random slopes or interaction effects, it may be interesting to investigate its performance more thoroughly under such conditions (see also Andridge, 2011; Enders et al., 2016). In the context of multilevel models with random slopes, it has also been recommended that single-level MI be performed separately within each group (Graham, 2009). However, this strategy has been shown to be inefficient (i.e., low power) and should be avoided (Taljaard et al., 2008)

Every simulation study has its limitations, and owing to their smaller frame, the simulation studies presented here are no exception. In each study, we focused on varying the sample sizes rather than creating a diverse pattern of possible effects and effect sizes. However, this came at the price of choosing constant values for many of the population parameters. Therefore, the results should not be generalized to arbitrary patterns of effects. On the other hand, it is nearly impossible to address the diversity of possible research designs in a single study. Future studies should investigate the performance of multilevel MI in more specialized applications, including settings with very small samples at the individual level (e.g., dyadic data) or the group level (e.g., research in large organizations), a larger variety of patterns of effects and missing data mechanisms (see Newman, 2009), low ICCs, or a large number of continuous and categorical

variables (see Vermunt, 2003; Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008). Further topics for future research also include the application of multilevel MI in longitudinal data, which share many but not not all of the features of cross-sectional data, and in models with additional levels of hierarchy (see Yucel, 2008). In principle, however, these models can be addressed with existing statistical software.

Summing up, we believe that MI is already a powerful tool for treating missing data in multilevel research. Several procedures that make MI both generally applicable and easy to use have become available. In the present article, we attempted to provide guidance on the application of multilevel MI in research practice by providing both simulation results and recommendations for different applications of multilevel models. Our findings suggest natural directions for future research. For example, even though multilevel MI yielded reliable results in most applications, this was not the case in multilevel models with random slopes or interaction effects when data were missing in explanatory variables. Several procedures that might alleviate these problems have been proposed, but before these procedures can widely be adopted in practice, they must be evaluated more thoroughly in the context of multilevel designs, and they must be implemented in standard software. In this spirit, we hope that the present study and the materials provided with it will stimulate further research in this area and contribute to the regular use of MI in research practice.

Appendix A: Simulation design

In all simulation studies, the data were generated on the basis of the following model

$$Y_{ij} = \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + \gamma_{02}W_j + \gamma_{11}W_j(X_{ij} - \bar{X}_{\bullet j}) + u_{0j} + u_{1j}(X_{ij} - \bar{X}_{\bullet j}) + e_{ij}\,, \quad \text{(A1)}$$

where u_{0j} and u_{1j} followed a multivariate normal distribution with mean zero and covariance matrix $\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$, and e_{ij} was normally distributed with mean zero and variance σ^2. All variables were assumed to be standardized with mean zero and unit variance.

The data were generated in several steps. First, we simulated the between- and within-group components of X, which followed a normal distribution with mean zero and variances τ_x^2 and σ_x^2, respectively, which were determined by fixing the intraclass correlation (ICC) of X ($\rho_{I,x}$).

Then, we applied group-mean centering in order to obtain $\bar{X}_{\bullet j}$ and $(X_{ij} - \bar{X}_{\bullet j})$. In the next step, we generated W_j from a linear regression on $\bar{X}_{\bullet j}$ according to a correlation coefficient $\rho_{\bar{x}w}$. The regression coefficient was given by $\rho_{\bar{x}w}/\sqrt{\tau_x^2 + \sigma_x^2/n}$ and the residual variance by $(1 - \rho_{\bar{x}w}^2)$. Finally, Y was simulated from Equation A1. The residual variance was set to

$$\sigma^2 = 1 - \gamma_{01}^2\left(\tau_x^2 + \frac{\sigma_x^2}{n}\right) - \gamma_{02}^2 - 2\gamma_{01}\gamma_{02}\rho_{\bar{x}w}\sqrt{\tau_x^2 + \frac{\sigma_x^2}{n}} - \left(\gamma_{10}^2 + \gamma_{11}^2 + \tau_1^2\right)\left(\sigma_x^2 - \frac{\sigma_x^2}{n}\right) - \tau_0^2 \,. \quad \text{(A2)}$$

The parameters of the simulations were chosen in such a way as to imply certain values for the ICC of Y (see also Aguinis & Culpepper, 2015, for a very general decomposition of variance). For this purpose, the ICC was calculated as follows

$$\rho_{I,y} = \gamma_{01}^2\left(\tau_x^2 + \frac{\sigma_x^2}{n}\right) + \gamma_{02}^2 + 2\gamma_{02}\left(\gamma_{01} - \gamma_{01}\frac{\sigma_x^2/n}{\tau_x^2 + \sigma_x^2/n}\right)\rho_{\bar{x}w}\sqrt{\tau_x^2 + \frac{\sigma_x^2}{n}} - (\gamma_{10}^2 + \gamma_{11}^2 + \tau_1^2)\frac{\sigma_x^2}{n} + \tau_0^2 \,. \quad \text{(A3)}$$

Missing data were simulated according to a missing data mechanism $T \sim P$, where T is a target and P is a predictor variable. For person i in group j, we simulated a latent response propensity

$$R_{ij} = \lambda_0 + \lambda_1 P_{ij} + r_{ij} \quad \text{(A4)}$$

where R denotes T's response propensity, λ_0 is a quantile of the standard normal distribution according to a missing data probability, and λ_1 is the weight with which P drives the missing data mechanism. The residuals r_{ij} were drawn from a normal distribution with mean zero and variance $1 - \lambda_1^2$. A value T_{ij} was deleted if $R_{ij} > 0$. To generate MCAR and MAR data, we set $\lambda_1 = 0$ and .50, respectively. In Study 3b, the missing data mechanism was based on the group means.

In Study 3a, which included a dichotomous variable D, we generated the data as follows: First, we simulated within- and between-group components for a latent background variable D^* according to the ICC of D ($\rho_{I,d}$). We set D_{ij} to one if $D_{ij}^* > 0$ and zero otherwise. Then, X was simulated conditional on D, and Y was simulated conditional on D and X. Because the within- and between-group components of D could not be partitioned in the same way as for continuous variables, we generated X and Y by specifying residual ICCs for both variables ($\rho_{I,x|d}$ and $\rho_{I,y|dx}$). These were chosen in such a way that they implied marginal ICCs for both variables that were similar to those in Studies 1, 2, and 3b.

Appendix B: Step-by-step examples for multilevel MI

Here, we provide examples for multilevel MI in the statistical software R, using the packages `jomo` (JM), `mice` (FCS), and `miceadds` (FCS). We make use of the `justice` data set included in the `mitml` package, which contains simulated data from 1,400 employees organized in 200 organizations (`id`) denoting employees' sex (`sex`), their organizational satisfaction (`sat`), their orientation toward procedural justice (`jor`), and scores for justice climate at the level of organizations (`jcl`).

```
library(jomo)       # load "jomo" multilevel JM
library(mice)       # load "mice" for multilevel FCS
library(miceadds)   # load "miceadds" for multilevel FCS

library(mitml)                  # load "mitml"
data(justice, package="mitml")  # load "justice" data set
```

The structure of the data set is as follows.

```
# id    sex   sat    jor    jcl
#  1 female 4.783  0.886 -0.336
#  1   male 5.392  0.205 -0.336
#  1   male 6.696  0.328 -0.336
```

All variables contain missing data. For each of the following examples, we focus on one model of interest. For the purpose of illustration, we use a subset of the data in each example in which missing data occur *only* in the variables included in the model of interest. In practice, these examples can (and should) be combined to treat missing data for all variables simultaneously.

Example 1.1: Random intercept model

In this example, we demonstrate multilevel MI in a multilevel random intercept model with only continuous variables. Assume a researcher is interested in estimating the following model

$$SAT_{ij} = \gamma_{00} + \gamma_{10}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{01}\overline{JOR}_{\bullet j} + u_{0j} + e_{ij} \ . \tag{B1}$$

Either the JM or the FCS approach can be used in this case. In order to set aside the treatment of missing data in categorical and group-level variables (see Examples 1.2 and 1.3), the data set is reduced in such a way that only the variables in the model of interest require imputation.

```
ex1.1 <- subset(justice, !is.na(sex) & !is.na(jcl))
```

Option 1: multilevel FCS. In the FCS approach, it is necessary to specify the imputation model for each variable as well as the predictor variables for each model. Using `mice`, the imputation model is specified by defining an imputation method and predictor variables for each target variable with missing data. The imputation methods are specified in a character vector (`impMethod`). Here, we assign a two-level model to each variable with missing data.

```
# set up imputation methods
impMethod <- character(ncol(ex1.1))   # create empty vector for
names(impMethod) <- colnames(ex1.1)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"   # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"   # use '2l.pan' for 'jor' (two-level normal)
```

The predictors in each model are then defined in the predictor matrix (`predMatrix`), where each line corresponds to one target variable, and integer values denote the relations between the target and predictor variables.

```
# set up predictor matrix
predMatrix <- matrix(0, ncol(ex1.1), ncol(ex1.1))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex1.1)  # matrix

# ... define predictors for each variable
predMatrix[ "sat" , c("id","sex","jor","jcl") ] <- c(-2,3,3,1)   # -2 = cluster variable
predMatrix[ "jor" , c("id","sex","sat","jcl") ] <- c(-2,3,3,1)   # 1 = overall effect
                                                                 # 3 = overall + group-level effect
```

The imputation is then carried out by issuing the following command.

```
imp <- mice::mice(data=ex1.1, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix)
```

Option 2: multilevel JM. Because the model of interest is a multilevel random intercept model, all variables (including auxiliary variables) can be treated as target variables in the JM approach. In the `mitml` package, imputation models can be specified as a list of two model formulas, pertaining to individual- and group-level variables, respectively (for an alternative model specification similar to FCS, see the package documentation).

```
fml <- list( sat + jor + sex ~ 1 + (1|id),   # level-1 targets = sat + jor + sex (no predictors)
             jcl ~ 1 )                       # level-2 targets = jcl (no predictors)
```

The imputation is then carried out by issuing the following command.

```
imp <- mitml::jomoImpute(data=ex1.1, formula=fml, n.burn=5000, n.iter=500, m=20)
```

Example 1.2: Random intercept model with categorical variables

In this example, we demonstrate multilevel MI with a mix of continuous and categorical variables. Assume a researcher us interested in the following model

$$SAT_{ij} = \gamma_{00} + \gamma_{10} JOR_{ij} + \gamma_{20} Sex_{ij} + u_{0j} + e_{ij} \ . \tag{B2}$$

Either the JM or the FCS approach can be used in this case. As before, the data set is reduced in such a way that only these variables require imputation.

```
ex1.2 <- subset(justice, !is.na(jcl))
```

Option 1: multilevel FCS. The specification of the FCS approach is similar to the previous example but now includes a logistic two-level model for the categorical target variable (`sex`).

```
# set up imputation methods
impMethod <- character(ncol(ex1.2))   # create empty vector for
names(impMethod) <- colnames(ex1.2)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"      # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"      # use '2l.pan' for 'jor' (two-level normal)
impMethod[ "sex" ] <- "2l.binary"   # use '2l.binary' for 'sex' (two-level logistic)

# set up predictor matrix
predMatrix <- matrix(0, ncol(ex1.2), ncol(ex1.2))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex1.2)   # matrix

# ... define predictors for each variable
predMatrix[ "sat" , c("id","sex","jor","jcl") ] <- c(-2,3,3,1)   # -2 = cluster variable
predMatrix[ "jor" , c("id","sex","sat","jcl") ] <- c(-2,3,3,1)   # 1 = overall effect
predMatrix[ "sex" , c("id","sat","jor","jcl") ] <- c(-2,3,3,1)   # 3 = overall + group-level effect
```

The imputation is then run as follows. Note that categorical variables must be converted into a numeric data type beforehand in order to be imputed using `mice`.

```
# convert "sex" to dummy variable
ex1.2 <- within(ex1.2, sex <- as.integer(sex)-1)

imp <- mice::mice(data=ex1.2, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix)
```

Option 2: multilevel JM. The specification of the JM approach is identical to the previous example, treating all variables (including auxiliary variables) as target variables in the imputation model.

```
fml <- list( sat + jor + sex ~ 1 + (1|id),   # level-1 targets = sat + jor + sex (no predictors)
             jcl ~ 1 )                       # level-2 targets = jcl (no predictors)
```

The imputation is then carried out by issuing the following command. Not that, in contrast to `mice`, `jomo` requires that categorical variables are formatted as factor variables in R, as is the case in the original data set.

```
imp <- mitml::jomoImpute(data=ex1.2, formula=fml, n.burn=5000, n.iter=500, m=20)
```

Example 1.3: Random intercept model with group-level variables

In this example, we demonstrate multilevel MI with explanatory variables at both the individual and the group level. Assume a researcher is interested in the following model

$$SAT_{ij} = \gamma_{00} + \gamma_{10}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{01}\overline{JOR}_{\bullet j} + \gamma_{02}JCL_j + u_{0j} + e_{ij} \; . \tag{B3}$$

Either the JM or the FCS approach can be used in this case. As before, the data set is reduced in such a way that only these variables require imputation.

```
ex1.3 <- subset(justice, !is.na(sex))
```

Option 1: multilevel FCS. The specification of the FCS approach is similar to the previous example but now includes a group-level regression model for the target variable at the group level (`jcl`).

```
# set up imputation methods
impMethod <- character(ncol(ex1.3))   # create empty vector for
names(impMethod) <- colnames(ex1.3)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"        # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"        # use '2l.pan' for 'jor' (two-level normal)
impMethod[ "jcl" ] <- "2lonly.norm"   # use '2lonly.norm' for 'jcl' (group-level regression)

# set up predictor matrix
predMatrix <- matrix(0, ncol(ex1.3), ncol(ex1.3))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex1.3)   # matrix

# ... define predictors for each variable
predMatrix[ "sat" , c("id","sex","jor","jcl") ] <- c(-2,3,3,1)   # -2 = cluster variable
predMatrix[ "jor" , c("id","sex","sat","jcl") ] <- c(-2,3,3,1)   # 1 = overall effect
predMatrix[ "jcl" , c("id","sex","sat","jor") ] <- c(-2,1,1,1)   # 3 = overall + group-level effect
```

The imputation is then carried out by issuing the following command.

```
imp <- mice::mice(data=ex1.3, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix)
```

Option 2: multilevel JM. The specification of the JM approach is identical to the previous example, treating all variables (including auxiliary variables) as target variables in the imputation model.

```
fml <- list( sat + jor + sex ~ 1 + (1|id),   # level-1 targets = sat + jor + sex (no predictors)
             jcl ~ 1 )                       # level-2 targets = jcl (no predictors)
```

The imputation is then carried out by issuing the following command.

```
imp <- mitml::jomoImpute(data=ex1.3, formula=fml, n.burn=5000, n.iter=500, m=20)
```

Example 1.4: Random intercept model with interaction effects

In this example, we demonstrate multilevel MI for a multilevel random intercept model with two explanatory variables and an interaction effect. Assume a researcher is interested in the following model

$$\begin{aligned} SAT_{ij} =& \gamma_{00} + \gamma_{10}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{20} Sex_{ij} + \gamma_{30} Sex_{ij}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{01}\overline{JOR}_{\bullet j} + \\ & u_{0j} + e_{ij} \ . \end{aligned} \tag{B4}$$

In order to include the interaction effect in the imputation model, we make use of passive imputation steps in the FCS approach. As before, the data set is reduced in such a way that only these variables require imputation.

```
ex1.4 <- subset(justice, !is.na(jcl))
```

To define passive steps, new variables must be created for the components of organizational satisfaction (`sat`) and justice orientation (`jor`) and for the product terms that allow relations between individual-level organizational satisfaction (`sat`) and justice orientation (`jor`) to vary by sex (`sex`) and vice versa.

```
ex1.4 <- within(ex1.4,{
  sex <- as.integer(sex)-1   # recode 'sex' as a dummy variable
  sat.GRP <- sat.IND <- NA   # passive variables for components of 'sat'
  jor.GRP <- jor.IND <- NA   # passive variables for components of 'jor'
  sat.IND_sex <- NA          # passive variable for interaction of 'sat' and 'sex'
  jor.IND_sex <- NA          # passive variable for interaction of 'jor' and 'sex'
```

```
  sat.IND_jor.IND <- NA      # passive variable for interaction of 'sat' and 'jor'
})
```

The specification of the imputation methods proceeds in a similar manner as before. In addition, passive steps are defined for the components of organizational satisfaction (`sat`) and justice orientation (`jor`) as well as for the product terms by employing the `~I()` identity function.

```
# set up imputation methods
impMethod <- character(ncol(ex1.4))   # create empty vector for
names(impMethod) <- colnames(ex1.4)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"      # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"      # use '2l.pan' for 'jor' (two-level normal)
impMethod[ "sex" ] <- "2l.binary"   # use '2l.binary' for 'sex' (two-level logistic)

impMethod[ "sat.GRP" ] <- "2l.groupmean"      # passive step for updating the group means
impMethod[ "jor.GRP" ] <- "2l.groupmean"      # means and within-group deviations of 'sat'
impMethod[ "sat.IND" ] <- "~I(sat-sat.GRP)"   # and 'jor'
impMethod[ "jor.IND" ] <- "~I(jor-jor.GRP)"

impMethod[ "sat.IND_sex" ] <- "~I(sat.IND*sex)"           # passive steps for updating the
impMethod[ "jor.IND_sex" ] <- "~I(jor.IND*sex)"           # product terms involving  'sat',
impMethod[ "sat.IND_jor.IND" ] <- "~I(sat.IND*jor.IND)"   # 'jor', and 'sex'
```

The predictors for each model are specified as before. However, two additional entries that allow the group means to be updated are added.

```
# set up predictor matrix
predMatrix <- matrix(0, ncol(ex1.4), ncol(ex1.4))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex1.4)   # matrix

# ... define predictors for each variable
predMatrix[ "sat.GRP" , c("id","sat") ] <- c(-2,1)   # -2 = cluster variable
predMatrix[ "jor.GRP" , c("id","jor") ] <- c(-2,1)   # 1 = variable to be aggregated

predMatrix[ "sat" , c("id","sex","jor","jcl","jor.IND_sex") ] <- c(-2,1,3,1,1)
predMatrix[ "jor" , c("id","sex","sat","jcl","sat.IND_sex") ] <- c(-2,1,3,1,1)
predMatrix[ "sex" , c("id","sat","jor","jcl","sat.IND_jor.IND") ] <- c(-2,3,3,1,1)
```

Finally, the imputation is carried out by issuing the following command. In addition, we specify a visit sequence (`visitSeq`), which defines the order in which the variables are to be imputed and the passive steps are to be carried out. Passive steps are updated after obtaining a new imputation to the extents to which they are needed in subsequent imputation steps.

```
# set up visit sequence
visitSeq <- c(
  "sat","sat.GRP","sat.IND","sat.IND_sex",      # impute 'sat', then update terms for 'jor'
  "jor","jor.GRP","jor.IND","sat.IND_jor.IND",  # impute 'jor', then update terms for 'sex'
```

```
  "sex","jor.IND_sex"                          # impute 'sex', then update terms for 'sat'
  )
visitSeq <- match(visitSeq, colnames(ex1.4))

imp <- mice::mice(data=ex1.4, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix, allow.na=TRUE)
```

Example 2.1: Random slope model

In this example, we demonstrate multilevel MI for a multilevel model with random slopes. Assume a researcher is interested in the following model

$$SAT_{ij} = \gamma_{00} + \gamma_{10}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{01}\overline{JOR}_{\bullet j} + u_{0j} + u_{1j}(JOR_{ij} - \overline{JOR}_{\bullet j}) + e_{ij} \ . \qquad \text{(B5)}$$

To accommodate the fact that the individual-level relation of organizational satisfaction (`sat`) and justice orientation (`jor`) varies across groups, the FCS approach will be used. As before, the data set is reduced in such a way that only the variables in the model of interest require imputation.

```
ex2.1 <- subset(justice, !is.na(sex) & !is.na(jcl))
```

The specification of the FCS approach is very similar to Example 1.1 but now includes a random slope in the imputation model for both organizational satisfaction (`sat`) and justice orientation (`jor`), which is denoted by the number 4 in the predictor matrix.

```
# set up imputation methods
impMethod <- character(ncol(ex2.1))   # create empty vector for
names(impMethod) <- colnames(ex2.1)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"   # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"   # use '2l.pan' for 'jor' (two-level normal)

# set up predictor matrix
predMatrix <- matrix(0, ncol(ex2.1), ncol(ex2.1))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex2.1)   # matrix

# ... define predictors for each variable
predMatrix[ "sat" , c("id","sex","jor","jcl") ] <- c(-2,3,4,1)   # 4 = individual-level (random)
predMatrix[ "jor" , c("id","sex","sat","jcl") ] <- c(-2,3,4,1)   # and group-level (fixed) effect
                                                                 # ... otherwise as above
```

The imputation is then carried out by issuing the following command.

```
imp <- mice::mice(data=ex2.1, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix)
```

Example 2.2: Random slope model with interaction effects

In this example, we demonstrate multilevel MI with two explanatory variables, one at the individual and one at the group level, including a random slope and interaction effects. Assume a researcher is interested in the following model

$$\begin{aligned} SAT_{ij} = \gamma_{00} + \gamma_{10}(JOR_{ij} - \overline{JOR}_{\bullet j}) + \gamma_{01}\overline{JOR}_{\bullet j} + \gamma_{02}JCL_j + \gamma_{11}(JOR_{ij} - \overline{JOR}_{\bullet j})\,JCL_j + \\ \gamma_{03}\overline{JOR}_{\bullet j}JCL_j + u_{0j} + u_{1j}(JOR_{ij} - \overline{JOR}_{\bullet j}) + e_{ij}\ . \end{aligned} \tag{B6}$$

In order to include the random slopes and the interaction effects in the imputation model, we make use of passive imputation steps in the FCS approach. As before, the data set is reduced in such a way that only these variables require imputation.

```
ex2.2 <- subset(justice, !is.na(sex))
```

To define passive steps, new variables must be created for the components of organizational satisfaction (`sat`) and justice orientation (`jor`) as well as for the product terms at the individual and the group level.

```
ex2.2 <- within(ex2.2,{
  sat.GRP <- sat.IND <- NA   # passive variables for components of 'sat'
  jor.GRP <- jor.IND <- NA   # ... for components of 'jor'
  sat.IND_jcl <- NA          # ... for individual-level interaction of 'sat' and 'jcl'
  sat.GRP_jcl <- NA          # ... for group-level interaction of 'sat' and 'jcl'
  jor.IND_jcl <- NA          # ... for individual-level interaction of 'jor' and 'jcl'
  jor.GRP_jcl <- NA          # ... for group-level interaction of 'jor' and 'jcl'
  sat.GRP_jor.GRP <- NA      # ... for group-level interaction of 'sat' and 'jor'
})
```

The specification of the imputation methods proceeds as before. Passive steps are defined for the components of organizational satisfaction (`sat`) and justice orientation (`jor`) as well as for the product terms by imploying `~I()`.

```
# set up imputation methods
impMethod <- character(ncol(ex2.2))   # create empty vector for
names(impMethod) <- colnames(ex2.2)   # imputation methods

# ... define method for each variable
impMethod[ "sat" ] <- "2l.pan"       # use '2l.pan' for 'sat' (two-level normal)
impMethod[ "jor" ] <- "2l.pan"       # use '2l.pan' for 'jor' (two-level normal)
impMethod[ "jcl" ] <- "2lonly.norm"  # use '2lonly.norm' for 'jcl' (group-level regression)

impMethod[ "sat.GRP" ] <- "2l.groupmean"      # passive step for updating the group means
impMethod[ "jor.GRP" ] <- "2l.groupmean"      # means and within-group deviations of 'sat'
impMethod[ "sat.IND" ] <- "~I(sat-sat.GRP)"   # and 'jor'
```

```
impMethod[ "jor.IND" ] <- "~I(jor-jor.GRP)"

impMethod[ "sat.IND_jcl" ] <- "~I(sat.IND*jcl)"          # passive steps for updating the
impMethod[ "jor.IND_jcl" ] <- "~I(jor.IND*jcl)"          # individual-level product terms

impMethod[ "sat.GRP_jcl" ] <- "~I(sat.GRP*jcl)"          # passive steps for updating
impMethod[ "jor.GRP_jcl" ] <- "~I(jor.GRP*jcl)"          # the group-level product terms
impMethod[ "sat.GRP_jor.GRP" ] <- "~I(sat.GRP*jor.GRP)"
```

The predictors in each model are specified as before.

```
# set up predictor matrix
predMatrix <- matrix(0, ncol(ex2.2), ncol(ex2.2))                # create empty predictor
rownames(predMatrix) <- colnames(predMatrix) <- colnames(ex2.2)   # matrix

# ... define predictors for each variable
predMatrix[ "sat.GRP" , c("id","sat") ] <- c(-2,1)   # -2 = cluster variable
predMatrix[ "jor.GRP" , c("id","jor") ] <- c(-2,1)   # 1 = variable to be aggregated

predMatrix[ "sat" , c("id","sex","jor","jcl","jor.IND_jcl","jor.GRP_jcl") ] <- c(-2,3,3,1,1,1)
predMatrix[ "jor" , c("id","sex","sat","jcl","sat.IND_jcl","sat.GRP_jcl") ] <- c(-2,3,3,1,1,1)
predMatrix[ "jcl" , c("id","sex","sat","jor","sat.GRP_jor.GRP") ] <- c(-2,1,1,1,1)
```

The imputation is carried out by issuing the following command. The visit sequence (`visitSeq`) is specified in such a way that passive steps are updated after obtaining a new imputation to the extents to which they are needed in subsequent imputation steps.

```
# set up visit sequence
visitSeq <- c(
  "sat","sat.GRP","sat.IND","sat.IND_jcl","sat.GRP_jcl", # impute 'sat', then update terms for 'jor'
  "jor","jor.GRP","jor.IND","sat.GRP_jor.GRP",           # impute 'jor', then update terms for 'jcl'
  "jcl","jor.IND_jcl","jor.GRP_jcl"                      # impute 'jcl', then update terms for 'sat'
)
visitSeq <- match(visitSeq, colnames(ex2.2))

imp <- mice::mice(data=ex2.2, maxit=10, m=20, imputationMethod=impMethod,
                  predictorMatrix=predMatrix, allow.na=TRUE)
```

Article 3: Multiple imputation of missing data at Level 2: A comparison of fully conditional and joint modeling in multilevel designs

Grund, S., Lüdtke, O., & Robitzsch, A. (2017b). *Multiple imputation of missing data at* Level 2*: A comparison of fully conditional and joint modeling in multilevel designs*. Manuscript submitted for publication.

Multiple imputation (MI) can be used to address missing data at Level 2 in multilevel research. In this article, we compare joint modeling (JM) and the fully conditional specification (FCS) of MI as well as different strategies for including auxiliary variables at Level 1 using either their manifest or latent cluster means. We show with theoretical arguments and computer simulations that (a) an FCS approach that uses latent cluster means is comparable to JM, and (b) using manifest cluster means provides similar results except in relatively extreme cases with unbalanced data. We outline a computational procedure for including latent cluster means in an FCS approach using plausible values and provide an example using data from the PISA 2012 study.

Multiple imputation (MI) of missing data has received considerable attention in the methodological and applied missing data literature (e.g., Allison, 2001; Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002). However, many open questions remain when the data have a multilevel structure (e.g., when students are clustered within schools; for recent reviews, see Enders et al., 2016; Hox et al., 2016). Most studies to date have focused on missing data that occur at Level 1 (e.g., when students do not answer all items on a questionnaire). These studies have shown that the multilevel structure must be taken into account during MI because ignoring the multilevel structure in the imputation model may lead to biased estimates in subsequent analyses (Andridge, 2011; Black et al., 2011; Drechsler, 2015; Enders et al., 2016; Lüdtke et al., 2017; Taljaard et al., 2008; for a more general discussion, see Carpenter & Kenward, 2013; Meng, 1994).

Much less attention has been paid to missing data at Level 2, even though the treatment of missing data at Level 2 can be of great practical importance when the model of interest includes variables at both Level 1 and 2. For example, in a study of teacher effects on student achievement, a whole class of students would have to be dropped from the analysis if a certain teacher's data are missing. Currently, the methodological literature provides little guidance about how to carry out MI when data are missing at Level 2 (see also van Buuren, 2011). In one of the first studies to consider this topic, Gibson and Olejnik (2003) applied single-level MI to a

subset of the data that included only variables at Level 2 but ignored the contribution of variables at Level 1. Later, Cheung (2007) applied single-level MI to the data set as a whole (also known as "flat-file" imputation; see also van Buuren 2011), thus including variables at both levels but ignoring the multilevel structure. In contrast to most of the missing data literature, these studies concluded that "the performance of MI was (...) poorest among all of the methods that were studied" (Cheung 2007, p. 625; see also Gibson and Olejnik 2003, p. 233). This illustrates that the performance of MI depends on the specification of the imputation model; if the model does not reflect the characteristics of the data or the intended analysis, then using MI may even be harmful. In recent years, however, more advanced methods that specifically take into account the multilevel structure of the data as well as missing data at different levels of analysis have been developed for MI (e.g., Asparouhov & Muthén, 2010b; Carpenter & Kenward, 2013; Goldstein et al., 2009).

The present article pursues three different goals. First, we compare two popular approaches for MI of missing data at Level 2, joint modeling (JM) and the fully conditional specification (FCS) of MI, as well as two popular ad hoc procedures, single-level MI and listwise deletion (LD; see also Enders et al. 2016). Second, we discuss different strategies for including variables at Level 1 when specifying the imputation model for missing data at Level 2. More precisely, we evaluate the consequences of including the manifest or latent cluster means of variables at Level 1 as auxiliary variables (i.e., covariates) in the imputation model at Level 2 (see also Asparouhov & Muthén, 2006; Lüdtke et al., 2008). In this context, we present a procedure for including latent cluster means in the FCS paradigm using the method of plausible values (Mislevy, 1991). In two simulation studies, we investigate the performance of each of these approaches in various conditions, including applications with small samples and unbalanced data, and the role of Level 1 variables when treating missing data at Level 2. Finally, we provide an empirical example using data from the Programme for International Student Assessment (PISA; OECD, 2014) and conclude with a discussion of our findings.

Cluster-level components in multilevel data

In two-level data with observations (e.g., students) nested within clusters (e.g., school classes),

variables can be measured directly at Level 1 (e.g., student self-concept) or Level 2 (e.g., class size, teacher qualification). In addition, variables at Level 1 can be decomposed into one part that varies only within clusters (within-cluster component), and a second part that varies only between clusters (cluster-level component), the latter of which can be used to estimate cluster-level effects of Level 1 variables (e.g., Cronbach, 1976; Preacher et al., 2010). In the following, we identify two ways of including the cluster-level component of predictor variables at Level 1 in multilevel models. In the first approach, the cluster mean of the Level 1 variable is calculated and included as a manifest predictor variable. However, the methodological literature has pointed out that the observed cluster mean is sometimes not a reliable measure of the unobserved, true cluster mean (e.g., Croon & van Veldhoven, 2007; Shin & Raudenbush, 2010). Thus, in the second approach, the cluster-level component of the Level 1 variable is treated as a latent variable, correcting for the unreliability that comes from estimating cluster means with only a finite number of observations (Lüdtke et al., 2008). In the following, we provide a more formal comparison of the two approaches.

Consider a set of variables $(\mathbf{x}_{ij}, \mathbf{z}_j)$, where P variables $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijP})$ are recorded at Level 1, and Q variables $\mathbf{z}_j = (z_{j1}, \ldots, z_{jQ})$ are recorded at Level 2. Using manifest or latent cluster means, the values $\mathbf{x}_{ij}$ for an observation i in cluster j can be expressed as

$$\begin{aligned} \mathbf{x}_{ij} &= \bar{\mathbf{x}}_{\bullet j} + (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\bullet j}) \quad \text{(manifest)} \\ \mathbf{x}_{ij} &= \mathbf{u}_j + \mathbf{e}_{ij}\,, \qquad\qquad \text{(latent)} \end{aligned} \tag{1}$$

where $\bar{\mathbf{x}}_{\bullet j}$ denotes the *manifest* cluster mean, $\mathbf{u}_j$ denotes the *latent* component at Level 2, and $\mathbf{u}_j$ and $\mathbf{e}_{ij}$ are independent and distributed normally with mean zero and covariance matrices $\mathbf{T}$ and $\mathbf{\Sigma}$. Consequently, assuming latent cluster means, the joint distribution of $\mathbf{x}_{ij}$ and $\mathbf{z}_j$ can be expressed as

$$Var(\mathbf{x}_{ij}, \mathbf{z}_j) = \begin{pmatrix} \mathbf{T} + \mathbf{\Sigma} & \boldsymbol{\sigma}^T \\ \boldsymbol{\sigma} & \mathbf{\Phi} \end{pmatrix}, \tag{2}$$

where $\mathbf{\Phi}$ is the covariance matrix of $\mathbf{z}_j$, and $\boldsymbol{\sigma}$ denotes the covariance of $\mathbf{x}_{ij}$ with $\mathbf{z}_j$. The manifest and latent cluster means express the joint structure of $\mathbf{x}_{ij}$ and $\mathbf{z}_j$ in slightly different ways, which becomes clear when noting that $\bar{\mathbf{x}}_{\bullet j} = \mathbf{u}_j + \bar{\mathbf{e}}_{\bullet j}$. Although the covariances between variables at Level 1 and 2 are equivalent in complete data, $Cov(\bar{\mathbf{x}}_{\bullet j}, \mathbf{z}_j) = Cov(\mathbf{u}_j, \mathbf{z}_j)$, the

manifest means tend to have a larger variance across clusters, $Var(\bar{\mathbf{x}}_{\bullet j}) = Var(\mathbf{u}_j) + Var(\bar{\mathbf{e}}_{\bullet j})$. This is particularly important in multilevel analyses because the manifest and latent cluster means can imply different correlation and regression coefficients at Level 2 (Croon & van Veldhoven, 2007; Grilli & Rampichini, 2011; Lüdtke et al., 2008).

Substantive analysis models

Consider the case with only one variable at Level 1 (y_{ij}) and one variable at Level 2 (z_j), where $y_{ij} = u_j + e_{ij}$ with latent cluster means u_j. In the following, we consider two analysis models that can be used to describe the relation between y_{ij} and z_j. In the first model, y_{ij} is an outcome variable at Level 1 that is predicted by z_j,

$$y_{ij} = \beta_0 + \beta_{yz} z_j + \upsilon_j + \epsilon_{ij} \,, \tag{3}$$

where β_{yz} denotes the regression coefficient of y_{ij} regressed on z_j (see Snijders & Bosker, 2012b). In the second model, reusing some of the same notation, z_j is an outcome variable at Level 2 that is predicted by the latent cluster means of y_{ij},

$$z_j = \beta_0 + \beta_{zy} u_j + \nu_j \,, \tag{4}$$

where β_{zy} denotes the regression coefficient of z_j regressed on y_{ij} (for a discussion, see Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Notice that the model in Equation 4 could also be estimated on the basis of the manifest cluster means (i.e., with $\bar{y}_{\bullet j}$ instead of u_j), yielding an alternative estimate of the regression coefficient of z_j on y_{ij}, say $\tilde{\beta}_{zy}$. In general, β_{zy} and $\tilde{\beta}_{zy}$ will not be identical unless either the clusters become large or the variance of y_{ij} at Level 1 becomes small in comparison with the variance at Level 2 (Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Specifically, if the u_j were known, the population values of the two regression coefficients could be expressed as follows. For balanced clusters of size n,

$$\beta_{zy} = Var(u_j)^{-1} Cov(u_j, z_j) = \mathbf{T}^{-1}\boldsymbol{\sigma} \quad \text{and} \quad \tilde{\beta}_{zy} = Var(\bar{y}_{\bullet j})^{-1} Cov(\bar{y}_{\bullet j}, z_j) = (\mathbf{T} + \tfrac{1}{n}\boldsymbol{\Sigma})^{-1}\boldsymbol{\sigma} \,. \tag{5}$$

The fact that the two regression coefficients differ is well known in the multilevel literature (e.g., Lüdtke et al., 2008; Preacher et al., 2010; Shin & Raudenbush, 2010). In the present article, we elaborate on the consequences of this finding for the treatment of missing data: When dealing

with missing data at Level 2, the manifest and latent cluster means offer two different ways of incorporating the cluster-level components of variables at Level 1 in the imputation model for missing data at Level 2.

Imputation models for missing data at Level 2

In the following section, we present two popular approaches to multilevel MI: joint modeling (JM) and the fully conditional specification of MI (FCS). In order to discuss how the two approaches take the cluster-level component of variables at Level 1 into account when dealing with missing data at Level 2, we also compare the main features of the computational algorithms underlying the two approaches (see also Enders et al., 2016). For the purpose of this article, we focus on applications with normally distributed variables and missing data at Level 2. However, either approach can be used to deal with missing data at both Level 1 and 2; nonnormal and categorical variables can also be addressed and will be considered in the Discussion section.

Joint modeling (JM)

The general idea of MI is to draw multiple replacements for the missing values from the conditional distribution of the missing data, given the observed data and a statistical model (the imputation model). In JM, a single imputation model is specified for all variables with missing data, and imputations are generated for all variables simultaneously (Carpenter & Kenward, 2013; Goldstein et al., 2009; see also Schafer & Yucel, 2002). To simplify[1] its presentation, we consider a variant of the JM that does not include predictor variables but instead treats all variables as target variables in the imputation procedure. This model can be written as

$$\begin{aligned} \mathbf{y}_{1ij} &= \boldsymbol{\mu}_1 + \mathbf{u}_{1j} + \mathbf{e}_{1ij} \\ \mathbf{y}_{2j} &= \boldsymbol{\mu}_2 + \mathbf{u}_{2j} \,, \end{aligned} \tag{6}$$

[1] In the general formulation of the JM, predictor variables can be included in the model if they do not contain any missing data (i.e., they are completely observed). The simplified model discussed here was chosen because (a) it facilitates the presentation and comparison of the computational procedures, (b) it allows for arbitrary patterns of missing data, and (c) it can be applied in any situation in which the analysis model is a multilevel random intercept model.

where $\mathbf{y}_{1ij}$ denotes a number of target variables at Level 1, taking on values for observation i in cluster j, with mean vector $\boldsymbol{\mu}_1$, random intercepts $\mathbf{u}_{1j}$ at Level 2, and residuals $\mathbf{e}_{1ij}$ at Level 1. Likewise, $\mathbf{y}_{2j}$ denotes target variables at Level 2, with mean vector $\boldsymbol{\mu}_2$ and residuals $\mathbf{u}_{2j}$ at Level 2. The random intercepts and residuals at Level 2 combined, $\mathbf{u}_j = (\mathbf{u}_{1j}, \mathbf{u}_{2j})$, are assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Psi}$. The residuals at Level 1, $\mathbf{e}_{1ij}$, are assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

To illustrate the computational procedure for sampling from the JM, we consider the case where there are J clusters ($j = 1, \ldots, J$) each with n_j observations ($i = 1, \ldots, n_j$), P completely observed variables at Level 1, and Q variables at Level 2 with arbitrary patterns of missing data (see also Carpenter & Kenward, 2013; Goldstein et al., 2009). Then, for each missing data pattern, $\mathbf{y}_{2j}$ can be decomposed into an observed and an unobserved part, $\mathbf{y}_{2j} = (\mathbf{y}_{2j}^{\text{obs}}, \mathbf{y}_{2j}^{\text{mis}})$. The goal of MI is to draw replacements $\mathbf{y}_{2j}^{\text{imp}}$ for the $\mathbf{y}_{2j}^{\text{mis}}$ on the basis of the observed data $\mathbf{y}_{1ij}$ and $\mathbf{y}_{2j}^{\text{obs}}$, and the parameters of the imputation model, $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. The covariance matrix at Level 2 is a $(P+Q) \times (P+Q)$ matrix which, for computational convenience, can be partitioned by reordering its rows and columns as $\begin{bmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_2 \end{bmatrix}$, with subscripts referring to variables at Level 1 and 2, or $\begin{bmatrix} \boldsymbol{\Psi}_j^{\text{obs}} & \boldsymbol{\Psi}_j^{\text{obs,mis}} \\ \boldsymbol{\Psi}_j^{\text{mis,obs}} & \boldsymbol{\Psi}_j^{\text{mis}} \end{bmatrix}$, with superscripts referring to observed and missing data for each cluster j. From a set of starting values and given appropriate prior distributions, the Gibbs sampler iterates along the following steps. At iteration t,

1. Draw $\mathbf{u}_{1j}^{(t+1)} \sim P(\mathbf{u}_{1j} | \mathbf{y}_{1ij}, \mathbf{u}_{2j}^{(t)}, \boldsymbol{\theta}^{(t)})$ from a multivariate normal distribution $N(\tilde{\mathbf{u}}_{1j}^{(t)}, \mathbf{U}_{1j}^{(t)})$, conditional on $\mathbf{u}_{2j}$, with mean and covariance matrix as follows.
   i) $\tilde{\mathbf{u}}_{1j}^{(t)} = (\mathbf{I}_P - \boldsymbol{\Lambda}_{1|2j}^{(t)})\boldsymbol{\mu}_{1|2j}^{(t)} + \frac{1}{n_j}\boldsymbol{\Lambda}_{1|2j}^{(t)} \sum_{i=1}^{n_j} (\mathbf{y}_{1ij} - \boldsymbol{\mu}_1^{(t)})$, where $\boldsymbol{\Lambda}_{1|2j}^{(t)} = \boldsymbol{\Psi}_{1|2}^{(t)} \left[\boldsymbol{\Psi}_{1|2}^{(t)} + \frac{1}{n_j}\boldsymbol{\Sigma}^{(t)} \right]^{-1}$ is the reliability of the conditional cluster mean of $\mathbf{y}_{1ij}$ given $\mathbf{y}_{2j}$, $\boldsymbol{\mu}_{1|2j}^{(t)} = \boldsymbol{\Psi}_{12}^{(t)} \left[\boldsymbol{\Psi}_2^{(t)} \right]^{-1} \mathbf{u}_{2j}^{(t)}$ is the expected value of $\mathbf{u}_{1j}$ given $\mathbf{u}_{2j}$, and $\boldsymbol{\Psi}_{1|2}^{(t)} = \boldsymbol{\Psi}_1^{(t)} - \boldsymbol{\Psi}_{12}^{(t)} \left[\boldsymbol{\Psi}_2^{(t)} \right]^{-1} \boldsymbol{\Psi}_{21}^{(t)}$ is the conditional variance of $\mathbf{u}_{1j}$ given $\mathbf{u}_{2j}$
   ii) $\mathbf{U}_{1j}^{(t)} = \frac{1}{n_j}\boldsymbol{\Lambda}_{1|2j}^{(t)}\boldsymbol{\Sigma}^{(t)}$, where $\boldsymbol{\Lambda}_{1|2j}^{(t)}$ is as defined above
2. Calculate the residuals $\mathbf{u}_{2j}^{\text{obs},(t+1)} = \mathbf{y}_{2j}^{\text{obs}} - \boldsymbol{\mu}_2^{(t)}$ for observed cases at Level 2 by subtraction.

3. Impute $\mathbf{u}_{2j}^{\text{imp},(t+1)} \sim P(\mathbf{u}_{2j}^{\text{mis}} | \mathbf{u}_{1j}^{(t+1)}, \mathbf{u}_{2j}^{\text{obs},(t+1)}, \boldsymbol{\theta}^{(t)})$ for the $\mathbf{y}_{2j}^{\text{mis}}$ by drawing from a multivariate normal distribution $N(\boldsymbol{\mu}_{2j}^{\text{mis|obs},(t)}, \boldsymbol{\Psi}_{j}^{\text{mis|obs},(t)})$, with mean and covariance matrix as follows.
   i) $\boldsymbol{\mu}_{2j}^{\text{mis|obs},(t)} = \boldsymbol{\Psi}_{j}^{\text{obs,mis},(t)} \left[\boldsymbol{\Psi}_{j}^{\text{obs},(t)}\right]^{-1} \mathbf{u}_{j}^{\text{obs},(t+1)}$, the expected value of $\mathbf{u}_{2j}^{\text{mis}}$ given $\mathbf{u}_{j}^{\text{obs}}$ with $\mathbf{u}_{j}^{\text{obs},(t+1)} = (\mathbf{u}_{1j}^{(t+1)}, \mathbf{u}_{2j}^{\text{obs},(t+1)})$
   ii) $\boldsymbol{\Psi}_{j}^{\text{mis|obs},(t)} = \boldsymbol{\Psi}_{j}^{\text{mis},(t)} - \boldsymbol{\Psi}^{\text{obs,mis},(t)} \left[\boldsymbol{\Psi}_{j}^{\text{obs},(t)}\right]^{-1} \boldsymbol{\Psi}_{j}^{\text{mis,obs},(t)}$, the conditional variance of $\mathbf{u}_{2j}^{\text{mis}}$ given $\mathbf{u}_{j}^{\text{obs}}$
4. Form $\mathbf{u}_{2j}^{(t+1)} = (\mathbf{u}_{2j}^{\text{obs},(t+1)}, \mathbf{u}_{2j}^{\text{imp},(t+1)})$, and calculate $\mathbf{y}_{2j}^{(t+1)} = \boldsymbol{\mu}_{2}^{(t)} + \mathbf{u}_{2j}^{(t+1)}$.
5. Draw $\boldsymbol{\theta}^{(t+1)} \sim P(\boldsymbol{\theta} | \mathbf{y}_{1ij}, \mathbf{y}_{2j}^{(t+1)}, \mathbf{u}_{1j}^{(t+1)}, \mathbf{u}_{2j}^{(t+1)})$, given appropriate priors, where $P(\cdot)$ is an inverse-Wishart distribution for $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$, and multivariate normal for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Two important steps in this procedure ensure that the relations between variables are taken into account when performing MI. First, the random effects $\mathbf{u}_{1j}$ of variables at Level 1 are drawn conditionally on the variables at Level 2 (Step 1). Second, the missing residuals at Level 2, $\mathbf{u}_{2j}$, are drawn conditionally on the random effects of $\mathbf{y}_{1ij}$ and the observed $\mathbf{y}_{2j}$ (Step 3). Here, it becomes clear that the JM uses the *latent* cluster means (i.e., random effects) of $\mathbf{y}_1$ to predict missing values in $\mathbf{y}_2$.[2] Formally, the expression in Step 1a can be seen as a shrinkage estimator for the cluster means of $\mathbf{y}_1$. Using this estimator, the latent means (i.e., random effects) are "pulled" away from the observed (i.e., manifest) means and toward the grand mean to an extent that is determined by the reliability of the cluster means (see also de Leeuw & Kreft, 1995; Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004).

Fully conditional specification (FCS)

As an alternative to JM, the joint distribution of the variables with missing data can be approximated by imputing one variable at a time using a sequence of univariate imputation models, where each model conditions on the other variables in the data set (or a subset of them). This procedure is known as the fully conditional specification of MI (FCS) but sometimes also re-

[2] In the more general formulation of the JM, which includes additional predictor variables on the right-hand side of the model, it is possible to include manifest cluster means as predictor variables as long as the respective variables are completely observed. This specification of the JM is conceptually similar to the FCS approach and will not be considered further (for a discussion, see Enders et al., 2016).

ferred to as "chained equations" or sequential MI (Raghunathan et al., 2001; Royston & White, 2011; van Buuren et al., 2006).

Let y_{1ijp} denote observation i in cluster j for the p-th variable at Level 1 ($p = 1, \ldots, P$), and let y_{2jq} denote the value of cluster j for the q-th variable at Level 2 ($q = 1, \ldots, Q$). Then, imputations for missing values in individual-level variables may be generated from a set of conditional distributions

$$y_{1ijp} \sim P(y_{1ijp}|\mathbf{y}_{1ij(-p)}, \tilde{\mathbf{y}}_{1j(-p)}, \mathbf{y}_{2j}, \boldsymbol{\theta}_p)\,, \tag{7}$$

where the subscript $(-p)$ denotes the set of variables from which p is excluded, $\tilde{\mathbf{y}}_{1j}$ denotes the cluster-level components of variables at Level 1 (e.g., manifest or latent means), and $\boldsymbol{\theta}_p$ denotes the parameters of the p-th imputation model. Similarly, for missing values at Level 2, imputations may be generated from

$$y_{2jq} \sim P(y_{2jq}|\tilde{\mathbf{y}}_{1j}, \mathbf{y}_{2j(-q)}, \boldsymbol{\theta}_q)\,, \tag{8}$$

where the subscript $(-q)$ denotes the set of variables from which q is excluded, and $\boldsymbol{\theta}_q$ denotes the parameters of the q-th imputation model. For example, the imputation model at Level 1 may be a multilevel random intercept model (e.g., Schafer & Yucel, 2002; Snijders & Bosker, 2012b; van Buuren, 2011), and the imputation model at Level 2 may be a regression model based on the other variables and cluster-level components at Level 2 (e.g., Rubin, 1987; van Buuren, 2012). The relations between the variables are preserved in the FCS approach by iterating across variables and using each variable and its cluster-level components as predictor variables in every other imputation model. In contrast to JM, however, the FCS approach makes it possible to extract the cluster-level components of variables at Level 1 in different ways, that is, $\tilde{\mathbf{y}}_{1j}$ may include either manifest or latent cluster means of $\mathbf{y}_{1j}$ (or a mixture thereof). Once new imputations have been drawn, the cluster-level components must be updated accordingly.

To illustrate the computational procedure used in the FCS approach, we first describe the general procedure for imputing missing data at Level 2. Then, we describe how manifest and latent cluster means can be generated and incorporated into that procedure. Consider the scenario above, where there are P completely observed variables at Level 1, and Q partially observed variables at Level 2. For the q-th variable, denote observed and missing values as

y_{2jq}^{obs} and y_{2jq}^{mis}, and denote the parameters of the imputation model as $\boldsymbol{\theta}_q = \{\beta_{0q}, \boldsymbol{\beta}_{1q}, \phi_q^2\}$. For variable q at iteration t,

1. Calculate $\tilde{\mathbf{y}}_{1j}^{(t)}$ from $\mathbf{y}_{1ij}$ either as manifest or latent cluster means (see below).
2. Draw $\boldsymbol{\theta}_q^{(t+1)} \sim P(\boldsymbol{\theta}|\tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j}^{(t)})$ given appropriate priors, where $P(\cdot)$ is inverse-Gamma for ϕ_q^2 and multivariate normal for β_{0q} and $\boldsymbol{\beta}_{1q}$ combined.
3. Impute $y_{2jq}^{\text{imp},(t+1)} \sim P(y_{2jq}^{\text{mis}}|\boldsymbol{\theta}_q^{(t+1)}, \tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j(-q)}^{(t)})$ from a univariate normal distribution $N\big(\beta_{0q}^{(t+1)} + \tilde{\mathbf{x}}_{j(-q)}^{(t)}\boldsymbol{\beta}_{1q}^{(t+1)}, \phi_q^{2,(t+1)}\big)$, conditional on the predictor variables $\tilde{\mathbf{x}}_{j(-q)}^{(t)} = (\tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j(-q)}^{(t)})$.

In order to include the manifest means in $\tilde{\mathbf{y}}_{1j}$, an additional step is carried out which simply calculates the manifest mean based on the current scores of $\mathbf{y}_{1ij}$. In the literature, this strategy is more widely known as *passive* imputation (Royston, 2005; van Buuren, 2012). For the p-th variable at Level 1,

1. Calculate $\tilde{y}_{1jp}^{(t)} = \bar{y}_{1\bullet jp} = \dfrac{1}{n_j}\displaystyle\sum_{i=1}^{n_j} y_{1ijp}$.

Alternatively, latent means may be included in $\tilde{\mathbf{y}}_{1j}$. To this end, the latent means are drawn from their posterior distribution, given the other variables and cluster-level components at Level 2. Here, we present a procedure for sampling the latent means using the *plausible value* technique, which regards the observed responses at Level 1 as indicators of an unobserved, latent variable at Level 2 (Mislevy, 1991; Yucel et al., 2007). For the p-th variable at Level 1, let $\boldsymbol{\theta}_p^* = \{\beta_{0p}, \boldsymbol{\beta}_{1p}, \psi_{p|(-p)}^2, \sigma_p^2\}$. Then,

1. Fit[3] the multilevel random intercept model $y_{1ijp} = \hat{\beta}_{0p}^{(t)} + \hat{\boldsymbol{\beta}}_{1p}^{(t)}\tilde{\mathbf{x}}_{j(-p)}^{(t)} + \upsilon_j + \epsilon_{ij}$, obtaining estimates for the conditional mean $\mu_{p|(-p)j}^{(t)} = \hat{\beta}_{0p}^{(t)} + \hat{\boldsymbol{\beta}}_{1p}^{(t)}\tilde{\mathbf{x}}_{j(-p)}^{(t)}$ and the (residual) conditional variances $\hat{\psi}_{p|(-p)}^{2,(t)}$ (at Level 2) and $\hat{\sigma}_p^{2,(t)}$ (at Level 1) of y_{1ijp}, given the predictor variables $\tilde{\mathbf{x}}_{j(-p)}^{(t)} = (\tilde{\mathbf{y}}_{1j(-p)}^{(t)}, \mathbf{y}_{2j}^{(t)})$.
2. Draw $u_{1jp}^{(t)} \sim P(u_{1jp}|\tilde{\mathbf{y}}_{1j(-p)}^{(t)}, \mathbf{y}_{2j}^{(t)}, \hat{\boldsymbol{\theta}}_p^{*(t)})$ from a univariate normal distribution $N(\tilde{u}_{1jp}^{(t)}, U_{1jp}^{(t)})$, where the mean and variance are calculated as follows.

[3]This approach is similar to obtaining "empirical Bayes" estimates for random effects in multilevel modeling (e.g., Laird & Ware, 1982; Morris, 1983). The estimation of the model parameters can be achieved by maximum-likelihood (ML) or Bayesian methods. Here, we used Bayesian estimates of the model parameters because ML led to convergence issues in smaller samples. As an alternative, a fully Bayesian procedure can be used in which the estimates $\hat{\boldsymbol{\theta}}^{*(t)}$ are replaced with Bayesian posterior draws $\boldsymbol{\theta}^{*(t)}$ (see the Discussion section).

i) $\tilde{u}_{1jp}^{(t)} = \mu_{p|(-p)j}^{(t)} + \lambda_{p|(-p)j}^{(t)} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{1ijp} - \mu_{p|(-p)j}^{(t)})$, where $\lambda_{p|(-p)j}^{(t)} = \frac{\hat{\psi}_{p|(-p)}^{2,(t)}}{\hat{\psi}_{p|(-p)j}^{2,(t)} + \hat{\sigma}_p^{2,(t)}/n_j}$ is the reliability of the conditional cluster means of y_{1ijp}, given $\tilde{\mathbf{y}}_{1j(-p)}$ and $\mathbf{y}_{2j}$

ii) $U_{1jp}^{(t)} = \lambda_{p|(-p)j}^{(t)} \cdot \frac{\hat{\sigma}_p^{2,(t)}}{n_j}$, where $\lambda_{p|(-p)j}^{(t)}$ is as defined above

3. Set $\tilde{y}_{1jp}^{(t)} = u_{1jp}^{(t)}$.

Because the latent cluster means are regarded as unobservable in the plausible value approach, new values for the latent means must be generated at each iteration even if the underlying variable is completely observed. This acknowledges the fact that, because only a finite number of observations are used to estimate the cluster-level component, any single estimate of the (latent) cluster mean is subject to uncertainty (for related approaches involving plausible values, see Blackwell et al., 2017b; Yang & Seltzer, 2016).

Notice that, when using latent cluster means, FCS becomes very similar to JM. Only in Step 2a above does the expression appear to be slightly different from the corresponding step in JM (Step 1a). However, the similarity becomes fully visible when the expression in Step 2a is rearranged:

$$\tilde{u}_{1jp} = (1 - \lambda_{p|(-p)j}) \cdot \mu_{p|(-p)j} + \lambda_{p|(-p)j} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \ . \tag{9}$$

This illustrates that the FCS approach with latent cluster means employs the same kind of shrinkage that is also used in JM. The handling of the overall mean of y_{1ijp} differs because the conditional mean $\mu_{p|(-p)j}$ is redefined in FCS to include the overall mean.

Using manifest versus latent cluster means. The fact that either manifest (FCS-MAN) or latent cluster means (FCS-LAT) can be used in FCS raises the question of which procedure is most appropriate in a given scenario. For the purposes of this article, we assume that the distributional assumptions of the JM hold in the population, and FCS is used to treat missing data at Level 2. For FCS to be consistent with the JM, the conditional distributions employed in FCS must imply the same joint distribution as the JM. Even though several authors have argued that this is the case for balanced data (i.e., with clusters of the same size; Carpenter & Kenward, 2013; Lüdtke et al., 2017; Mistler, 2015), it has been suggested that the same does not hold in unbalanced data for FCS-MAN (i.e., with clusters of different sizes; Resche-Rigon & White, in

press). More precisely, Resche-Rigon and White demonstrated that the conditional distribution implied by the JM does not depend solely on the manifest means but also on cluster size, to the effect that FCS-MAN would need to account for the Level 1 heteroscedasticity that is due to differences in cluster size.

In the present article, we extend this line of reasoning in two different ways. First, we show in the Appendix that, when missing data at Level 2 are treated with FCS-MAN, (a) variance estimates for variables at Level 2 remain unbiased, but (b) estimates of covariances at Level 2 are biased towards zero in unbalanced data. Second and in contrast to FCS-MAN, we argue that FCS-LAT provides estimates that are consistent with the JM regardless of whether or not the data are balanced because the "shrinkage" estimates of the latent cluster means take the differences in cluster size into account (i.e., the differences in reliability of the cluster means; see also Raudenbush & Bryk, 2002). The bias under FCS-MAN is difficult to evaluate in detail because it depends on the distribution of clusters sizes in the sample. In the Appendix, the bias is derived under the assumption that the number of clusters goes to infinity and that the missing data occur completely at random (MCAR) and independently of cluster size. Consider again the case with only one variable at Level 1 (y) and one variable at Level 2 (z). Then, the bias of the estimator of the covariance of y with z can be expressed as

$$\%Bias(\hat{\sigma}_{yz}) = \alpha \left[\sum_{k\in\mathcal{S}} \left(\tfrac{k}{\bar{n}} - 1\right) \pi_k \left(\tau_y^2 + \tfrac{\sigma_y^2}{k}\right)\right] \left[\sum_{k\in\mathcal{S}} \pi_k \left(\tau_y^2 + \tfrac{\sigma_y^2}{k}\right)\right]^{-1}, \tag{10}$$

where α denotes the probability of missing data, $\mathcal{S}$ denotes the set of cluster sizes (k) uniquely present in the data, π_k the proportion of clusters with size k, $\bar{n}$ the average cluster size, and σ_y^2 and τ_y^2 the variance components of y at Level 1 and 2, respectively. The fraction in this expression relates the variability of the cluster means for each $k \in \mathcal{S}$ to the variability of the cluster means overall. Because smaller clusters, which tend to have larger variability in the observed cluster means, receive negative weights ($\frac{k}{\bar{n}} - 1$) as opposed to larger clusters with less variability, the bias in $\hat{\sigma}_{yz}$ tends to be negative (i.e., towards zero). In balanced data ($k = \bar{n}$ for all $k \in \mathcal{S}$), the bias is zero. However, even with unbalanced data, the bias appears to be relatively small. This is illustrated in Figure 1 for the special case of uniformly distributed cluster sizes (n_j), different levels of the average cluster size ($\bar{n}$), different choices for the range

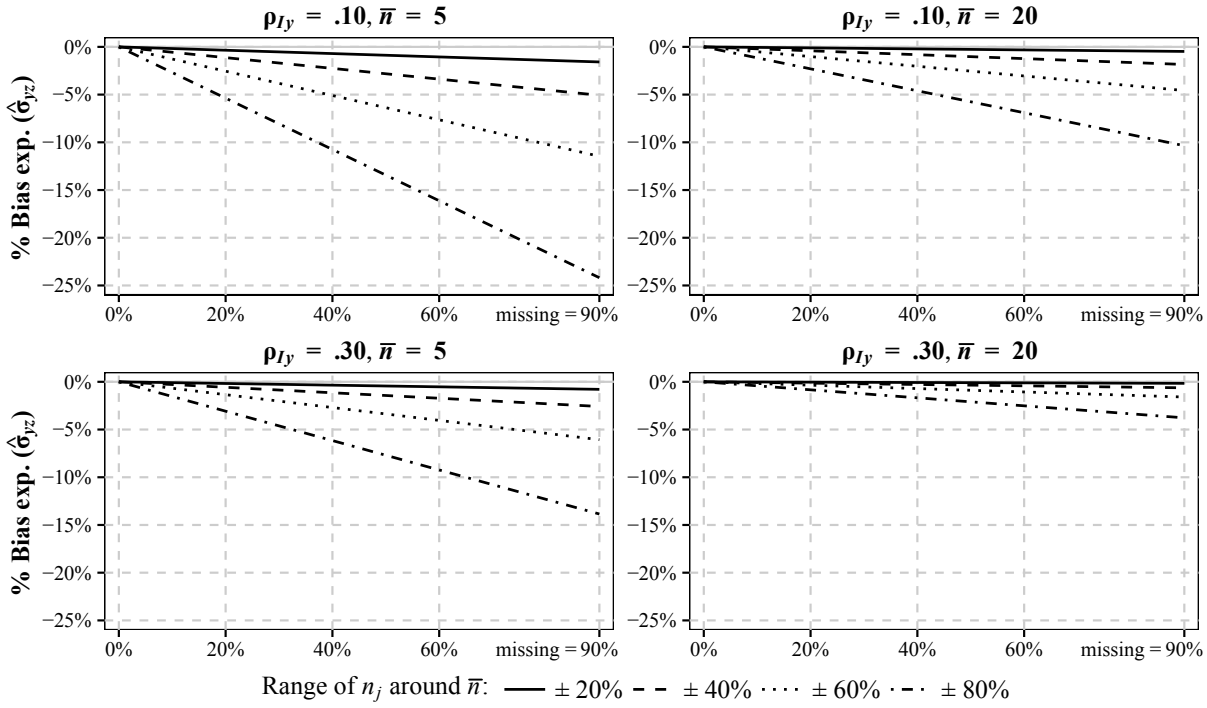

Figure 1: Expected bias for the covariance of y with z under FCS-MAN, for varying amounts of missing data, different intraclass correlations of y (ρ_{I_y}), different average cluster sizes ($\bar{n}$), and different ranges of cluster sizes (n_j, assuming a uniform distribution).

of the cluster sizes, different amounts of missing data, and different values for the intraclass correlation (ICC) of y (ρ_{I_y}). Relatively extreme conditions appear to be necessary in order for the parameter estimates to be distorted to a degree that is no longer tolerable (e.g., $< -10\%$). Note also that, although the nonequivalence of FCS-LAT and FCS-MAN in unbalanced data holds in general, the expression for the bias was derived under relatively strong assumptions and should not be generalized to more general conditions.

Even though the use of FCS-LAT and JM may be preferred from a theoretical point of view, it is important to acknowledge that the statistical models underlying these procedures may be more difficult to estimate than those underlying FCS-MAN, especially in smaller samples or when variables at Level 1 have little variance at Level 2 (e.g., Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Similarly, the different procedures may be more or less accurate depending on the missing data mechanism, the proportion of missing values, and the information available from auxiliary variables at Level 1 (e.g., Andridge & Thompson, 2015). Thus, it is important to study the properties of the different procedures in less than ideal conditions (e.g., very few clusters, small vs. large ICCs, more or less informative data loss, different types of unbalanced

data). Finally, either procedure may provide substantial gains in accuracy and efficiency when compared with still popular but simpler methods such as single-level MI or LD. To this end, we conducted two computer simulation studies. In Study 1, we evaluated the performance of the different methods under a variety of conditions with balanced data. In Study 2, we focused on the more general case with unbalanced data and the potential bias associated with using manifest cluster means (i.e., with FCS-MAN).

Study 1

In the following section, we present the results of the first simulation study in which we compared the performance of JM and FCS for missing data at Level 2 with balanced data.

Simulation procedure

Data generation. For the purpose of this study, we focused on the special case where there is only one variable at Level 1 (y) and one variable at Level 2 (z). The data were generated using the model in Equation 6. For the two variables y and z, the model reads:

$$\begin{aligned} y_{ij} &= \mu_y + u_{yj} + e_{ij} \\ z_j &= \mu_z + u_{zj} \ . \end{aligned} \tag{11}$$

For simplicity, we assumed that all variables were standardized with mean zero ($\mu_y = \mu_z = 0$) and unit variance. To specify the variances and covariances at Level 1 and 2, we defined the ICC of y (ρ_{I_y}) and the correlation between the two variables at Level 2 (ρ_{yz}). Missing values were induced in z depending on the observed cluster means $\bar{y}_{\bullet j}$ using the following generalized linear model

$$r_j = \alpha_0 + \lambda \bar{y}_{\bullet j} + \delta_j \ , \tag{12}$$

where r_j denotes the latent propensity for observing z_j, α_0 is a quantile of the standard normal distribution according to some missing data probability α (e.g., $\alpha_0 = -.842$ for $\alpha = 20\%$ missing data), and λ is the effect of y on the response propensity of z. The variance of r_j was fixed at 1, and the residuals δ_j were drawn from a normal distribution with variance $1 - \lambda^2 \, Var(\bar{y}_{\bullet j})$. A value z_j was deleted if $r_j > 0$.

Table 1: Simulated conditions in Study 1 and Study 2

Design Conditions	Study 1	Study 2
Cluster size (n or $\bar{n}$)	5, 20	5, 20
Number of clusters (J)	30, 50, 100, 200, 500, 1000	50, 200, 1000
Range in cluster size	–	uniform (±40%, 80%), bimodal (±40%, 80%)
ICC of y (ρ_{Iy})	.10, .30	.10, .30
Correlation of y and z (ρ_{yz})	.5	.5
Effect of y on missingness (λ)[a]	0, 0.5, 1	0.5
Portion of missing data (α)	20%, 40%	20%, 40%

[a] The values for λ are given in a standardized metric. A value of 1 constitutes a strong, deterministic missing data mechanism, in which all values that lie beyond a certain cutoff are deleted.

Table 1 summarizes the simulation conditions. In Study 1, we simulated conditions with different cluster sizes ($n = 5, 20$) that are typical in educational research (e.g., students in school classes, repeated measurements). We varied the number of clusters between $J = 30$ and $1,000$ to examine both the small- and large-sample properties of the procedures. We varied the ICC of y in two steps ($\rho_{Iy} = .10, .30$) to reflect conditions with more or less information, respectively, in y located at Level 2 (see also Lüdtke et al., 2008). We simulated conditions in which data were missing completely at random (MCAR, $\lambda = 0$) or moderately or strongly missing at random (MAR, $\lambda = 0.5$ or 1), and either 20% or 40% of the data were missing.[4] Each condition was replicated 1,000 times.

Imputation. To impute missing values with JM, we used the R package `jomo` (Quartagno & Carpenter, 2016a). To implement the FCS approach, we used the R packages `mice` (van Buuren & Groothuis-Oudshoorn, 2011) and `miceadds` (Robitzsch, Grund, & Henke, 2017) for imputation with FCS-MAN and FCS-LAT, respectively. In addition, we included single-level MI with FCS (FCS-SL) and listwise deletion (LD) for the purpose of comparison. Single-level FCS was implemented as "flat-file" imputation thus treating all variables as variables at Level 1 (see also van Buuren, 2011); because this resulted in different imputations within clusters for variables at Level 2, imputations were averaged within clusters prior to being analyzed. With

[4]The values for λ are given here in a standardized metric. The actual values of λ in the data-generating model were different because they also depended on the ICC of y and the sample size at Level 1. The actual values were chosen in such a way that they implied a standardized effect of y of size 0, 0.5, and 1, respectively.

each procedure, we generated 10 imputed data sets. For JM, we chose 1,000 burn-in iterations and 500 iterations between imputations. For the FCS approach, we chose 20 iterations per imputation. These values were found to be sufficient to ensure convergence as determined by assessing diagnostic plots. Default flat prior distributions were used for all procedures.

Analysis and parameters of interest. The software M*plus* was used to analyze the data (L. K. Muthén & Muthén, 2012). Using M*plus*, we estimated the mean (μ_z) and the variance (σ_z^2) of z as well as the (latent) covariance between y and z (σ_{yz}). Furthermore, we estimated the regression coefficients relating y and z at Level 2 using two additional regression models with y regressed on z (β_{yz}) and z regressed on y (β_{zy}). For each parameter and each simulation condition, we calculated the bias, the root mean squared error (RMSE) and the coverage rate of the 95% confidence interval. To calculate the bias and RMSE, we used the average estimates from the complete data sets as a point of reference instead of the "true" values in the data-generating model. This was necessary because, even without missing data, the estimates of some parameters were biased in some conditions, rendering a comparison with the "true" values less useful. The complete set of results, including the raw bias and RMSE for all missing data procedures as well as those for the complete data sets, is provided in Supplement D of the supplemental online materials.

Results

We first focus on the estimates of the mean and variance of z ($\hat{\mu}_z$ and $\hat{\sigma}_z^2$), and the covariance of y with z ($\hat{\sigma}_{yz}$). The bias for the mean, variance and covariance is presented in Figure 2 for conditions with different cluster sizes (n) and numbers of clusters (J), different amounts of information at Level 2 as reflected by the ICC of y (ρ_{Iy}), and 20% missing data under moderate MAR ($\lambda = 0.5$). All procedures for multilevel MI provided approximately unbiased estimates of the three parameters in larger samples ($J \rightarrow 1,000$).The procedures differed only in the sample size needed to achieve these results. Whereas FCS-MAN and FCS-LAT provided approximately unbiased estimates of the population parameters even in small samples ($n = 5$, $J = 30$) and with little information at Level 2 ($\rho_{Iy} = .10$), JM required slightly larger samples to provide unbiased estimates of these parameters ($J \geq 100$). By contrast, FCS-SL and LD

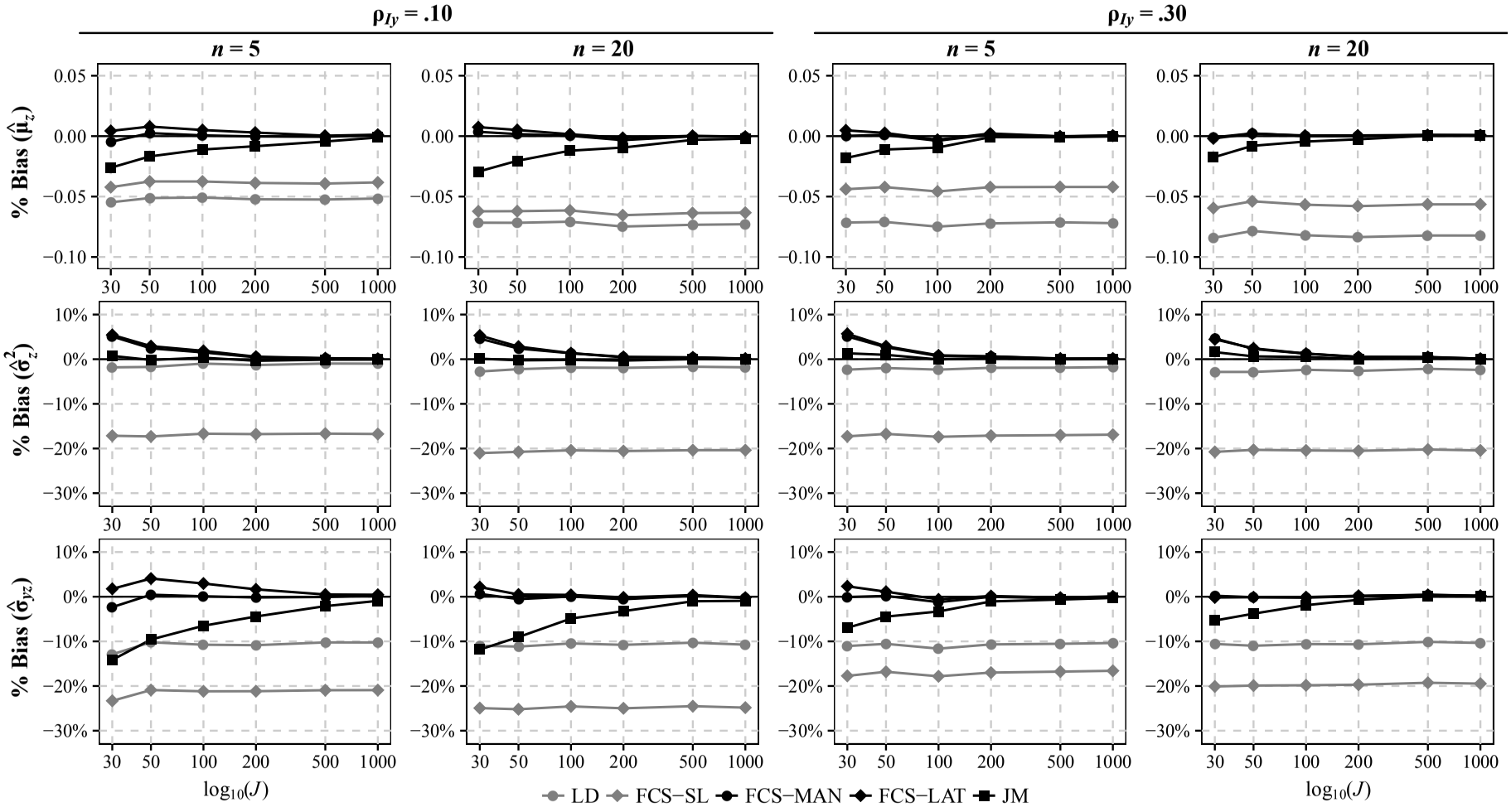

Figure 2: Estimated bias for the mean and the variance of z ($\hat{\mu}_z$ and $\hat{\sigma}^2_z$), and the covariance of y with z ($\hat{\sigma}_{yz}$) for varying sample sizes at Level 1 (n) and Level 2 (J), and ICC of y (ρ_{Iy}), with 20% missing data (MAR, $\lambda = 0.5$). LD = listwise deletion; FCS-SL = single-level FCS; FCS-MAN = two-level FCS with manifest cluster means; FCS-LAT = two-level FCS with latent cluster means; JM = joint modeling.

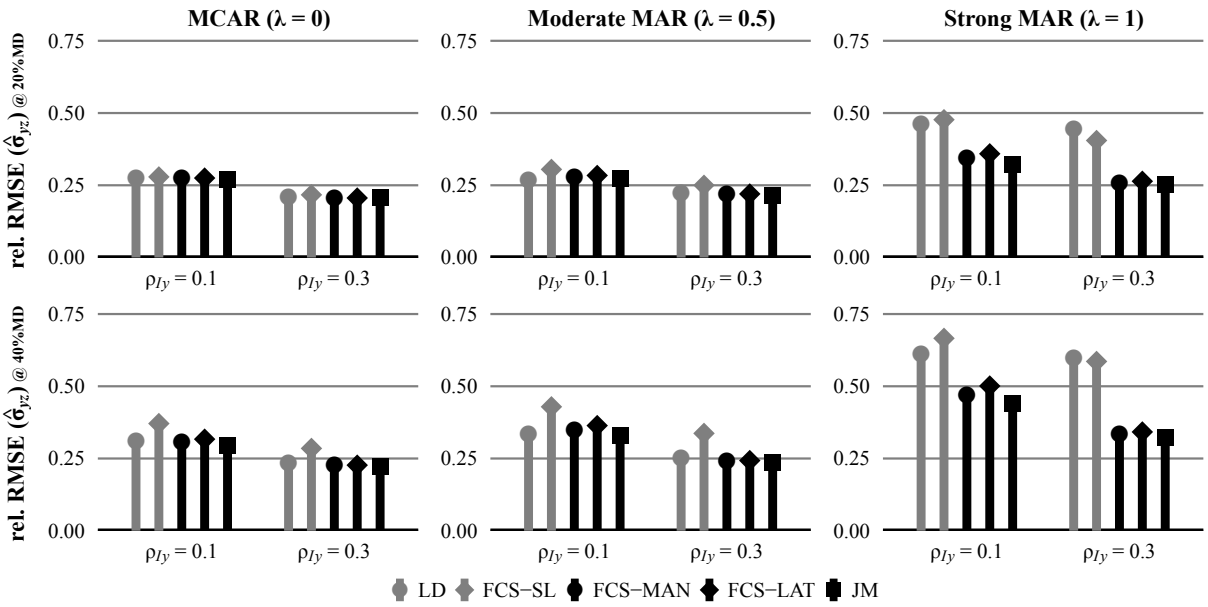

Figure 3: Relative RMSE for the covariance of y with z ($\hat{\sigma}_{yz}$) with $n = 5$ and $J = 200$, for varying ICCs of y (ρ_{Iy}), different missing data mechanisms (λ), and different portions of missing data (MD). LD = listwise deletion; FCS-SL = single-level FCS; FCS-MAN = two-level FCS with manifest cluster means; FCS-LAT = two-level FCS with latent cluster means; JM = joint modeling.

provided strongly biased results of the mean and covariance regardless of sample size, and FCS-SL also led to biased estimates of the variance of z.

The results obtained from the different procedures were also affected by the missing data mechanism (λ), the amount of missing data, and the ICC of y (ρ_{Iy}), as illustrated in Figure 3 for the RMSE of the covariance of y with z ($\hat{\sigma}_{yz}$). These factors can be regarded as determinants of the fraction of missing information (FMI), that is, the loss of precision associated with parameter estimation with missing data (e.g., Andridge & Thompson, 2015). Larger portions of missing data and more informative missing data mechanisms (λ) both increased the RMSE, whereas an increase in the ICC of y reduced it. In comparison with LD, the different MI procedures tended to benefit more from a larger ICC of y, especially under more severe losses of data (i.e., 40% missing data, MAR).

The results for the regression models with y regressed on z ($\hat{\beta}_{yz}$) and z regressed on y ($\hat{\beta}_{zy}$) are summarized in Table 2. Overall, the results were consistent with the results presented above; that is, we obtained approximately unbiased estimates of β_{yz} and β_{zy} in larger samples, but the estimates had slight downward biases in smaller samples. By contrast, estimates obtained from FCS-SL and LD were biased, especially under MAR and regardless of sample size (see also

Table 2: Study 1: Bias (in %), Relative RMSE, and Coverage of the 95% Confidence Interval for the Regression Coefficients of y on z and z on y ($\hat{\beta}_{yz}$ and $\hat{\beta}_{zy}$) for Small ICC of y ($\rho_{I_y} = .10$) and 20% Missing Data (MAR, $\lambda = 0.5$)

	Bias (%)				Rel. RMSE				Coverage (%)			
	FCS-SL	FCS-MAN	FCS-LAT	JM	FCS-SL	FCS-MAN	FCS-LAT	JM	FCS-SL	FCS-MAN	FCS-LAT	JM
						Regression $y \sim z$ ($\hat{\beta}_{yz}$)						
$n = 5$												
$J = 30$	−7.0	−6.7	−3.6	−14.1	0.708	0.712	0.727	0.666	90.4	90.4	89.6	93.6
$J = 50$	−4.2	−2.0	0.7	−9.2	0.520	0.525	0.539	0.504	93.1	93.4	91.6	95.0
$J = 100$	−5.3	−1.6	0.8	−6.8	0.373	0.381	0.383	0.371	93.9	94.3	93.9	95.3
$J = 200$	−5.2	−0.7	0.9	−4.2	0.253	0.254	0.257	0.253	94.6	95.1	94.1	95.2
$J = 500$	−5.1	−0.3	0.3	−2.0	0.166	0.162	0.163	0.161	93.3	94.8	94.5	95.0
$J = 1000$	−5.0	0.1	0.3	−0.8	0.121	0.115	0.115	0.114	92.8	94.6	94.2	94.7
$n = 20$												
$J = 30$	−4.5	−3.4	−2.8	−11.1	0.481	0.473	0.471	0.458	90.7	92.3	92.0	94.5
$J = 50$	−5.7	−3.0	−2.6	−8.8	0.348	0.349	0.349	0.345	93.1	92.9	93.1	95.1
$J = 100$	−5.3	−1.5	−1.1	−4.8	0.256	0.255	0.252	0.252	92.6	94.0	93.6	94.3
$J = 200$	−5.5	−1.0	−0.7	−3.0	0.185	0.180	0.179	0.179	92.6	94.4	93.8	95.0
$J = 500$	−5.2	−0.2	−0.1	−1.1	0.119	0.108	0.107	0.107	91.3	95.3	95.8	95.9
$J = 1000$	−5.6	−0.4	−0.4	−0.9	0.096	0.079	0.079	0.079	88.0	94.0	93.9	94.5
						Regression $z \sim y$ ($\hat{\beta}_{zy}$)						
$n = 5$												
$J = 30$	−20.3	−12.9	−9.5	−17.1	1.067	0.961	1.025	1.001	86.3	93.3	93.4	94.2
$J = 50$	−16.8	−3.2	−0.2	−8.8	0.822	0.760	0.797	0.794	84.9	93.5	93.1	92.9
$J = 100$	−19.6	−1.0	1.8	−6.7	0.585	0.589	0.605	0.576	82.4	94.2	94.7	94.0
$J = 200$	−21.0	−0.5	1.7	−4.7	0.389	0.379	0.400	0.382	80.5	95.1	95.3	93.8
$J = 500$	−20.9	0.0	0.6	−2.1	0.273	0.220	0.223	0.211	72.2	96.8	96.2	96.1
$J = 1000$	−20.9	0.3	0.4	−1.0	0.243	0.156	0.155	0.151	59.0	96.0	95.5	95.8
$n = 20$												
$J = 30$	−24.0	0.3	2.0	−11.9	0.544	0.584	0.595	0.541	82.7	92.7	92.7	94.4
$J = 50$	−24.6	−0.4	0.7	−9.2	0.408	0.409	0.415	0.379	81.5	93.7	93.9	94.5
$J = 100$	−24.4	−0.0	0.3	−5.1	0.333	0.286	0.284	0.271	75.3	94.2	94.2	95.2
$J = 200$	−24.9	−0.5	−0.0	−3.2	0.292	0.196	0.197	0.190	60.4	94.3	93.4	94.8
$J = 500$	−24.5	0.2	0.4	−1.0	0.262	0.119	0.118	0.115	27.8	95.1	95.4	95.0
$J = 1000$	−24.8	−0.3	−0.2	−0.9	0.258	0.086	0.086	0.085	6.0	95.2	94.4	94.7

Note. n = cluster size; J = number of clusters; FCS-MAN = two-level FCS with manifest cluster means; FCS-LAT = two-level FCS with latent cluster means; JM = joint modeling.

Supplement D). It is interesting that, even though the bias observed in smaller samples was largest for JM, the estimates under JM were also the most accurate overall in these conditions as reflected by the RMSE, indicating that the variability of the estimates was lower under JM as compared with FCS-MAN and FCS-LAT. The coverage of the 95% confidence interval was

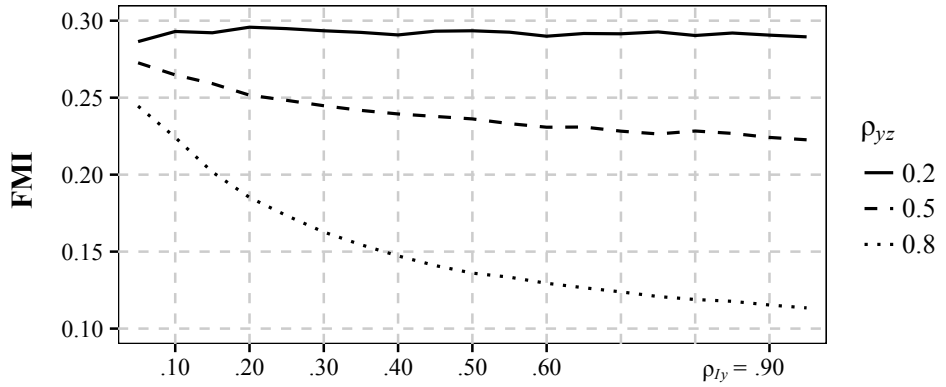

Figure 4: Fraction of missing information (FMI) for estimates of the covariance of y with z ($\hat{\sigma}_{yz}$) with $n = 5$ and $J = 200$, and 20% missing data (MAR, $\lambda = 0.5$), for varying ICCs of y (ρ_{I_y}) and different correlations between y and z (ρ_{yz}).

close to the nominal 95% in most conditions but was sometimes too low under FCS in very small samples ($n = 5$ and $J = 30$, with $\rho_{I_y} = .10$) or in conditions with larger portions of missing data (see Supplement D). However, as might be expected from this collection of results, the near-optimal coverage under JM (and to a lesser extent under FCS) occurred at the expense of standard errors that were sometimes too large as compared with the variance of the parameter estimates in smaller samples ($n = 5$, $J \leq 100$).

Summary

Taken together, these results indicate that (a) the overall performance of FCS-MAN, FCS-LAT, and JM is similar in terms of bias, RMSE, and coverage of the 95% confidence interval, (b) the performance of the procedures may differ in smaller samples or when larger portions of the data are missing, and (c) including variables with substantial variance between clusters (i.e., large ICC) can be extremely beneficial for MI because these variables can provide crucial information about missing values at Level 2. To further illustrate the importance of including variables at Level 1 for imputing variables at Level 2, we conducted an additional simulation study in which we varied the ICC of y in a range from .05 to .95 and the correlation of y and z between .20 and .80; we then estimated the fraction of missing information (FMI) under JM in each condition (otherwise $n = 5$, $J = 200$, $\lambda = 0.5$, 20% missing data). The results are shown in Figure 4. As can be seen, the FMI tended to decrease as the ICC of y (ρ_{I_y}) increased depending on the

correlation between y and z (ρ_{yz}). For example, increasing the ICC of y from .10 to .30 reduced the FMI by approximately 7.6% when the correlation was moderate ($\rho_{yz} = .5$) and by 27.4% when the correlation was strong ($\rho_{yz} = .8$). With weak correlation ($\rho_{yz} = .2$), increasing the ICC of y did not noticeably change the FMI, as may be expected from the fact that such weakly correlated variables are not able to explain much variance associated with missing values. This illustrates that researchers who want to treat missing data at Level 2 by means of MI should include auxiliary variables at Level 1, especially when the auxiliary variables (a) are strongly related to the variables with missing data and (b) contain substantial variance between clusters as indicated by their ICCs.

Study 2

In Study 1, we evaluated the performance of JM and FCS in balanced data. In practice, however, most research conducted with multilevel data is based on unbalanced data. For this reason, in Study 2, we focused on the more general case with unbalanced data (i.e., clusters of different sizes).

Simulation procedure

Following the same general procedures as in Study 1, we generated clusters of varying size n_j in Study 2, where n_j was drawn either from a uniform distribution in the range of ±40% or 80% around the average cluster size $\bar{n}$ (e.g., for $\bar{n} = 5$ and range ±80%, $n_j = 1, 2, \ldots, 9$) or a bimodal distribution that included only the extreme points of this range (e.g., for $\bar{n} = 5$ and range ±80%, $n_j = 1$ or 9; see Table 1). Even though the resulting range of n_j is quite typical in educational research, the two distributions should be regarded as extreme choices given that the distribution of cluster sizes in practice is often bell-shaped and possibly asymmetrical. For this reason, the results presented here should be regarded as a lower bound for the performance of the different MI procedures in practice.

Imputation. We used the same procedures as in Study 1. In addition, on the basis of Resche-Rigon and White's (in press) suggestions, we included an MI procedure that used

manifest cluster means similar to FCS-MAN but also acknowledged heteroscedasticity at Level 1 by including n_j and the interaction of n_j with $\bar{y}_{\bullet j}$ as additional predictor variables in the imputation model (FCS-NJ).

Results

To avoid redundancy, we focus on reporting the results for the covariance of y with z ($\hat{\sigma}_{yz}$; for the remaining results, see Supplement D). These results are summarized in Table 3 for conditions with a low ICC of y ($\rho_{I_y} = .10$) and 20% missing data. Consistent with our expectations, FCS-MAN provided slightly biased estimates of the covariance, even in conditions with very large samples ($J \rightarrow 1,000$). However, the bias usually remained relatively small and was restricted to conditions with few observations per cluster ($n = 5$) and strongly unbalanced data (±80%). Biases larger than −10% were obtained under FCS-MAN only in conditions with 40% missing data and strongly unbalanced data (±80%, uniform or bimodal). In line with our expectations, the bias was approximately twice as large in conditions with 40% missing data than in the conditions displayed in Table 3 (see Supplement D). In line with the recommendations in the literature, the bias under FCS-MAN was reduced to essentially zero when the cluster size was included in the imputation model (FCS-NJ). Similarly, under FCS-LAT or JM, the bias was approximately zero in larger samples even with strongly unbalanced data ($J \rightarrow 1,000$). In smaller samples, estimates of the covariance were slightly biased upwards under FCS-LAT ($J = 50$) and downwards under JM ($J \leq 200$). In terms of the RMSE, estimates obtained from JM were slightly more accurate in smaller samples ($J = 50$). In larger samples, differences in the RMSE tended to be very small. Similarly, the coverage of the 95% confidence interval was very close to the nominal 95% for all procedures in all but very extreme conditions (e.g., for FCS-MAN; see Supplement D). Under JM (and to a lesser extent under FCS), we again observed standard errors that were sometimes too large as compared with the variance of the parameter estimates in smaller samples ($n = 5$, $J = 50$). Taken together, these results indicate that (a) covariance estimates obtained under FCS-MAN can be biased in unbalanced data, (b) this bias is likely to be very small in any practical application of multilevel MI, (c) FCS-NJ, FCS-LAT, and JM all provide approximately unbiased results when samples are sufficiently

Table 3: Study 2: Bias (in %), Relative RMSE, and Coverage of the 95% Confidence Interval for Covariance of y and z ($\hat{\sigma}_{yz}$) in Unbalanced Data for Small ICC of y ($\rho_{Iy} = .10$) and 20% Missing Data (MAR, $\lambda = 0.5$)

	Bias (%)				Rel. RMSE				Coverage (%)			
	FCS-MAN	FCS-NJ	FCS-LAT	JM	FCS-MAN	FCS-NJ	FCS-LAT	JM	FCS-MAN	FCS-NJ	FCS-LAT	JM
Moderately unbalanced (uniform, ±40%)												
$\bar{n} = 5$												
$J = 50$	0.4	0.9	5.1	−8.4	0.591	0.601	0.611	0.547	94.5	95.3	94.0	94.5
$J = 200$	0.1	0.3	2.1	−3.8	0.285	0.284	0.289	0.273	96.1	96.5	95.5	96.7
$J = 1000$	−0.9	−0.4	−0.1	−1.5	0.132	0.132	0.131	0.130	93.8	94.3	94.5	94.1
$\bar{n} = 20$												
$J = 50$	−0.0	1.0	1.0	−8.0	0.446	0.466	0.452	0.421	92.6	93.9	92.4	91.4
$J = 200$	−0.5	−0.5	−0.3	−3.4	0.215	0.214	0.213	0.211	94.1	94.4	94.1	93.9
$J = 1000$	−0.1	−0.1	0.0	−0.6	0.094	0.095	0.093	0.094	94.7	94.2	94.8	95.0
Strongly unbalanced (uniform, ±80%)												
$\bar{n} = 5$												
$J = 50$	−0.8	2.5	4.5	−7.5	0.608	0.639	0.621	0.566	94.3	96.1	93.9	94.2
$J = 200$	−2.3	0.4	1.7	−3.6	0.300	0.302	0.307	0.291	94.1	94.1	93.8	93.6
$J = 1000$	−3.2	−0.3	0.2	−1.2	0.136	0.132	0.133	0.131	94.2	94.7	94.1	94.3
$\bar{n} = 20$												
$J = 50$	1.1	1.5	2.2	−7.5	0.437	0.449	0.435	0.401	94.1	95.0	93.7	93.6
$J = 200$	−0.4	0.2	0.5	−2.5	0.219	0.216	0.218	0.213	94.3	94.7	94.7	94.5
$J = 1000$	−0.9	−0.3	−0.1	−0.8	0.099	0.098	0.098	0.097	94.4	94.2	93.8	94.1
Moderately unbalanced (bimodal, ±40%)												
$\bar{n} = 5$												
$J = 50$	−0.4	1.4	3.8	−8.5	0.629	0.644	0.648	0.570	92.9	93.1	92.4	92.5
$J = 200$	−0.8	0.2	1.6	−3.8	0.290	0.289	0.294	0.278	93.8	95.0	94.0	94.3
$J = 1000$	−1.1	0.0	0.3	−1.1	0.132	0.132	0.132	0.130	94.4	94.2	94.6	94.3
$\bar{n} = 20$												
$J = 50$	0.5	0.5	1.3	−8.5	0.427	0.435	0.439	0.403	93.5	94.0	92.9	92.4
$J = 200$	−0.8	−0.7	−0.6	−3.5	0.209	0.207	0.208	0.205	95.1	95.5	95.5	95.2
$J = 1000$	−0.1	0.1	0.2	−0.4	0.096	0.096	0.096	0.095	94.3	94.6	94.5	94.6
Strongly unbalanced (bimodal, ±80%)												
$\bar{n} = 5$												
$J = 50$	−3.6	1.8	2.9	−7.2	0.627	0.653	0.652	0.600	93.7	94.2	92.6	92.7
$J = 200$	−6.0	0.1	1.5	−3.2	0.298	0.296	0.298	0.290	93.8	95.3	95.3	95.0
$J = 1000$	−7.1	−0.7	0.0	−1.0	0.149	0.135	0.137	0.135	91.0	94.0	94.0	94.5
$\bar{n} = 20$												
$J = 50$	−1.6	0.7	2.0	−7.7	0.471	0.473	0.479	0.444	93.1	94.2	93.3	92.6
$J = 200$	−2.7	−0.4	0.1	−2.8	0.231	0.227	0.228	0.223	94.4	95.3	94.3	94.1
$J = 1000$	−2.8	−0.4	0.3	−0.6	0.110	0.104	0.105	0.105	92.9	94.0	93.6	93.7

Note. $\bar{n}$ = average cluster size; J = number of clusters; FCS-MAN = two-level FCS with manifest cluster means; FCS-NJ = two-level FCS with manifest cluster means and cluster size (n_j); FCS-LAT = two-level FCS with latent cluster means; JM = joint modeling.

large, and (d) the performance of the procedures may differ in smaller samples in terms of bias and overall accuracy (RMSE).

Empirical example

To illustrate the treatment of missing data at both Level 1 and 2 using MI, we applied the procedures used in Study 1 to the German subsample of the Programme for International Student Assessment (PISA; OECD 2014). We were interested in the effects of the availability of computers at school on students' mathematics achievement when controlling for general aspects of students' learning environment. We controlled for students' gender, their economic, social, and cultural status (ESCS), and ratings on classroom management and student-teacher relations. To control for confounding effects of school size, we also included the number of students who were 15 years of age as an additional covariate. For the purpose of illustration, we used only the first plausible value for students' mathematics achievement and ignored issues related to unequal probabilities of being selected into the sample that may have been due to the sampling design.[5]

The data set included a total of 5,001 students nested within 230 schools with 3 to 25 students participating per school (with 90% of the schools having between 11 and 25 participants; $\bar{n} = 21.7$). The number of computers at school (Level 2) was missing for 17.4% of the schools and the number of students at age 15 for 14.8%. At the student level (Level 1), observations were missing for ESCS (17.2%), classroom management (45.1%), and student-teacher relations (44.6%; see OECD, 2014). We generated 20 imputations for the missing data using (a) JM as implemented in the R package `jomo`, (b) the FCS approach implemented in the R package `mice` with manifest cluster means for all student-level variables using passive imputation (FCS-MAN), (c) the FCS approach similar to approach (b) but with latent cluster means for students' math achievement and their ratings on classroom management and student-teacher relations using the plausible value approach implemented in `miceadds` (FCS-LAT), and (d) single-level

[5] In practice, the procedure would need to be repeated for each plausible value, resulting in imputations "nested" within plausible values (Rubin, 2003; Weirich et al., 2014). Unless differences in selection probability were fully accounted for by the observed variables, these issues would need to be addressed by including survey weights in the imputation and the analysis model (Rust, 2013; Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

Table 4: Parameter Estimates Obtained From the PISA 2012 Data in the Empirical Example Using Different MI Procedures

	FCS-SL		FCS-MAN		FCS-LAT		JM	
Parameter	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Intercept	482.410	(11.096)	474.541	(7.299)	475.043	(7.668)	474.585	(7.870)
Level 1								
Gender	25.287	(1.754)	25.234	(1.760)	25.152	(1.756)	25.237	(1.751)
ESCS	12.458	(1.327)	10.203	(1.340)	10.260	(1.328)	10.304	(1.294)
Classroom management	3.582	(1.069)	3.697	(1.439)	3.572	(1.215)	3.696	(1.362)
Student-teacher relations	1.488	(1.184)	4.168	(1.398)	4.565	(1.485)	3.888	(1.380)
Level 2								
Number of computers	0.166	(0.111)	0.198	(0.091)	0.213	(0.097)	0.197	(0.090)
Number of students	0.034	(0.090)	0.104	(0.068)	0.093	(0.071)	0.101	(0.066)
ESCS	112.520	(8.126)	106.306	(5.930)	106.177	(6.278)	108.139	(5.828)
Classroom management	98.069	(49.026)	42.964	(20.510)	44.705	(20.019)	39.807	(19.320)
Student-teacher relations	−47.962	(39.633)	−32.630	(13.032)	−30.334	(15.044)	−30.597	(14.557)
Intercept variance	1140.753	(202.834)	1257.663	(158.599)	1280.906	(173.783)	1265.579	(163.697)
Residual variance	4044.420	(86.424)	4058.046	(87.204)	4055.189	(87.459)	4060.389	(87.049)

Note. FCS-SL = single-level FCS; FCS-MAN = two-level FCS with manifest cluster means; FCS-LAT = two-level FCS with latent cluster means; JM = joint modeling.

FCS using `mice` ("flat-file", FCS-SL). We used M*plus* to fit the multilevel analysis model in which students' mathematics achievement was regressed on students' gender, ESCS, and ratings on classroom management and student-teacher relations as well as the number of students and computers at school. The analysis model included latent cluster means for the ratings on classroom management and student-teacher relations as well as manifest cluster means for ESCS, centering the individual scores around the cluster-level components. The computer code and the M*plus* syntax file are provided in Supplement A of the supplemental online materials.

The results are presented in Table 4. The estimates obtained from FCS-MAN, FCS-LAT, and JM as well as their standard errors were very similar to each other. For example, the effect of the number of computers at school when confounding variables at Level 2 were controlled for was 0.197 for JM (SE = 0.090, $p = .028$), 0.198 for FCS-MAN (SE = 0.091, $p = .029$), and 0.213 for FCS-LAT (SE = 0.097, $p = .027$). Estimates of the remaining parameters were also close, and the same pattern of results was observed for these procedures. By contrast, the results obtained from FCS-SL oftentimes did not agree with the results from the other procedures, and the standard errors tended to be smaller at Level 1 and larger at Level 2. Overall, these results

illustrate that FCS-MAN, FCS-LAT, and JM may provide similar results in many applications, especially when compared with simpler methods such as single-level MI (FCS-SL).

Discussion

The goals of the present article were (a) to compare the computational procedures underlying JM and FCS for MI of missing data at Level 2, (b) to examine the different options (manifest vs. latent cluster means) for including the cluster-level components of variables at Level 1 in the imputation model for variables at Level 2, and (c) to provide recommendations for research practice by conducting an evaluation of the different procedures in a computer simulation study. We showed that JM and FCS are conceptually similar when both use latent cluster means, and we outlined a computational procedure for including latent cluster means in the FCS approach using plausible values (FCS-LAT). Using theoretical arguments, and building on the previous literature, we showed that using manifest means (FCS-MAN) is equivalent to using latent means in balanced data but produces slightly biased estimates of covariances at Level 2 in unbalanced data. In line with previous research, we found that (a) controlling for cluster size (FCS-NJ) or (b) using latent cluster means during MI (FCS-LAT and JM) provides unbiased results regardless of whether or not the cluster sizes are balanced. However, it was also evident that the bias obtained under FCS-MAN was relatively small and limited to conditions with few observations per cluster ($n = 5$), low ICCs of variables at Level 1 ($\rho_{I_y} = .10$), and extremely unbalanced data. On the basis of our findings, we believe that all three procedures provide effective tools for dealing with missing data at Level 2 in most applications in practice. Especially when compared with procedures that delete cases with missing data (LD) or ignore the multilevel structure of the data (FCS-SL), all procedures for multilevel MI provide tremendous improvements in the accuracy of parameter estimates and inferences.

Even though both JM and FCS can be used to treat missing data at Level 2, the use of FCS has often been discouraged because software solutions that iterate back and forth between variables at Level 1 and 2 while still acknowledging the cluster-level components of variables at Level 1 have not been available (e.g., Enders et al., 2016). However, the FCS procedures discussed in

this article all fulfill these requirements. Moreover, using FCS may even have advantages for applications in practice (for a comparison, see Carpenter & Kenward, 2013). Specifically, FCS-MAN allows for flexible selection of auxiliary variables and is computationally very efficient even for large data sets. At least in the context of educational research, which often features cross-sectional data with moderate ICCs and relatively large clusters, it may be argued that FCS-MAN provides a good compromise between accuracy and computational speed. In addition, it is straightforward to extend FCS-MAN to address categorical variables as well as three-level or cross-classified data structures without greatly increasing computational demands.

On the other hand, FCS-LAT can be especially useful for applications that make specific use of the latent cluster means (e.g., Croon & van Veldhoven, 2007) because their plausible values are directly added to the imputed data sets and can be treated, stored and made available in a similar way as imputations for missing data (Yang & Seltzer, 2016). In addition, the use of FCS-LAT may be advised when working with constructs that exhibit low ICCs or with samples that include a small but variable number of observations per cluster. In the present article, FCS-LAT was implemented in an "empirical Bayes" approach on the basis of a posteriori Bayesian estimates (e.g., Laird & Ware, 1982). However, it may be argued that the properties of FCS-LAT can further be improved by adopting a fully Bayesian approach that includes an additional posterior draw in the model that is used to generate plausible values for latent cluster means. Additional simulations conducted over the course of this study indicated that the efficiency and coverage properties in smaller samples improve noticeably under such an approach at the cost of only a slight increase in bias (see Supplement B in the supplemental online materials). The software implementation of FCS-LAT in the R package `miceadds` allows either of the two methods to be used (Robitzsch et al., 2017).

It is interesting that the results obtained from JM were relatively sensitive to small-sample bias. We believe that this may be due to the standard least-informative priors employed in JM. Depending on the number of variables in the model, these priors can imply variance components at Level 2 that are much larger than those that might be expected from the data (Grund, Lüdtke, & Robitzsch, 2016a; McNeish, 2016). Consequently, it may be possible to improve parameter estimates by adjusting the prior to cover a more plausible range of values (see also Schafer &

Yucel, 2002). In an additional simulation study reported in Supplement C of the supplemental online materials, we evaluated the effects of using data-dependent priors, where the priors for $\mathbf{\Psi}$ and $\mathbf{\Sigma}$ were based on empirical estimates obtained from the complete data. Using these priors strongly reduced the small-sample bias under JM, providing results similar to those of FCS, even in relatively small samples (i.e., for $J = 50$). However, note that the use of data-dependent priors is not without criticism (e.g., Gelman et al., 2014) and should not be adopted lightheartedly when there are other sources of prior information available.

As in all of research, the present study comes with several limitations and points to consider. For example, the simulation studies were based on $M = 10$ imputations. However, larger numbers of imputations are often recommended for practice (e.g., Graham et al., 2007; see also the Empirical Example). Choosing a value larger than $M = 10$ may be beneficial in terms of efficiency and coverage properties, especially in applications with large fractions of missing information (Bodner, 2008). Furthermore, the procedures for multilevel MI featured in the present study all used standard (i.e., conjugate) families of prior distributions (e.g., see Schafer & Yucel, 2002). Alternative priors have been suggested in the context of Bayesian analyses and may also improve the results obtained with MI (Barnard, McCulloch, & Meng, 2000; Gelman, 2006). Future research may choose to elaborate on the sensitivity of MI to the specification of different prior distributions, particularly under JM (see also H. Liu, Zhang, & Grimm, 2015; Schuurman, Grasman, & Hamaker, 2016).

The present study also suggests several possible extensions and topics for future research. Throughout the study, we assumed that the latent model—that is, the JM—holds in the population (see also Carpenter & Kenward, 2013; Lüdtke et al., 2017; Resche-Rigon & White, in press). However, the manifest model can often be considered "true" as well, and manifest cluster means may be the preferred choice for estimating cluster-level effects in some multilevel analysis models (Lüdtke et al., 2008). Although we expect that the procedures considered here for the treatment of missing data at Level 2 would again provide results similar to one another, future research should elaborate on the properties of estimators under each method when the manifest model holds in the population (see also Grund et al., in press-b; Mistler, 2015).

Finally, we assumed that all variables followed a multivariate normal distribution, which is

often not appropriate when working with categorical and nonnormal data. In principle, all of the procedures presented here can be applied or adapted to categorical data, for example, by defining a set of underlying latent variables (e.g., with threshold parameters or an appropriate link function) that represent different categories (Carpenter & Kenward, 2013). This approach has been implemented for multilevel JM for missing categorical data at Level 1 and 2 (Asparouhov & Muthén, 2010b; Quartagno & Carpenter, 2016a). In multilevel FCS, the same procedures as for single-level data can be used for missing data at Level 2 in conjunction with FCS-MAN (i.e., on the basis of cluster means at Level 2; see Robitzsch et al., 2017). Finally, the generation of plausible values under FCS-LAT can be adapted to categorical data by employing an appropriate model for the underlying variables at Level 1 (e.g., binary, multinomial or ordered logit). Nonnormal data can be addressed by performing MI on the basis of transformed variables (Carpenter & Kenward, 2013; He & Raghunathan, 2006; Schafer, 1997; Schafer & Olsen, 1998); however, it has also been shown that normal-distribution-based MI is fairly robust against departures from normality (e.g., Demirtas, Freels, & Yucel, 2008; von Hippel, 2013).

To summarize, we believe that the current state of statistical software offers several options for treating missing data at Level 2 in an adequate way. Especially when compared with simpler methods such as LD or single-level MI, both of which ignore important characteristics of the data, the current procedures for multilevel MI are useful and effective additions to the researcher's toolbox. Instead of arguing for the use of only one of these procedures, we believe that it is most important for researchers to be aware of the specific challenges that arise during multilevel MI and make an informed decision about which procedure best fits the structure of their data and their respective research question. Finally, we hope that the thoughts presented in this article will open up and motivate questions for future research on the treatment of missing data in multilevel studies.

Appendix

This Appendix provides additional theoretical arguments regarding the use of manifest versus latent cluster means under FCS for missing data at Level 2.

Population model

Let z_j denote the values of a centered variable at Level 2 and $\mathbf{x}_{ij} = \mathbf{u}_j + \mathbf{e}_{ij}$ denote values for a set of centered variables at Level 1 with independent components $\mathbf{u}_j$ and $\mathbf{e}_{ij}$. Then, for cluster j of size n_j we can write z_j as

$$z_j = \mathbf{u}_j\boldsymbol{\gamma} + w_j \ . \tag{A1}$$

where w_j is independent of $\mathbf{u}_j$ and $\mathbf{e}_{ij}$. Further defining $Var(\mathbf{u}_j) \equiv \mathbf{T}$, $Var(\mathbf{e}_{ij}) \equiv \boldsymbol{\Sigma}$, and $Var(w_j) \equiv \phi^2$ as well as $\boldsymbol{\sigma} \equiv \mathbf{T}\boldsymbol{\gamma}$, the joint distribution of all variables $\mathbf{y}_{ij} = (\mathbf{x}_{ij}, z_j)$ can be summarized as

$$Var(\mathbf{y}_{ij}) = \begin{pmatrix} \mathbf{T} + \boldsymbol{\Sigma} & \boldsymbol{\sigma}^T \\ \boldsymbol{\sigma} & \boldsymbol{\gamma}^T\mathbf{T}\boldsymbol{\gamma} + \phi^2 \end{pmatrix} \ . \tag{A2}$$

We introduce the following notation for further development. Specifically, we define a probability distribution for the cluster sizes n_j independent of $\mathbf{x}_{ij}$ and z_j, where $\mathcal{S}$ denotes the set of unique cluster sizes, so that $P(n_j = k) \equiv \pi_k$ with $0 \leq \pi_k \leq 1$ for all $k \in \mathcal{S}$ and $\sum_{k\in\mathcal{S}} \pi_k = 1$. We assume that z is partially missing, $z = (z^{\text{mis}}, z^{\text{obs}})$, with probability α whereas $\mathbf{x}$ is observed. For simplicity, we omit superscripts for $\mathbf{x}$ where possible. We further assume (a) that the number of clusters approaches infinity ($J \to \infty$), so that posterior variances become zero, and (b) that z_j is MCAR so that α is independent of $\mathbf{x}_{ij}$ and n_j. With no loss of generality, we assume that the first J_0 clusters have z_j missing, the other J_1 observed ($j = 1, \ldots, J_0, J_0 + 1, \ldots, J$), where the proportion $\frac{J_0}{J}$ of missing values in z converges to α as the sample size goes to infinity (i.e., $\lim_{J\to\infty} \frac{J_0}{J} = \alpha$).

FCS with latent cluster means (FCS-LAT)

To generate imputations z_j^{imp}, a regression model on the basis of the latent cluster means ($\mathbf{u}_j$) of $\mathbf{x}_{ij}$ can be used. To show that the joint distribution of $\mathbf{y}_{ij}$ is preserved during MI, one must show that the distribution of the completed data $\mathbf{y}_{ij}^{\text{com}} = (\mathbf{y}_{ij}^{\text{imp}}, \mathbf{y}_{ij}^{\text{obs}})$ including z_j^{imp} is identical to Equation A2. As argued by van Buuren (2012) and Hughes et al. (2014), sampling from a sequence of univariate conditional normal distributions is equivalent to sampling from a joint multivariate normal distribution. This can be applied to the joint distribution $P(\mathbf{x}_{ij}, \mathbf{u}_j, z_j)$ with unknown $\mathbf{u}_j$ and missing z_j by sampling from the following conditional distributions

$$\begin{aligned} u_{jp}^{\text{imp}} &\sim P(u_{jp}|\mathbf{x}_{ij}, \mathbf{u}_{j(-p)}, z_j) \\ z_j^{\text{imp}} &\sim P(z_j|\mathbf{x}_{ij}, \mathbf{u}_j)\,, \end{aligned} \tag{A3}$$

with the notation as defined in the main text. The conditional distributions can further be simplified as $P(z_j|\mathbf{x}_{ij}, \mathbf{u}_j) = P(z_j|\mathbf{u}_j)$ because z_j is conditionally independent of $\mathbf{x}_{ij}$ given $\mathbf{u}_j$. Consequently, under FCS-LAT, imputations z_j^{imp} are generated from the conditional model

$$z_j^{\text{imp}} = \mathbf{u}_j^{\text{imp}}\,\boldsymbol{\gamma} + w_j^{\text{imp}}\,, \tag{A4}$$

where estimates of $\boldsymbol{\gamma}$ and ϕ^2 are obtained from the observed data, and posterior draws for $\mathbf{u}_j^{\text{imp}}$ are obtained as described in the main text (e.g., using the plausible value approach by Mislevy, 1991). This is sufficient because (a) all u_{jp} are conditionally independent of $\mathbf{x}_{ij}$ given z_j and $\mathbf{u}_{j(-p)}$, and (b) z_j is conditionally independent of $\mathbf{x}_{ij}$ given $\mathbf{u}_j$ (see above). As a result, the model in Equation A4 is consistent with Equations A2 and A3, and FCS-LAT on the basis of $\mathbf{u}_j^{\text{imp}}$ is consistent with drawing imputations directly from the joint model (Equation A2).

FCS with manifest cluster means (FCS-MAN)

Alternatively, the imputation model can be based on the manifest cluster means ($\bar{\mathbf{x}}_{\bullet j} = \mathbf{u}_j + \bar{\mathbf{e}}_{\bullet j}$) of $\mathbf{x}_{ij}$, and imputations can be generated from the following equation

$$z_j^{\text{imp}} = \bar{\mathbf{x}}_{\bullet j}\boldsymbol{\beta} + \epsilon_j^{\text{imp}}\,, \tag{A5}$$

where the ϵ_j^{imp} are distributed normally with mean zero and variance $Var(\epsilon_j^{\text{imp}})$. In general, the regression coefficients in the manifest ($\boldsymbol{\beta}$) and latent imputation model ($\boldsymbol{\gamma}$) do not coincide (Croon & van Veldhoven, 2007). The regression coefficients in Equation A5 are estimated as

$$\hat{\boldsymbol{\beta}} = \left[\frac{1}{J_1}\sum_{j=J_0+1}^{J} \bar{\mathbf{x}}_{\bullet j}^T\bar{\mathbf{x}}_{\bullet j}\right]^{-1}\left(\frac{1}{J_1}\sum_{j=J_0+1}^{J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right)\,. \tag{A6}$$

Note that $E(\bar{\mathbf{x}}_{\bullet j}^T\bar{\mathbf{x}}_{\bullet j}) = \mathbf{T} + \frac{1}{n_j}\boldsymbol{\Sigma}$ and $E(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}) = \mathbf{T}\boldsymbol{\gamma}$. Then, as the number of clusters goes to infinity ($J \to \infty$), the expected value of $\hat{\boldsymbol{\beta}}$ can then be expressed as

$$E(\hat{\boldsymbol{\beta}}) \overset{J\to\infty}{=} \left[\sum_{k\in\mathcal{S}} \pi_k(\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma})\right]^{-1}\mathbf{T}\boldsymbol{\gamma}\,. \tag{A7}$$

In the special case with balanced data with a constant cluster size $n_j = k_0$, it is further worth noting that $E(\bar{\mathbf{x}}_{\bullet j}^T \bar{\mathbf{x}}_{\bullet j}) = (\mathbf{T} + \frac{1}{k_0}\mathbf{\Sigma})$, in which case Equation A7 reduces to

$$E(\hat{\boldsymbol{\beta}}) \overset{J\to\infty}{=} (\mathbf{T} + \tfrac{1}{k_0}\mathbf{\Sigma})^{-1}\mathbf{T}\boldsymbol{\gamma}\,, \tag{A8}$$

Carpenter and Kenward (2013) showed that for the case with balanced data, that the conditional independence of z_j and $\mathbf{x}_{ij}$ also holds given $\bar{\mathbf{x}}_{\bullet j}$, that is, $P(z_j|\mathbf{x}_{ij}) = P(z_j|\bar{\mathbf{x}}_{\bullet j})$, so that FCS-MAN would be consistent with the joint model (Equation A2). However, it may be expected that this no longer holds in the general, unbalanced case (see also Resche-Rigon & White, in press). In the following, we show which aspects of the joint distribution are preserved under FCS-MAN in balanced and unbalanced data.

Variance of z. The fact that the variance of z is unbiased can easily be shown with the decomposition of variance in the linear model. Let $\hat{z}_j = \bar{\mathbf{x}}_{\bullet j}\hat{\boldsymbol{\beta}}$. Under the given assumptions, it holds that $Var(\hat{z}_j^{\text{mis}}) = Var(\hat{z}_j^{\text{obs}})$ and $Var(\epsilon_j^{\text{imp}}) = Var(\epsilon_j^{\text{obs}})$. As a result, $Var(z_j^{\text{imp}}) = Var(\hat{z}_j^{\text{mis}}) + Var(\epsilon_j^{\text{imp}}) = Var(z_j)$, showing that the variance of $z_j^{\text{com}} = (z_j^{\text{obs}}, z_j^{\text{imp}})$ is unbiased.

Estimators of the covariance of $\mathbf{x}$ *and* z. To elaborate on the estimation of the covariance, we focus on maximum likelihood (ML) estimation. However, because the standard ML estimator cannot be expressed in closed form in the general case with unbalanced data, we study Muthén's ML estimator (MUML; B. O. Muthén, 1990). The MUML estimator ($\hat{\boldsymbol{\sigma}}$) allows estimating $\boldsymbol{\sigma}$ in closed form and can be expressed as

$$\hat{\boldsymbol{\sigma}} = \frac{1}{J}\sum_{j=1}^{J}\frac{n_j}{c_J}\bar{\mathbf{x}}_{\bullet j}^T z_j\,, \tag{A9}$$

where $c_J = \left[(\sum_{j=1}^J n_j)^2 - \sum_{j=1}^J n_j^2\right]\left[\sum_{j=1}^J n_j(J-1)\right]^{-1}$ is a function of the cluster sizes with $\lim_{J\to\infty} c_J = \bar{n}_\infty = \sum_{k\in\mathcal{S}}\pi_k \cdot k$ (i.e., the average cluster size). In complete data, $\hat{\boldsymbol{\sigma}}$ is identical to the ML estimator in the case with balanced data (B. O. Muthén, 1990) and remains an asymptotically ($J \to \infty$) unbiased estimator of $\boldsymbol{\sigma}$ in the unbalanced case (Yuan & Hayashi, 2005). In balanced data with cluster size $n_j = k_0$, the estimator reduces to

$$\hat{\boldsymbol{\sigma}} = \frac{1}{J}\sum_{j=1}^{J}\bar{\mathbf{x}}_{\bullet j}^T z_j\,. \tag{A10}$$

In the following, we use this estimator to show the potential bias in estimating $\boldsymbol{\sigma}$ from the completed data z_j^{com}, where imputations z_j^{imp} have been generated under FCS with manifest cluster means.

Covariance of $\mathbf{x}$ *and* z *in balanced data.* In balanced data with cluster size $n_j = k_0$, the covariance in Equation A10 is estimated on the basis of both the observed and imputed data $z_j^{\text{com}} = (z_j^{\text{obs}}, z_j^{\text{imp}})$ as follows

$$\hat{\boldsymbol{\sigma}} = \frac{1}{J}\left(\sum_{j=1}^{J_0} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j=J_0+1}^{J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right). \tag{A11}$$

The expected value of $\hat{\boldsymbol{\sigma}}$ can then be expressed as

$$E(\hat{\boldsymbol{\sigma}}) = E\left[\frac{1}{J}\left(\sum_{j=1}^{J_0} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j=J_0+1}^{J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right)\right] = E\left(\frac{J_0}{J}\right) E\left(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}\right) + E\left(\frac{J_1}{J}\right) E\left(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right). \tag{A12}$$

In the limit of $J \to \infty$, it holds that $E(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}) \overset{J\to\infty}{=} E(\bar{\mathbf{x}}_{\bullet j}^T \bar{\mathbf{x}}_{\bullet j}) E(\hat{\boldsymbol{\beta}})$. Then, by further noting that $E(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}) = \mathbf{T}\boldsymbol{\gamma}$ as before and by plugging in Equation A8, it can be shown that Equation A12 converges to

$$E(\hat{\boldsymbol{\sigma}}) \overset{J\to\infty}{=} \alpha\left(\mathbf{T} + \tfrac{1}{k_0}\boldsymbol{\Sigma}\right)\left[\mathbf{T} + \tfrac{1}{k_0}\boldsymbol{\Sigma}\right]^{-1}\mathbf{T}\boldsymbol{\gamma} + (1-\alpha)\,\mathbf{T}\boldsymbol{\gamma} = \mathbf{T}\boldsymbol{\gamma} = \boldsymbol{\sigma}\,, \tag{A13}$$

which shows that $\hat{\boldsymbol{\sigma}}$ is asymptotically unbiased in balanced data.

Covariance of $\mathbf{x}$ *and* z *in unbalanced data.* In the general case with unbalanced data, the potential bias in $\hat{\boldsymbol{\sigma}}$ is more difficult to evaluate because (a) the cluster sizes included in Equation A9 complicate calculations, and (b) the z_j^{imp} are not independent of $\mathbf{e}_{ij}$ under FCS-MAN as would be the case in complete data (Croon & van Veldhoven, 2007). Instead, we follow the law of total expectation by averaging over the conditional expectations with fixed cluster sizes $n_j = k$. Let $\hat{\boldsymbol{\sigma}}_k$ denote the value of $\hat{\boldsymbol{\sigma}}$ for clusters of size $n_j = k$. By conditioning on cluster size, we also obtain balanced subsets of the data, in which we can use Equation A10 instead of A9. Consequently, $\hat{\boldsymbol{\sigma}}_k$ can be expressed as

$$\hat{\boldsymbol{\sigma}}_k = \hat{\boldsymbol{\sigma}}_{|n_j=k} = \frac{1}{J_{(k)}}\left(\sum_{j\in\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j\in\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right), \tag{A14}$$

where $\mathcal{J}_{0(k)}$ and $\mathcal{J}_{1(k)}$ denote two sets of clusters with size k and with missing and observed z_j, respectively, $J_{0(k)}$ and $J_{1(k)}$ denote the number of clusters therein, and $J_{(k)}$ denotes the total

number of clusters of size k. By noting that $E(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}) \overset{J\to\infty}{=} E(\bar{\mathbf{x}}_{\bullet j}^T \bar{\mathbf{x}}_{\bullet j}) E(\hat{\boldsymbol{\beta}})$ as before and by plugging in Equation A7, the expected value of $\hat{\boldsymbol{\sigma}}_k$ can be written as

$$\begin{aligned} E(\hat{\boldsymbol{\sigma}}_k) &= E\left[\frac{1}{J_{(k)}}\left(\sum_{j\in\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j\in\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right)\right] = E\left(\frac{J_{0(k)}}{J_{(k)}}\right) E\left(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}\right) + E\left(\frac{J_{1(k)}}{J_{(k)}}\right) E\left(\bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right) \\ &\overset{J\to\infty}{=} \alpha\left(\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}\right)\left[\sum_{k'\in\mathcal{S}} \pi_{k'}\left(\mathbf{T} + \tfrac{1}{k'}\boldsymbol{\Sigma}\right)\right]^{-1} \mathbf{T}\boldsymbol{\gamma} + (1-\alpha)\,\mathbf{T}\boldsymbol{\gamma}\,, \end{aligned} \tag{A15}$$

where $k' \in \mathcal{S}$ is used to denote all cluster sizes besides and including k. This expression is generally not equal to $\boldsymbol{\sigma}$ unless $\mathbf{T} + \frac{1}{k}\boldsymbol{\Sigma} = \left[\sum_{k'\in\mathcal{S}} \pi_{k'}\left(\mathbf{T} + \frac{1}{k'}\boldsymbol{\Sigma}\right)\right]^{-1}$.

In the full data set, $\hat{\boldsymbol{\sigma}}$ is again based on both the observed and imputed data, z_j^{obs} and z_j^{imp}, and can be written as

$$\hat{\boldsymbol{\sigma}} = \frac{1}{J}\left(\sum_{j=1}^{J_0} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j=J_0+1}^{J} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right) \tag{A16}$$

The expected value of $\hat{\boldsymbol{\sigma}}$ can then be expressed as

$$\begin{aligned} E(\hat{\boldsymbol{\sigma}}) &= E\left[\frac{1}{J}\left(\sum_{j=1}^{J_0} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}} + \sum_{j=J_0+1}^{J} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right)\right] \\ &= E\left(\frac{J_0}{J}\right) E\left(\sum_{k\in\mathcal{S}} \frac{J_{0(k)}}{J_0} \frac{k}{c_J} \frac{1}{J_{0(k)}} \sum_{\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}\right) + E\left(\frac{J_1}{J}\right) E\left(\sum_{k\in\mathcal{S}} \frac{J_{1(k)}}{J_1} \frac{k}{c_J} \frac{1}{J_{1(k)}} \sum_{\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right) \\ &= E\left(\frac{J_0}{J}\right) \sum_{k\in\mathcal{S}} kE\left(\frac{J_{0(k)}}{J_0}\right) E\left(\frac{1}{c_J}\right) E\left(\frac{1}{J_{0(k)}} \sum_{j\in\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{imp}}\right) \\ &\qquad + E\left(\frac{J_1}{J}\right) \sum_{k\in\mathcal{S}} kE\left(\frac{J_{1(k)}}{J_1}\right) E\left(\frac{1}{c_J}\right) E\left(\frac{1}{J_{1(k)}} \sum_{j\in\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^T z_j^{\text{obs}}\right), \end{aligned} \tag{A17}$$

which illustrates the contribution of the conditional expectations given in Equation A15. In the limit as $J \to \infty$, this expression converges to

$$\begin{aligned} E(\hat{\boldsymbol{\sigma}}) &\overset{J\to\infty}{=} \alpha \sum_{k\in\mathcal{S}} \frac{k}{\bar{n}_\infty} \pi_k\left(\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}\right)\left[\sum_{k'\in\mathcal{S}} \pi_{k'}\left(\mathbf{T} + \tfrac{1}{k'}\boldsymbol{\Sigma}\right)\right]^{-1} \boldsymbol{\sigma} + (1-\alpha) \sum_{k\in\mathcal{S}} \frac{k}{\bar{n}_\infty} \pi_k\, \boldsymbol{\sigma} \\ &= \alpha\left(\sum_{k\in\mathcal{S}} \frac{k}{\bar{n}_\infty} \pi_k\left(\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}\right)\right)\left[\sum_{k\in\mathcal{S}} \pi_k\left(\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}\right)\right]^{-1} \boldsymbol{\sigma} + (1-\alpha)\,\boldsymbol{\sigma}\,. \end{aligned} \tag{A18}$$

This expression is generally not equal to $\boldsymbol{\sigma}$. Consequently, the asymptotic bias of $\hat{\boldsymbol{\sigma}}$ as an estimator of $\boldsymbol{\sigma}$ can be expressed as

$$
\begin{aligned}
Bias(\hat{\boldsymbol{\sigma}}) &= \alpha \left\{ \sum_{k \in \mathcal{S}} \frac{k}{\bar{n}_\infty} \pi_k (\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}) \left[\sum_{k \in \mathcal{S}} \pi_k (\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}) \right]^{-1} - 1 \right\} \boldsymbol{\sigma} \\
&= \alpha \left\{ \sum_{k \in \mathcal{S}} \left(\frac{k}{\bar{n}_\infty} - 1 \right) \pi_k (\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}) \right\} \left[\sum_{k \in \mathcal{S}} \pi_k (\mathbf{T} + \tfrac{1}{k}\boldsymbol{\Sigma}) \right]^{-1} \boldsymbol{\sigma} \,,
\end{aligned}
\tag{A19}
$$

which is not generally zero in unbalanced data. Because the expected value of $\hat{\boldsymbol{\sigma}}$ converges with that of the ML estimator as the number of cluster becomes large ($J \to \infty$), we expect that regression coefficients obtained under FCS-MAN should be biased as well.[6]

[6]It is interesting that unbiased estimates of $\boldsymbol{\sigma}$ might be obtained under FCS-MAN with an estimator that does not weight by cluster size, which can be seen by plugging in $\frac{k}{\bar{n}_\infty} = 1$ into Equation A19. However, because such an estimator is unlikely to perform well in general, this is left as a topic for future research.

Article 4: Pooling ANOVA results from multiply imputed datasets: A simulation study

Grund, S., Lüdtke, O., & Robitzsch, A. (2016b). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*, *12*, 75–88. doi:10.1027/1614-2241/a000111

The analysis of variance (ANOVA) is frequently used to examine whether a number of groups differ on a variable of interest. The global hypodissertation test of the ANOVA can be reformulated as a regression model in which all group differences are simultaneously tested against zero. Multiple imputation offers reliable and effective treatment of missing data; however, recommendations differ with regard to what procedures are suitable for pooling ANOVA results from multiply imputed datasets. In this article, we compared several procedures (known as D_1, D_2 and D_3) using Monte Carlo simulations. Even though previous recommendations have advocated that D_2 should be avoided in favor of D_1 or D_3, our results suggest that all procedures provide a suitable test of the ANOVA's global null hypodissertation in many plausible research scenarios. In more extreme settings, D_1 was most reliable, whereas D_2 and D_3 suffered from different limitations. We provide guidelines on how the different methods can be applied in one- and two-factorial ANOVA designs and information about the conditions under which some procedures may perform better than others. Computer code is supplied for each method to be used in freely available statistical software.

The analysis of variance (ANOVA) is a popular method for analyzing data in many fields of psychology and the social sciences (Cohen, Cohen, West, & Aiken, 2003; Maxwell & Delaney, 2004). One of the major goals of an ANOVA is to examine whether a number of groups (e.g., demographic features, experimental conditions) differ with respect to some variable of interest. The global null hypothesis, according to which all groups stem from the same population, is tested by comparing the portions of variance that reside between and within groups. Under the null hypothesis, the ratio of the mean squares between and within groups follows an F distribution. If group differences are reasonably large compared with individual differences, the global null hypothesis is rejected, and groups are believed to differ with respect to the variable of interest.

Missing data are a pervasive problem in the social sciences. Deleting the missing values (e.g., listwise deletion) is an easy but inefficient way of dealing with missing data that can seriously distort statistical analyses (Little & Rubin, 2002). Other techniques such as multiple imputation (Rubin, 1987) promise a more reliable and efficient treatment of missing data (Schafer & Graham, 2002). Multiple imputation (MI) draws a number of M replacements for

the missing values from their posterior predictive distribution, given the observed data and a statistical model. The completed datasets are then analyzed using regular complete-data methods, and the parameter estimates are pooled according to the rules described in Rubin (1987) to form final parameter estimates and inferences.

Rubin's rules are easily applied to one-dimensional estimands such as means or regression coefficients, but multidimensional estimands (e.g., comparing multiple groups in the ANOVA's F test) call for different methods. Several such methods are discussed in the literature, and clear recommendations can be found in various books and articles (Little & Rubin, 2002; Marshall, Altman, Holder, & Royston, 2009; Reiter & Raghunathan, 2007; Schafer, 1997). However, some authors' conclusions are less than definite and they emphasize the need for further research concerning realistic applications of these methods (Enders, 2010; Snijders & Bosker, 2012b; van Buuren, 2012). In addition, previous studies have often focused on a technical understanding of these methods without considering specific research designs. Using computer simulations, we compared several pooling methods for the F test in one- and two-factorial ANOVA designs. We examined the robustness of these methods as well as the conditions under which some methods may be more trustworthy than others. We attempted to complement the existing literature with simulation results that can be easily applied to research practice. Computer code is given for each method to be used in freely available software.

Pooling ANOVA results

The one-factorial ANOVA can be reformulated as a regression model in which the outcome variable is regressed on a number of dummy variables that represent the membership in a group i ($i = 1, \ldots, I$). For I groups, the group membership can be coded by $K = I - 1$ dummy variables such that the regression coefficients reflect differences between groups. In complete datasets, the Wald test of the K-dimensional vector of regression coefficients (without the intercept) is equivalent to testing the ANOVA's null hypothesis that there are no differences between groups (e.g., Cohen et al., 2003). Over the past years, several methods have become available for carrying out multiparameter hypothesis tests in multiply imputed datasets (e.g., Enders, 2010; Little & Rubin, 2002; Schafer, 1997; van Buuren, 2012). These methods build on different

aspects of the completed-data analyses and thus differ in behavior and ease of application. Here, we provide a brief overview of the procedures featured in our study, illustrated for the one-factorial ANOVA. The procedures extend naturally to two-factorial designs, with effect coding instead of dummy coding.

Moment based statistics (D_1 and D_1^*). The D_1 procedure extends Rubin's rules to multidimensional estimands such as the K-dimensional vector of regression coefficients in the ANOVA. Using D_1, the vectors of regression coefficients and their associated covariance matrices are pooled across the imputed datasets. Given a set of coefficient vectors $\hat{Q}_m$ ($m = 1, \ldots, M$) and estimates of their sampling covariance matrix $\hat{U}_m$, the D_1 statistic reads

$$D_1 = \frac{(\bar{Q} - Q_0)^T \bar{U}^{-1} (\bar{Q} - Q_0)}{K(1 + \text{ARIV}_1)} , \tag{1}$$

where $K = I - 1$ is the number of regression coefficients that represent group differences, $\bar{Q}$ and $\bar{U}$ are the average point and covariance estimates, and Q_0 is the vector of regression coefficients expected under the null hypothesis. The ARIV_1 denotes the average relative increase in variance due to nonresponse, that is, the extent to which the sampling variance of the estimator has increased due to missing data

$$\text{ARIV}_1 = \frac{(1 + M^{-1})\text{tr}(B\bar{U}^{-1})}{K} , \tag{2}$$

where B is the covariance matrix of the estimates $\hat{Q}_m$ across the imputed datasets (see Enders, 2010, for an illustration). The ARIV is conceptually related to the fraction of missing information (FMI; Rubin, 1987), which denotes the portion of the total sampling variance of an estimator that is due to missing data[1]. Rubin (1987) and Li, Raghunathan, and Rubin (1991) derived an F reference distribution for D_1, along with K numerator and v_1 denominator degrees of freedom. For $a = K(M - 1)$, the denominator degrees of freedom are calculated as

$$v_1 = \begin{cases} 4 + (a - 4)[1 + (1 - 2a^{-1})\,\text{ARIV}_1^{-1}]^2 & \text{if } a > 4 \\ (K + 1)(M - 1)(1 + \text{ARIV}_1^{-1})^2/2 & \text{otherwise} \end{cases} . \tag{3}$$

In its original formulation, the degrees of freedom for D_1 were derived under the assumption of infinite complete-data degrees of freedom. Reiter (2007) proposed a correction formula

[1]Estimates of the FMI were based on estimates of the ARIV such that FMI = ARIV/(1+ARIV).

that adjusts the denominator degrees of freedom v_1 for finite samples. The resulting test is henceforth called D_1^*. Calculating D_1 and D_1^* requires pooling the point and variance estimates across datasets, a task that is relatively simple and well documented (see Enders, 2010).

The D_1 procedure is frequently recommended in the literature (e.g., Allison, 2001; Enders, 2010; Graham, 2012; Little & Rubin, 2002; Schafer, 1997; van Buuren, 2012). Li, Raghunathan, and Rubin (1991) showed that D_1 is reliable and robust unless the FMI is very large and variable across parameters. Reiter (2007) showed that D_1^* produced accurate Type I error rates even in small samples. Licht (2010) proposed an adjustment of D_1 and replicated the favorable results of Li, Raghunathan, and Rubin (1991) for finite samples and larger K. van Ginkel and Kroonenberg (2014) illustrated the use of D_1^* in empirical datasets. However, simulation results regarding D_1 and D_1^* are still relatively scarce, and van Buuren (2012) suggests evaluating them "in more general settings" (p. 157). Enders (2010) found it "difficult to assess the trustworthiness of the D_1 statistic in realistic research scenarios" (p. 236).

p *values from Wald-like hypothesis tests (D_2).* Li, Meng, et al. (1991) developed a test statistic that is computed from a series of Wald tests (or their p values, equivalently) rather than from point and variance estimates. This is especially useful if K is large or variance estimates (e.g., standard errors) are not available. Given a number of M Wald-like test statistics W_m, the D_2 statistic reads

$$D_2 = \frac{\overline{W}K^{-1} - (M+1)(M-1)^{-1}\text{ARIV}_2}{1 + \text{ARIV}_2} , \tag{4}$$

where $\overline{W}$ is the average test statistic across datasets and K is again the number of parameters that represent group differences. The ARIV_2 is another estimate of the average relative increase in variance that is based solely on the individual test statistics W_m

$$\text{ARIV}_2 = (1 + M^{-1}) \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\sqrt{W}_m - \overline{\sqrt{W}} \right)^2 \right] , \tag{5}$$

where $\overline{\sqrt{W}}$ denotes the average $\sqrt{W}_m$ across the imputed datasets (see Enders, 2010). Li, Meng, et al. (1991) proposed an F reference distribution for D_2 with K numerator and v_2 denominator degrees of freedom

$$v_2 = K^{-3/M}(M-1)(1 + \text{ARIV}_2^{-1})^2 . \tag{6}$$

In order to apply D_2, the individual test statistics (W_m) should follow a χ^2 distribution. Hence, in ANOVA models, the F values for all datasets (F_m) must be transformed such that $W_m = KF_m$, each of which approach a χ^2 distribution as the denominator degrees of freedom go to infinity. The D_2 statistic is easily calculated by pooling the test statistics across datasets. No specialized software or programming skills are required in order to calculate D_2, and only the M test statistics from the imputed datasets must be entered into the formulae, which are routinely included in the output of most statistical software.

However, the literature often advises against D_2. Li, Meng, et al. (1991) suggested that it be used only as a rough guide because its Type I error rates can be too high or too low depending on the FMI. It is usually recommended that D_1 be used whenever possible because D_2 is less precise, less powerful, and only loosely correlated with the "more nearly optimal" D_1 (Schafer, 1997, p. 116; Enders, 2010; Little & Rubin, 2002). Nonetheless, D_2 has been acknowledged for its ease of implementation because it operates directly on the test statistics (e.g., Allison, 2001; Snijders & Bosker, 2012b). Van Buuren (2012) advised that D_2 may be used if nothing but the test statistics are available but that D_2 is "considerably less reliable" than other pooling methods (p. 159).

Pooled likelihood-ratio tests (D_3). Coming from the perspective of model comparison, hypotheses about a set of parameters can be tested using likelihood-ratio tests (LRTs). The D_3 procedure was developed by Meng and Rubin (1992) to enable LRTs with multiply imputed datasets. The procedure does not require variance estimates; instead, it operates on the likelihood. Meng and Rubin (1992) showed that it is not sufficient to simply combine the individual LRT statistics L_m into an average $\bar{L}$. In addition, the LRT statistic needs to be evaluated at the average estimates of the model parameters for all imputed datasets. The D_3 statistic reads

$$D_3 = \frac{\tilde{L}}{K(1 + \text{ARIV}_3)}\,, \tag{7}$$

where $\tilde{L}$ is the mean LRT statistic across the imputed datasets evaluated at the average parameter estimates, and K is the number of parameters being tested. Estimating the ARIV_3 includes the two pooled LRTs evaluated at the individual and pooled estimates, respectively (see Enders, 2010)

$$\text{ARIV}_3 = \frac{M+1}{K(M-1)}(\bar{L} - \tilde{L}) \; . \tag{8}$$

According to Meng and Rubin (1992), the F reference distribution for D_3 has K numerator and v_3 denominator degrees of freedom. For $a = K(M-1)$,

$$v_3 = \begin{cases} 4 + (a-4)[1 + (1-2a^{-1})\,\text{ARIV}_3^{-1}]^2 & \text{if } a > 4 \\ (K+1)(M-1)(1+\text{ARIV}_3^{-1})^2/2 & \text{otherwise} \end{cases} \; . \tag{9}$$

Calculating D_3 can be tedious because it requires that users have access to the likelihood function and that it is possible to evaluate it at user-defined values. Due to its complexity, the procedure is not frequently used, but it has been implemented in likelihood-oriented software such as M*plus* (Asparouhov & Muthén, 2008), SAS (Mistler, 2013b) and the `semTools` package for R (Pornprasertmanit, 2014). The D_3 statistic is frequently recommended when D_1 cannot be calculated, that is, in the absence of standard errors (Little & Rubin, 2002; van Buuren, 2012). It has been argued that D_1 and D_3 should behave similarly, and more reliably than D_2, because the two are approximately equal (Meng & Rubin, 1992; Schafer, 1997). However, Enders (2010) pointed out that "virtually no research studies have compared the two test statistics" (p. 241).

Present study

Even though recommendations regarding D_1, D_1^*, D_2 and D_3 can be found in the literature, the behavior of these methods is still not fully understood. Earlier studies focused on the general properties of these methods, and simulation studies considered the FMI as a pivotal point (e.g., Li, Meng, et al., 1991; Li, Raghunathan, & Rubin, 1991). Their usual recommendation is that, in general, some procedures should be preferred (D_1, D_1^*, D_3), while others should be avoided (D_2). However, in the present article, we argue that all of these methods provide suitable tests for ANOVA models in most conditions that are encountered in psychological research. We conducted computer simulations that explore their performance from the perspective of practical research. Our results are intended to complement the existing literature with results that can be easily applied to practical research, and to assist researchers in their statistical decision making.

We examined the Type I error rates and the statistical power of the four pooling methods. Study 1 features a fully crossed simulation design in which the number of groups, the group

size, the effect size, the missing data mechanism, and the amount of information available from an auxiliary variable were varied. This design allowed us to examine possible interactions between the simulation factors. However, in order to reduce computational effort, some of its conditions had to be restricted. The conditions were chosen to mimic what frequently occurs in applications of the ANOVA in psychological research. Two additional studies were conducted that relaxed some of the restrictions made in Study 1. This made it possible to examine specific findings in greater detail. Study 2a provides details on how including an auxiliary variable into the imputation model may influence statistical power (Collins et al., 2001). For this purpose, we varied the correlation between outcome and auxiliary variable in very fine steps, thus exploring the conditions in which the ANOVA might benefit from using MI. In Study 2b, we examined the effects of larger FMIs on the Type I error rates, that is, for larger amounts of missing data and given different amounts of auxiliary information. In this context, we elaborate on the "link" between the simulation factors and the FMI in our simulation design. This was deemed helpful for judging the severity of missing data problems in research practice and for providing a reference frame for the results of earlier studies. Study 3 extends the paradigm of Study 1 to two-factorial ANOVA designs. In the two-factorial design we took special interest in testing the overall interaction effect, which, especially in large ANOVA designs, may involve a large number of parameters.

Study 1

The first simulation study was conducted to assess the performances of D_1, D_1^*, D_2, and D_3 under conditions that are commonly encountered in one-factorial ANOVA designs. All simulation factors were fully crossed in order to examine the factors that drive the performance of these methods.

Simulation procedure

Data generating model. The ANOVA provided the foundation for the data generating model. A continuous outcome Y was simulated from a normal distribution given the group

means μ_i for a factor A with groups $i = 1, \ldots, I$, that is,

$$Y = \mu_i + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma_\epsilon^2), \tag{10}$$

where σ_ϵ^2 denotes the variance within groups. According to Cohen (1988), the variance of the group means around the population mean (i.e., the grand mean) $\bar{\mu}$ can be defined as

$$\sigma_A^2 = \frac{\sum_{i=1}^{I}(\mu_i - \bar{\mu})^2}{I}. \tag{11}$$

The sum of the two variances (σ_A^2 and σ_ϵ^2) was defined to be one. The population mean was assumed to be zero. Differences between groups were simulated according to Cohen's (1988) f, here

$$f_A = \frac{\sigma_A}{\sigma_\epsilon}\ . \tag{12}$$

Thus, the two variances followed as

$$\sigma_A^2 = \frac{f_A^2}{1 + f_A^2} \quad \text{and} \quad \sigma_\epsilon^2 = 1 - \frac{f_A^2}{1 + f_A^2}\ . \tag{13}$$

Different patterns of group means were simulated in order to mimic plausible research scenarios. This was achieved by rephrasing all group means as $\mu_i = p_i d_A$, where the p_i form a pattern of group means $p_A = (p_1, \ldots, p_I)$ that sums to zero, and d_A is a scaling factor that enlarges this pattern so that it would imply the correct portions of variance as given by Equation 13. The scaling factor d_A was derived by rearranging Equation 11, which yields

$$d_A = \sigma_A \sqrt{\frac{I}{\sum_{i=1}^{I} p_i^2}}\ . \tag{14}$$

We simulated two patterns of group means labeled "difference" and "trend," respectively, in which either one third of the groups differed greatly from the others or all groups differed in such a way that they formed a linear trend. For example, with $I = 3$ groups, the two patterns can be written $p_{A,\text{difference}} = (-1/2, 1, -1/2)$ and $p_{A,\text{trend}} = (-1, 0, 1)$, respectively. To illustrate, suppose we wanted to establish an effect of size $f_A = .40$ forming a difference pattern $p_A = (-1/2, 1, -1/2)$. This implies a variance of group means $\sigma_A^2 = 0.16/1.16 = 0.14$; thus, the scaling factor would become $d_A = 0.37\sqrt{3/(0.25 + 1 + 0.25)} = 0.53$. Finally, the group means μ_i would be $(-0.26, 0.53, -0.26)$.

Table 1: Simulation Design of the Different Simulation Studies

Design conditions	Study 1	Study 2a	Study 2b	Study 3
Group size (n)	25, 50, 100	25	25, 100	10, 30, 50
Levels of A and B (I, J)	3, 6, 12	12	12	3×3, 5×5
Main effect A (f_A)	0, .10, .25, .40	.25	0	0, .10, .25
Main effect B (f_B)	–	–	–	0
Interaction effect (f_{AB})	–	–	–	0, .10, .25
Effect patterns	difference, trend	difference	difference	difference
Correlation XY (ρ_{xy})	0, .35, .70	0, .05, . . . , .95	0, .20, .35, .50, .70, .90	0, .35, .70
MD effect of X (λ)	0, .35, .70	0	0	0, .35, .70
MD probability	25%	25%, 50%	5%, 10%, . . . , 80%	25%
Number of Imputations	5, 10, 20, 50, 100	100	100	5, 10, 20, 50, 100

Note. The correlation ρ_{xy} and the MD probability were varied in steps of .05 and 5% in Studies 2a and 2b, respectively. MD = missing data.

A second continuous variable X was simulated to allow for different missing data mechanisms and to mimic situations in which auxiliary information can be included in the imputation model. The covariate X was simulated as

$$X = \rho_{xy} Y + \epsilon_X \quad \text{with} \quad \epsilon_X \sim N(0, 1 - \rho_{xy}^2) \,, \tag{15}$$

where ρ_{xy} denotes the correlation between X and Y. Table 1 provides an overview of the simulation design of all studies. In Study 1, we varied the number of groups ($I = 3, 6, 12$), the sample size within each group ($n = 25, 50, 100$), the effect size ($f_A = 0, .10, .25, .40$), and the correlation between X and Y ($\rho_{xy} = 0, .35, .70$).

Imposition of missing values. Missing data were imposed on the outcome Y, whereas the covariate X and the group membership of each person were fully observed. Different missing data mechanisms were defined according to Rubin (1976). In this classification, the hypothetical complete data Y are divided into observed and unobserved portions, Y_{obs} and Y_{mis}, respectively. An indicator variable R denotes which values in Y are observed. Rubin (1976) introduced several broad classes of missing data mechanisms. If the missing values are simply a random sample of the hypothetical completely observed Y, then the values are missing completely at random (MCAR), that is, $P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R)$. If the chance of observing Y depends on the observed data but does not further depend on the missing part, then the values are missing at random (MAR), that is, $P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R|Y_{\text{obs}})$. The two are often called *ignorable* missing

data because the exact missing data mechanism need not be known in order to perform MI. Treating nonignorable missing data requires making strong assumptions about the missing data mechanism and thus was not considered in this study (see Carpenter & Kenward, 2013).

The missing values were simulated using a latent response variable R^*, which determined whether values in Y were missing dependent on the covariate X under a linear model

$$R^* = \lambda X + \epsilon_{R^*} \quad \text{where} \quad \epsilon_{R^*} \sim N(0, 1 - \lambda^2) \, . \tag{16}$$

Values in Y were set missing if $R^* < z$, where z is a quantile of the standard normal distribution according to the desired probability of missing data (e.g., $z = -0.67$ for 25% missing data). As presented in Table 1, we varied the effect of X on the latent response indicator to simulate different missing data mechanisms. For Y to be MCAR, we set $\lambda = 0$, and for Y to be MAR given X, we set $\lambda = .35$ or .70. The probability of missing data was held constant at 25% but was varied in Study 2b. Note that our simulation design implicitly varies the FMI by varying population and sample characteristics that influence the FMI. This is in contrast to previous studies, in which the FMI was varied explicitly (e.g., Li, Meng, et al., 1991; Li, Raghunathan, & Rubin, 1991; Licht, 2010). As mentioned before, the simulation design was chosen to mimic situations that are encountered in real-world applications of the ANOVA. Thus, we manipulated the severity of the missing data problem in terms of the design factors (e.g., amount of missing data, presence of auxiliary variables) rather than the FMI. This perspective was chosen so that the simulation design would directly relate to research practice, whereas the FMI would occur only insofar as it emerged from the simulated conditions.

Imputation and analysis. Imputations were carried out using the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011) in the statistical software R (R Core Team, 2014). The "norm" imputation method was used; therefore, missing values on Y were assumed to be normally distributed given the group membership and the covariate X. Following recent recommendations, we created $M = 100$ imputed datasets for each simulated dataset (Bodner, 2008; Graham et al., 2007). However, all analyses were repeated with different subsets of M, that is, with the first 5, 10, 20, and 50 of the total 100 datasets, respectively (see Table 1). The ANOVA model was fitted by dummy coding the grouping variable and regressing the outcome Y on the $K = I - 1$

dummy variables. All methods—D_1, D_1^*, D_2, and D_3—were implemented in the software R. The computer code is provided in the supplemental online material along with an example application to artificial data (see also Grund, Lüdtke, & Robitzsch, 2016a). In addition, listwise deletion (LD) was included as a strategy for handling missing data because it is still frequently used in research practice.

We compared the pooling methods with respect to Type I error rates and their power to detect nonzero effects. The Type I error rate is the relative frequency with which the null hypothesis is rejected when the population effect (f) is zero. Ideally, the Type I error rate should be close to the predefined significance level α (e.g., 5% or 1%). A procedure was considered liberal or conservative when its Type I error rate was higher or lower, respectively, than the nominal α. Bradley (1978) suggested a criterion for robustness, according to which Type I error rates within $\alpha \pm 0.5\alpha$ are considered acceptable (e.g., within 2.5% and 7.5% for $\alpha = 5\%$). In addition, we calculated the Type I error rates for the complete datasets (i.e., before imposing missing values) to provide a benchmark for the different pooling methods. The statistical power is the relative frequency with which the null hypothesis is rejected when the population effect is *not* zero. Assessing differences in statistical power is difficult because the expected power is not a fixed value for all conditions (Cohen, 1988). Thus, the expected power itself served as a benchmark for the pooling methods.

Results

The first study featured six simulation factors and 648 conditions in total. All conditions were replicated 10,000 times to ensure that the Type I error rates and the power to detect nonzero effects had stabilized. Reporting all results was not feasible due to the large number of conditions and because not all factors influenced the performance of the pooling methods. The complete results for $M = 100$ imputations are provided in the supplemental online material, intended as a repository for interested readers. We focus on the "difference" pattern of group means, and assume a level of $\alpha = 5\%$ throughout this section. The results were similar for $\alpha = 1\%$ and will be discussed whenever necessary.

Type I error rate. In all conditions and for all pooling methods, the Type I error rates varied

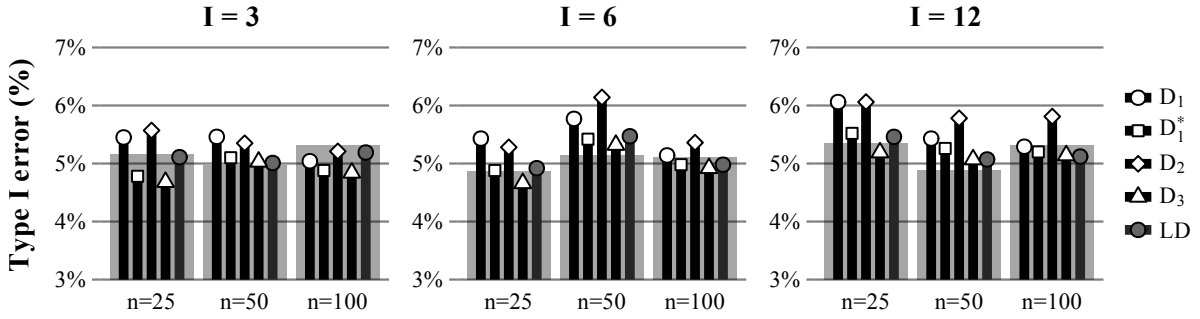

Figure 1: Type I error rates for different pooling methods and LD (α = 5%) depending on group size (n) and number of groups (I), given MCAR data ($\lambda = 0$) with no auxiliary information ($\rho_{xy} = 0$). The grey boxes indicate the Type I error rates obtained from complete datasets. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

within a reasonable range, that is, below 6.1% (D_2) and above 4.2% (D_3). Thus, no violations of Bradley's criterion for robustness were observed at α = 5%. Some methods were found to be liberal in some cases (D_1 and D_2), whereas others were slightly conservative (D_1^* and D_3). The extent to which the pooling methods were conservative or liberal was mostly influenced by the group size (n) and the number of groups (I). Figure 1 illustrates the Type I error rates of all procedures for different group sizes and different numbers of groups, when the correlation between X and Y ($\rho_{xy} = 0$) and the effect of X on missingness ($\lambda = 0$) were held constant.

The D_1 statistic was slightly liberal in small samples (i.e., small n or I) but otherwise provided nearly optimal results. The error rates obtained with D_1^* were nearly optimal under all conditions. D_2 was the most liberal of all pooling methods, but even for D_2, the Type I error rates were not seriously inflated. Contrary to D_1, however, D_2 remained somewhat liberal in larger samples, especially when the number of groups was large. Finally, D_3 produced reasonable Type I error rates but was somewhat conservative if the number of groups was large and the groups were relatively small (e.g., $I = 12$ and $n = 25$). Results obtained with LD were generally close to the ideal solutions and usually close to those obtained with D_1^*.

With increasing group size, the Type I error rates of the four pooling methods became more similar; that is, D_1 and to a lesser extent D_2 became less liberal, whereas D_3 became less conservative. Effects of the number of groups were more diverse because an increase in I increased both the sample size and the number of parameters of the global null hypothesis test. For D_1, D_1^*, and D_3, an increase in I led to more conservative results. Type I error rates for D_1^*

Table 2: Power to Detect Nonzero Effects (α=5%) for all Pooling Methods and LD

	λ = 0					λ = .70				
	LD	D_1	D_1^*	D_2	D_3	LD	D_1	D_1^*	D_2	D_3
	n = 25, I = 12, f_A = .25 (PE = .836)									
ρ_{xy} = 0	.683	.687	.669	.692	.658	.675	.680	.664	.685	.649
ρ_{xy} = .35	.675	.697	.682	.702	.674	.662	.698	.679	.711	.668
ρ_{xy} = .70	.677	.761	.747	.756	.749	.630	.758	.744	.762	.745
	n = 50, I = 3, f_A = .25 (PE = .780)									
ρ_{xy} = 0	.644	.646	.635	.649	.634	.645	.646	.635	.648	.631
ρ_{xy} = .35	.649	.671	.660	.668	.658	.631	.654	.644	.660	.640
ρ_{xy} = .70	.640	.704	.694	.704	.695	.611	.713	.704	.720	.703

Note. PE = power expected; n = group size; I = number of groups; f_A = size of main effect A; ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

and D_3 sometimes fell below those obtained from complete datasets and below the nominal α. D_2 on the other hand remained somewhat liberal for larger values of I unless the group size was very large in comparison (e.g., I = 12 and n = 25).

A lower level of α = 1% did not change the picture as a whole; that is, all pooling methods performed similarly when compared with one another. Bradley's criterion demands that Type I error rates vary within 0.5% and 1.5% in this case. Type I error rates could be as high as 1.5% (D_1) for smaller groups, thus violating Bradley's criterion for robustness, but they were usually close to the nominal value in larger samples (see the supplemental online material).

Statistical power. Assessing the power of the pooling methods entailed certain limitations due to floor and ceiling effects, that is, when the power approached 5% and 100%, respectively. Differences between methods were found to be consistent regardless of effect size, but naturally, these became smaller when the power approached its upper or lower bounds. Especially for large effects (f_A = .40), choosing a particular method became less important because the power was effectively 100% for all methods unless the samples were very small. Therefore, we will focus on small and moderate effect sizes (f_A = .10 and .25) in order to describe the results on a scale that is informative and meaningful for applied researchers (power between 60% and 80%).

The more liberal methods (D_1 for smaller samples, D_2) also scored highest in statistical power. Most importantly, the power obtained with MI was higher than with LD whenever the

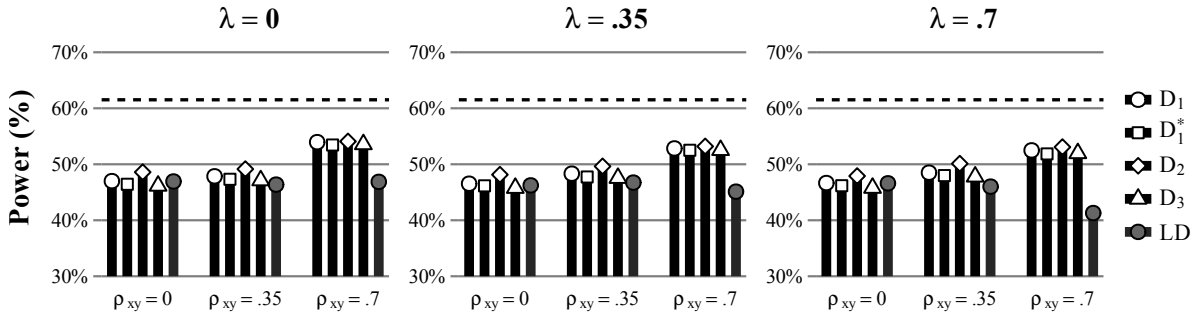

Figure 2: Power to detect nonzero main effect (f_A = .10) in larger samples (n = 100, I = 12) depending on the missing data mechanism. The expected power is indicated by a dashed line. ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

covariate X was somewhat informative about the missing values on Y, where a higher correlation between X and Y (ρ_{xy}) led to higher power when MI was used. The effect of X on missingness (λ) did not greatly influence the power by itself but moderated the aforementioned effects such that higher values of λ intensified the differences between LD and MI (see Collins et al., 2001).

Figure 2 illustrates the interplay of the correlation between X and Y and the effect of X on missingness in larger samples (n = 100, I = 12, f_A = .10). All pooling methods and LD were equally capable of detecting nonzero effects when the covariate carried no information about the missing outcome (ρ_{xy} = 0). As soon as the covariate provided information (ρ_{xy} = .35 or .70), higher statistical power was observed when MI was used. Similar results were obtained for moderate samples with small and large groups, as presented in Table 2. For small groups (n = 25, I = 12, f_A = .25), the more liberal pooling methods (D_1 and D_2) provided higher statistical power. With larger groups (n = 50, I = 3, f_A = .25), the difference between the pooling methods became smaller. Again, higher power was observed for MI when the covariate provided information about the missing Y. The conservative methods had lower power in general and thus relied more heavily on such information. Nonetheless, even the conservative methods had higher power than LD, given sufficient auxiliary information.

Number of imputed datasets. The number of imputations was varied within each simulation condition in order to provide an insight into how the results would have changed if fewer than M = 100 imputations had been used. The initial recommendation that M = 5 imputations would suffice for most applications of MI (Rubin, 1987) has been modified in the past by several

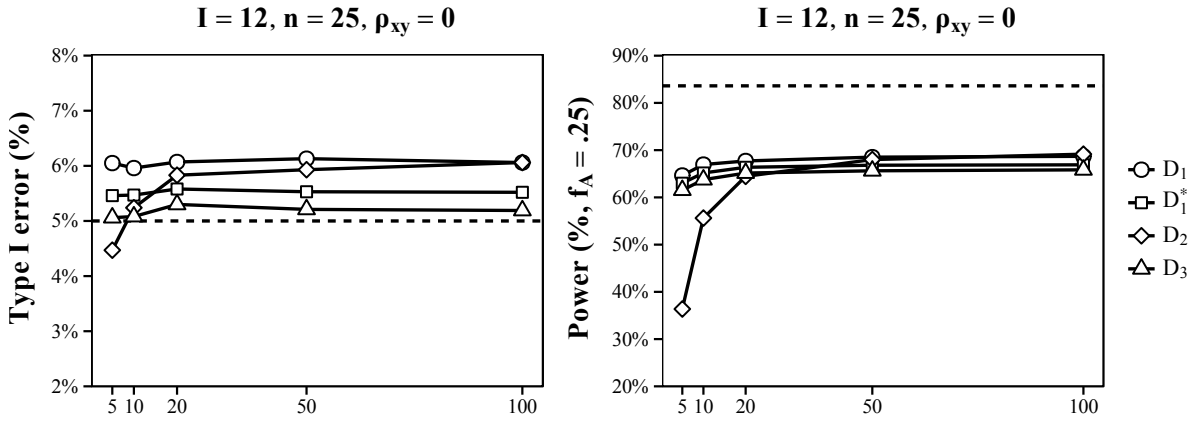

Figure 3: Type I error and statistical power of all pooling methods depending on the numbers of imputations. I = number of groups; n = group size; ρ_{xy} = correlation between X and Y; f_A = size of main effect A; D_1, D_1^*, D_2, D_3 = pooling methods.

authors (e.g., Bodner, 2008; Graham et al., 2007; Harel, 2007).

Interpreting the effect of different values of M proved to be challenging because its effect depended on the group size, the number of groups, the correlation between X and Y, and also differed between pooling methods. Figure 3 shows the results for different M in selected conditions. The results obtained with D_1, D_1^*, and D_3 were relatively insensitive to the number of imputations but were best when M was at least 20. For D_2, however, the performance changed substantially when more than 20 imputations were generated: Type I error rates from D_2 became slightly higher, and statistical power was much larger with M > 20, especially when the number of groups was large and the covariate X did not provide information about the missing Y ($\rho_{xy} = 0$). With fewer imputations ($M \leq 20$), D_2 tended to be conservative and suffered from a substantial loss of power. With a sufficient number of imputations, the power of the four methods was almost identical.

Discussion

The first simulation study compared different pooling methods for testing the global null hypothesis of the ANOVA with multiply imputed datasets. Differences emerged in terms of Type I error rates: Some methods tended to be slightly liberal (D_1 and D_2) or conservative (D_1^* and D_3), but no procedure led to Type I error rates far above or below the nominal value. The liberal methods also tended to detect nonzero effects more frequently. The biggest difference,

however, emerged between MI and LD when a covariate that provided information about the missing values was included in the imputation model. In such cases, using MI could be highly beneficial whereas potential losses from using MI when the covariate carried no information were not observed (see Collins et al., 2001).

Our study was able to replicate previous findings on the performances of D_1 and D_1^*, which were found to be stable and reliable in most cases (Li, Raghunathan, & Rubin, 1991; Reiter, 2007). Although seldom recommended, D_2 provided very reasonable results within the scope of the first study. Moreover, our results suggest that D_2 is equally powerful as D_1 and D_1^* when the number of imputations is sufficiently large. This is in stark contrast to current recommendations regarding D_2, which suggest that D_2 should generally be avoided because it was optimized for $M = 3$ imputations, less powerful than D_1, and unlikely to improve with larger M (Schafer, 1997; van Buuren, 2012). Our findings suggest that, due to its ease of application, D_2 might be a viable alternative in many applications of multiparameter tests, such as in the ANOVA, despite being theoretically less convincing than D_1. The D_3 procedure also provided good results but was unnecessarily conservative in small samples. Given that D_3 is rather difficult to implement, we believe that D_1 and D_1^* are better choices for ANOVA models unless researchers intend to use likelihood-based statistical software that already offers D_3 (see Enders, 2010). Care should be taken when the pooling methods are applied under more extreme conditions. The D_2 procedure was slightly more liberal when the number of groups I (and hence the number of parameters) was large. In such cases, D_3 was quite conservative unless the groups were very large in comparison.

Several limitations are noteworthy. First, due to the large simulation design, not all factors could be varied in very great detail. The simulation suggested that MI benefits when information about the missing values is available, but, at this point, it is unclear how much information a covariate must provide in order to be helpful. Thus, the purpose of Study 2a was to explore the potential gains in statistical power. Second, we chose a fixed value for the probability of missing data. The chosen value of 25% is quite large for many applications of the ANOVA, but the number of missing values can sometimes be higher depending on how the data were collected (e.g., Graham et al., 2006). Especially D_2 has been shown to be sensitive to very small and

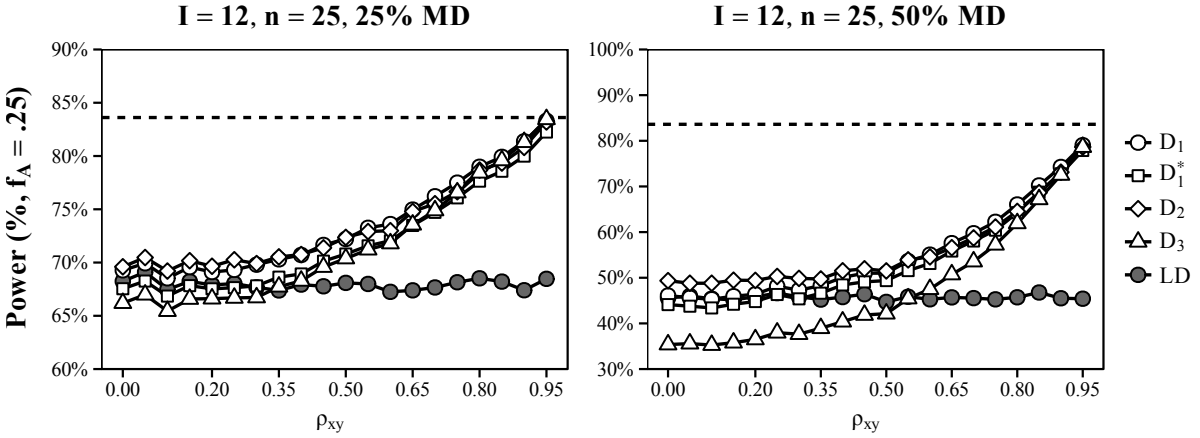

Figure 4: Power to detect main effect for all pooling methods and LD depending on the correlation between X and Y and the amount of missing data. The expected power is indicated by a dashed line. I = number of groups; n = group size; f_A = size of main effect A; ρ_{xy} = correlation between X and Y; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

large values of the FMI (Li, Meng, et al., 1991). In order to close the gap between our results and the existing literature, it must be elaborated upon how the amount of missing data and the presence of auxiliary variables influence the FMI and, as a result, the robustness of the pooling methods. This was the purpose of Study 2b. Finally, Study 1 was limited to one-factorial ANOVA designs. Therefore, Study 3 was conducted, which extended the paradigm of Study 1 to two-factorial ANOVA designs and the test of interaction effects.

Study 2a

To examine the effects of including a more or less useful covariate in the imputation model, we varied the correlation between X and Y in steps of .05, ranging from $\rho_{xy} = 0$ to $\rho_{xy} = .95$. Either 25% or 50% missing values were introduced into the dataset. The remaining factors were held constant, as shown in Table 1. One hundred imputations were created. These values were chosen to reflect practical research but also to avoid influences of sampling error and boundary conditions. The results were cross-checked for different conditions, but the main pattern of results was found to be comparable.

Figure 4 shows the statistical power to detect moderate effects ($f_A = .25$) for all pooling methods and LD as a function of the strength of the relationship between the covariate and the outcome (ρ_{xy}). The performance of the pooling methods differed only when the correlation

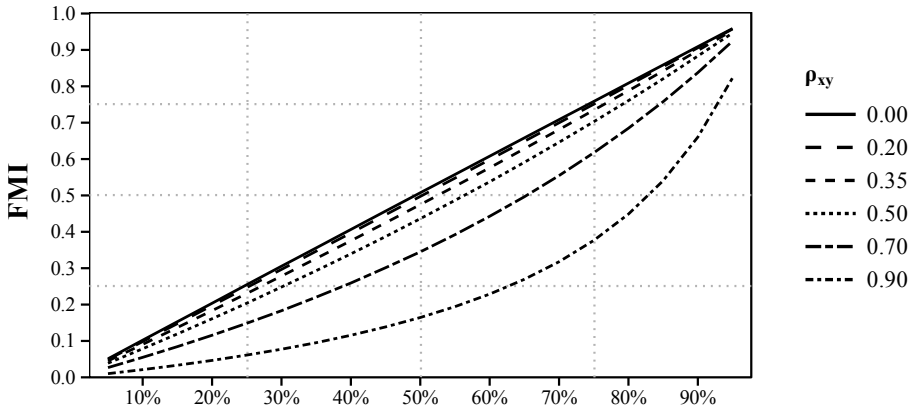

Figure 5: Estimates of the FMI obtained from D_1 in larger samples (n = 100, I = 12) with zero main effect (f_A = 0) depending on the amount of missing data and the correlation between X and Y. FMI = fraction of missing information; ρ_{xy} = correlation between X and Y.

was small and became increasingly similar as the correlation grew larger. This is not surprising because the FMI was largest when X and Y were uncorrelated (see Study 2b). Listwise deletion was comparably powerful as long as X was only weakly correlated with Y. For larger values of the correlation (ρ_{xy} = .35 and above), the pooling methods consistently outperformed LD in terms of statistical power. Whereas the advantages of using MI remained modest for correlations below .50, larger correlations greatly improved statistical power. When 50% of the data were missing, the differences between the pooling methods grew larger, especially when X and Y were only weakly correlated. In this case, D_2 appeared to be more powerful than D_1 and D_1^*, and D_3 appeared to be less powerful, essentially reflecting differences in Type I error rates.

This illustrates that the conclusions of Study 1 cannot be generalized to arbitrarily harsh conditions, and that more severe missing data problems must be met with more sophisticated methods (e.g., D_1 or D_1^*). Previous research has expressed these conditions in terms of the FMI. In Study 2b, we elaborate on how the FMI is related to the amount of missing data and auxiliary information, and how one's assessment of the missing data problem may guide one's choice among the pooling methods.

Study 2b

The FMI in our study was influenced by the amount of missing data and the correlation between X and Y. Figure 5 illustrates the relationship between these measures in our study. If auxiliary

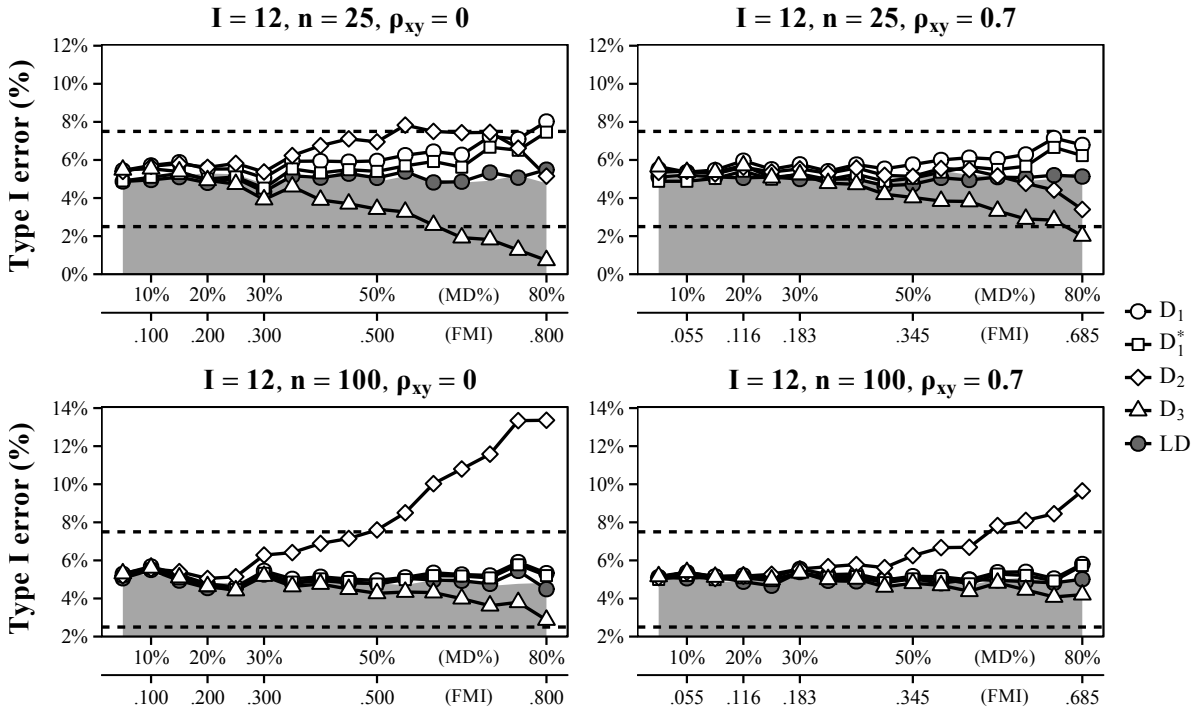

Figure 6: Type I error rates of all pooling methods and LD in moderate and larger samples dependent on the amount of missing data and the correlation between X and Y. The grey area indicates the Type I error obtained from complete datasets. I = number of groups; n = group size; ρ_{xy} = correlation between X and Y; FMI = fraction of missing information; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

information was not available ($\rho_{xy} = 0$), then the FMI was equal to the amount of missing data. Therefore, the FMI could be manipulated directly when $\rho_{xy} = 0$ by varying the missing data rate. However, if the covariate is predictive of the missing values, then the FMI is lowered depending on the strength of that relationship. In other words, the missing data problem becomes less severe the more information can be included into the imputation model. In Study 1, the missing data rate was fixed to 25%, which is already quite large for many applications of the ANOVA. As can be seen from Figure 5, this corresponds to an FMI of only .25 if $\rho_{xy} = 0$, or less if $\rho_{xy} = .35$ or .70. In earlier studies, values for the FMI up to .50 were often considered (see Figure 4). In Study 2b, we investigated the effects of the FMI more thoroughly by including different portions of missing data, ranging from 5% to 80% in increments of 5%, as well as different values for the correlation of X and Y, effectively varying the FMI between .03 and .80 (see Table 1). Type I error rates were calculated for each condition.

Figure 6 shows the Type I error rates for all methods in smaller and larger samples, given different amounts of missing data and a more or less useful covariate. D_1 and D_1^* were

robust even when large portions of data were missing and when the covariate did not provide information about the missing data. In such extreme cases, as predicted by Li, Meng, et al. (1991), D_2 was less reliable, and increasingly liberal in larger samples. The results remained acceptable for up to 50% missing data, at which point Bradley's liberal criterion for robustness was violated (FMI of .50). However, if the correlation between X and Y was large, then D_2 was more reliable, and the results remained acceptable for up to 65% missing data (also FMI of .50). Notice that, in smaller samples, D_2 became less liberal again when the amount of missing data became very large (above 70%)[2]. Surprisingly, D_3 was also affected by larger FMIs such that, for large amounts of missing data (above 40%), D_3 became more and more conservative. These results occurred most strongly in smaller samples, where results remained acceptable for up to 65% missing data when X provided no information about Y. This effect too became smaller as the correlation between X and Y grew larger.

Study 3

The third study was conducted in order to assess whether our results could be generalized to two-factorial ANOVA designs and, in particular, to tests of the interaction effect. For this purpose, we extended the procedure of Study 1 to two-factorial designs in which two factors A and B, with I and J levels, respectively, could influence the outcome Y. The two main effects and the interaction effect were each assigned an effect pattern, denoted p_A, p_B and p_{AB}, respectively, and an effect size, denoted f_A, f_B and f_{AB}, respectively. The difference pattern was employed for the two main effects. The interaction effect was defined in a similar fashion such that groups on the main diagonal of the $I \times J$ design would have larger values in Y compared to the off-diagonal groups. Scaling factors for each pattern were derived by the same logic as in Study 1. We chose similar values for the remaining simulation factors, as can be seen in Table 1. We examined the interaction effect in a 3×3 and 5×5 design with a different number of

[2] The behavior of D_2 for large FMIs appeared to be a result of two compensatory mechanisms. Liberal behavior of D_2 was associated with F-values slightly larger than 1. The inflation of F values was associated with values of the ARIV_2 that were lower than the respective ARIV_1 (see Equation 4), especially in larger samples. Conservative behavior of D_2, on the other hand, seemed to be induced by the denominator degrees of freedom, v_2, which tended to be smaller than v_1, and noticeably so in smaller samples (see the first term in Equation 6; cf. Equations 3 and 9).

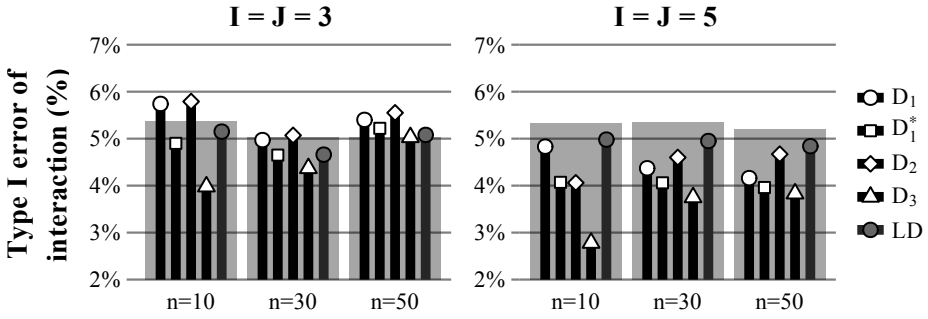

Figure 7: Type I error rates for different pooling methods and LD (α = 5%) for the interaction effect in the two-factorial design, depending on group size (n) and number of groups per factor ($I = J$). The grey boxes indicate the Type I error rates obtained from complete datasets. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

persons per group. Since the total sample size increased rapidly with I and J, we simulated smaller groups of size 10, 30 and 50, respectively, so that the range in total sample size was similar to Study 1.

The results for the main effects were consistent with those of Study 1. Therefore, we only report our findings concerning the interaction effect, that is, the Type I error rates if $f_{AB} = 0$ (α = 5%) and the power to detect nonzero interaction effects, given that the main effects were both zero. The test of the interaction effect involved 4 parameters in the 3×3 design and 16 parameters in the 5×5 design. Larger designs were not considered because they are rarely found in practice.

Figure 7 shows the Type I error rates obtained from the different pooling methods and LD when all effects are zero, and $\rho_{xy} = 0$ as well as $\lambda = 0$. For moderate ($n = 30$) and larger groups ($n = 50$) all methods were found to be robust. As in Study 1, the Type I error rates of D_1 and D_2 were slightly above those of D_1^* and D_3. For smaller groups ($n = 10$), D_1 and D_2 were found to be somewhat liberal in the 3×3 design (5.7% and 5.8%) and slightly conservative in the 5×5 design (4.8% and 4.1%), whereas D_1^* and D_3 performed conservatively in both cases (4.9% and 4.0% in the 3×3 design; 4.1% and 2.8% in the 5×5 design, respectively).

Similar differences were observed for the power to detect nonzero interaction effects, as is shown in Table 3. For smaller groups ($n = 10$, $I = J = 5$), D_1 and D_2 had greater power to detect nonzero interaction effects, whereas D_1^* and especially D_3 were less powerful; a pattern that was most pronounced if the covariate X did not provide information about the missing data

Table 3: Power to Detect Nonzero Interaction Effect (α=5%) in a Two-Factorial Design for all Pooling Methods and LD

	λ = 0					λ = .70				
	LD	D_1	D_1^*	D_2	D_3	LD	D_1	D_1^*	D_2	D_3
	n = 10, I = J = 5, f_{AB} = .25 (PE = .653)									
ρ_{xy} = 0	.489	.447	.415	.427	.355	.488	.438	.407	.435	.340
ρ_{xy} = .35	.488	.464	.430	.447	.394	.476	.448	.415	.458	.362
ρ_{xy} = .70	.474	.530	.500	.504	.508	.470	.543	.505	.546	.516
	n = 30, I = J = 3, f_{AB} = .25 (PE = .922)									
ρ_{xy} = 0	.803	.805	.802	.808	.796	.817	.815	.809	.822	.800
ρ_{xy} = .35	.807	.818	.809	.814	.806	.800	.817	.810	.820	.806
ρ_{xy} = .70	.797	.870	.863	.871	.866	.791	.880	.877	.888	.878

Note. PE = power expected; n = group size; I = number of groups by factor A; J = number of groups by factor B; f_{AB} = size of interaction effect; ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

(ρ_{xy} = 0). For moderate groups (n = 30, I = J = 3), the differences between the pooling methods were much smaller. In comparison with LD, and in larger samples, the power of the pooling methods was low if the covariate did not provide information about the missing values (ρ_{xy} = 0), but higher than with LD if the covariate was predictive of the missing data (ρ_{xy} = .70). In smaller samples, such low power was only observed for D_3. The missing data mechanism did not influence the power obtained with the pooling methods, but LD showed lower power if Y was MAR (λ = .70).

General discussion

By means of Monte Carlo simulation, we examined the performance of different pooling methods for the global null hypothesis test of the ANOVA with multiply imputed datasets. The goal of the present article was to complement the existing literature with simulation results that argue from the perspective of applied researchers. Similar to previous studies, we can conclude that D_1 and D_1^* are the most reliable pooling methods available and that D_3 behaves similarly in larger samples. However, we found that the use of D_2, at least for hypothesis tests in the ANOVA, is perfectly supported by many conditions that commonly occur in research practice. All pooling methods provided large potential gains over LD in terms of statistical power when a useful

covariate could be included in the imputation model, provided that the number of imputations was sufficiently large. Whereas the increase in statistical power depended on the presence of useful covariate information, there was usually no harm in using MI when the covariate did not provide any information at all. We hope that the simulation approach taken in this study will aid researchers in judging the severity of the missing data problem and in choosing the procedure which is the most fitting for their purpose.

In general, D_1 and D_1^* provided the most reliable hypothesis tests for the ANOVA, which replicates what previous studies concluded about D_1. Their Type I error rates varied within a small range around the optimal value, and reasonable gains in statistical power arose from including auxiliary variables. The slightly liberal behavior of D_1 was limited to small samples. Both methods appeared to be reliable even when large portions of data were missing.

The D_2 procedure performed well in Study 1 and Study 3. We observed similar gains in statistical power for D_1 and D_2, but the power of D_2 was much lower if the number of imputed datasets was not large enough. However, unlike previous research suggested, the power of D_2 improved drastically when the number of imputations was increased and was ultimately equal to that of D_1 (cf. Schafer, 1997; van Buuren, 2012). In line with previous research, we found that when the FMI was large, the robustness of D_2 was compromised (see Li, Meng, et al., 1991). Our simulation study suggests that researchers should refrain from using D_2 if large portions of the data are missing and no auxiliary variables can be included to compensate for the loss of information (e.g., 50% missing data, low correlation with other variables); the more information is supplied by covariates, the more missing data may be tolerated by D_2. All in all, the D_2 statistic appeared to be a reasonable choice for most applications of the ANOVA in psychological research. This is an encouraging result for applied researchers because D_2 is very easy to calculate using the test statistics alone, without requiring specialized software or programming experience.

The likelihood-based D_3 procedure performed well in most conditions, but it was quite conservative unless the samples were very large. This behavior was intensified if large portions of the data were missing. In general, the D_3 statistic can be recommended; however, at least for hypothesis tests in the ANOVA, larger gains in statistical power can be obtained using D_1, D_1^*

and D_2, which are often easier to implement.

Our results also have important implications for applications of MI in which large portions of the data are missing, for example, in "planned missing data" designs (Graham et al., 2006). In such designs, both MI and LD provide approximately unbiased parameter estimates because the data are MCAR. However, based on our results, it seems crucial that "planned missing data" designs include auxiliary variables which are correlated with the variables of interested, thus providing more informed imputations of missing values. Otherwise, analyses based on MI will be no more efficient than those based on LD (see Rhemtulla, Savalei, & Little, 2016). In other words, hypothesis tests based on MI can be much more powerful than those based on LD, but only if useful covariates are available that can be included in the imputation model.

As is true for any computer simulation, our study was limited in a number of ways. First, the complex simulation design limited the number of levels that could be studied for each factor in a fully crossed manner. In Study 1, we fixed the probability of missing data to 25%, and most other factors had a small number of levels. We addressed this problem by varying some simulation factors in two additional studies to explore their effects in better detail. Nonetheless, not all conditions were fully crossed, and therefore our results should not be generalized too readily to the vast diversity of conditions that can occur in practical research. Second, likelihood-based methods may be considered, which offer some advantages over LD, for example, to include auxiliary variables or to condition on possible causes of missing data (see Enders, 2010; Little & Rubin, 2002; von Hippel, 2007). Third, we assumed the covariate and the grouping variable to be fully observed at all times. This is often unlikely in practice, in which case, more general missing data methods must be considered (e.g., Enders, 2008; Little, 1992). Even though imputation itself was of minor interest in our study, results may differ for multivariate missing data problems. Finally, there are further alternatives to the four pooling methods considered here and they should be subjects of future research. Raghunathan and Dong (2011) proposed a pooling method which is solely based on the sum of squares. Variations and applications of D_1 and D_3 have been considered by Licht (2010), Kientoff (2011), and Consentino and Claeskens (2010).

In future studies, researchers may wish to address ANOVA designs with multiple or nested

factors, interaction effects, repeated measurements, or random effects (see van Ginkel & Kroonenberg, 2014). Effect size measures should be considered to allow for a more exhaustive treatment of missing data in ANOVA designs (Harel, 2009). However, the procedures featured in our study are not limited to the ANOVA. In structural equation modeling, researchers may utilize the same procedures that are featured in our study for various multiparameter tests with multiply imputed data (see Enders, 2010, for an overview). We believe that all pooling methods have good potential for reliable and efficient statistical inference when faced with missing data. The computer code for these methods is provided in the supplemental online material. We encourage researchers to use and extend these methods to thereby promote a wider application of missing data methods in psychology and the social sciences.

Article 5: Multiple imputation of multilevel missing data: An introduction to the R package pan

Grund, S., Lüdtke, O., & Robitzsch, A. (2016a). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open*, *6*(4), 1–17. doi:10.1177/2158244016668220

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. In order 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.

In recent years 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 in order 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 in order to ensure valid statistical inferences in subsequent multilevel analyses (Black et al., 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 paper 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, 2012b). In order 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 paper 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 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 nonstandard 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, G. 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-

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

	MA	RA	CA	SES	DPM	DPR	SC
MA		0.528	0.530	0.232	−0.234	−0.238	−0.217
RA			0.493	0.299	−0.291	−0.294	−0.327
CA				0.240	−0.265	−0.251	−0.221
SES					−0.154	−0.155	−0.123
DPM						0.782	0.399
DPR							0.419
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.

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 et al. (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 and reading achievement, a measure of cognitive ability, 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 reading achievement and cognitive ability 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 were missing for half of the sample due to a planned missing data design (Graham et al., 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

Table 2: Frequent Missing Data Patterns

Pattern	MA	RA	CA	SES	DPM	DPR	SC	Cases #	Rel. %	Cum. %
1	o	o	o	o	x	o	o	2306	26.3%	26.3%
2	o	o	o	o	o	o	o	2134	24.3%	50.6%
3	o	o	o	x	x	o	o	1173	13.4%	64.0%
4	o	o	o	x	o	o	o	1125	12.8%	76.9%
5	x	o	o	o	x	x	x	1027	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. o = observed; x = missing; 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.

25% of the sample (Patterns 3 and 4), SES was additionally missing. The remaining patterns were more diverse, and data were missing for math achievement scores, disciplinary problems in reading classes, or school climate. 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' math achievement scores (MA). In addition, we included SES in order 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 in order to separate within-group from between-group effects (see Enders & Tofighi, 2007; Raudenbush & Bryk, 2002). For student i in class j,

$$MA_{ij} = \beta_0 + \beta_1(DPM_{ij} - \overline{DPM}_j) + \beta_2\overline{DPM}_j + \beta_3(SES_{ij} - \overline{SES}_j) + \beta_4\overline{SES}_j + \upsilon_{0j} + \epsilon_{ij} \; . \quad (1)$$

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 accounts for the students' individual deviations from that average. For example, β_4 denotes the effect of a class' average SES on class-

level math achievement, whereas β_3 denotes the effect of students' individual deviations from the class average on their individual math achievement scores. The student- and class-level residuals are each assumed to follow a normal distribution with zero mean and variances $Var(\upsilon_{0j})$, independently and identically across classes, and $Var(\epsilon_{ij})$, independently and identically across students.

Example 2: Random-slope model

Our second model of interest examined the relationship between students' cognitive ability (CA) and their math achievement 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 in order to "reintroduce" the group-level construct into the model. The model reads

$$\begin{aligned} MA_{ij} = {} & \beta_0 + \beta_1(CA_{ij} - \overline{CA}_j) + \beta_2\overline{CA}_j + \beta_3(SES_{ij} - \overline{SES}_j) + \beta_4\overline{SES}_j \\ & + \upsilon_{0j} + \upsilon_{1j}(CA_{ij} - \overline{CA}_j) + \epsilon_{ij}\,, \end{aligned} \tag{2}$$

where υ_{1j} denotes the random effect of cognitive ability on math achievement 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.

Multiple imputation of incomplete multilevel data

Missing data could be addressed by restricting the analyses to completely observed cases (listwise deletion). 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). Multiple imputation 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 $\mathbf{Y}$ which is decomposed into observed and unobserved portions $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$. An indicator matrix $\mathbf{R}$ denotes whether values are observed or missing. If the missing data are simply a random sample of the hypothetical complete data, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R})$, 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, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{\text{obs}})$, then the data are missing at random (MAR). The two missing data mechanisms MCAR and MAR are often called "ignorable" because the exact missing data mechanism need not be known in order to perform MI (for a more general discussion of the role of "ignorability", see Enders, 2010). If neither condition holds, that is, $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, 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). In order 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 missing values themselves, without necessarily being part of the model of interest (Collins et al., 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 et al., 2014). Similarly, if one is interested in estimating the intraclass correlation (i.e., the variance within and between groups) with incomplete data, it would be crucial to take into account the clustered data structure (Taljaard et al., 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 et al., 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 et al., 2016; Lüdtke et al., 2017).

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. Formally, we refer to this model as the multivariate linear mixed-effects model (MLMM; see Schafer & Yucel, 2002). The model reads

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

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 of the

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

General notation

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

Multivariate empty model

$$\underset{\text{target variables}}{\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}_{ij}} = \underset{\text{fixed effects (intercepts)}}{\begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix}} + \underset{\text{random effects (intercepts)}}{\begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}_j} + \underset{\text{residuals}}{\begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}_{ij}}$$

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

Full mixed-effects model

$$\underset{\text{target variables}}{\begin{bmatrix} y_1 & y_2 \end{bmatrix}_{ij}} = \underset{\text{fixed effects (intercepts, slopes)}}{\begin{bmatrix} 1 & x_1 \end{bmatrix}_{ij} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \end{bmatrix}} + \underset{\text{random effects (intercepts, slopes)}}{\begin{bmatrix} 1 & x_1 \end{bmatrix}_{ij} \begin{bmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \end{bmatrix}_j} + \underset{\text{residuals}}{\begin{bmatrix} e_1 & e_2 \end{bmatrix}_{ij}}$$

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

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.

$\mathbf{y}_{ij}$ and allow for 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 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. In order 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* (L. K. 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, 2013a), `REALCOM` (Carpenter et al., 2011), and the R package `jomo` (Quartagno & Carpenter, 2016a). 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

In order 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`.

[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.

```
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. In order 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 listwise deletion (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 disciplinary problems in math classes (DPM) on math achievement (MA), while controlling for SES at the individual and class level.

$$MA_{ij} = \beta_0 + \beta_1(DPM_{ij} - \overline{DPM}_j) + \beta_2(SES_{ij} - \overline{SES}_j) + \beta_3\overline{DPM}_j + \beta_4\overline{SES}_j + \upsilon_{0j} + \epsilon_{ij} \quad \text{(1, revisited)}$$

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 $\mathbf{\Psi}$ and $\mathbf{\Sigma}$; 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, cognitive ability, ratings of disciplinary problems in reading classes 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 et al., 2014). 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 et al., 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. In order 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 in order 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 $\mathbf{\Psi}$, that is, the intercept variance for math achievement scores ($\hat{R} = 1.011$). 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")
```

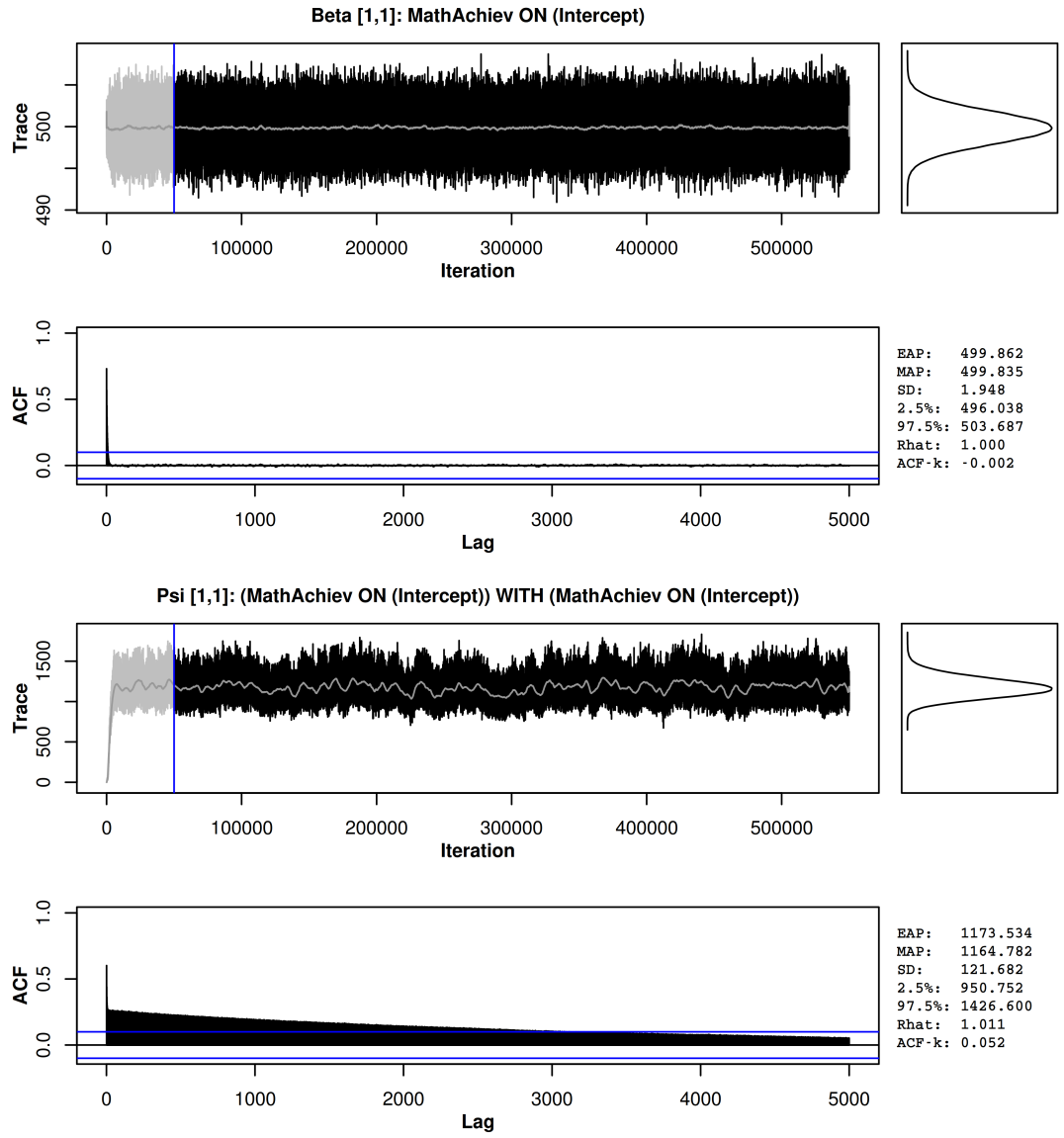

Figure 1: Diagnostic plots for the fixed intercept (top) and the intercept variance (bottom) of math achievement in the imputation model. The trace plot includes all iterations from the burn-in and the imputation phase. The autocorrelation plot and the posterior summaries are calculated only from the imputation phase.

Here, we discuss the diagnostic plots only for the fixed intercept and the intercept variance for math achievement, 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 could be considered independent. We concluded that the parameters had converged and that

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

	Multilevel MI	Single-level MI	Listwise Deletion
ICC_{MA}	0.121	0.111	0.115
ICC_{SES}	0.122	0.072	0.134
ICC_{DPM}	0.179	0.100	0.169

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

the imputed data sets constituted independent draws from the posterior predictive distribution of the missing data.

Intraclass correlations. Usually the first step in analyzing multilevel data is to estimate the intraclass correlation (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 math achievement, SES, and DPM to estimate their ICCs. In order 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 math achievement can be fit as shown below (for DPM and SES, see Supplement B). 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). 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

[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).

(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 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 each estimand that can be attributed to missing data (see Enders, 2010). The output for the model of interest is given below.

```
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.
```

The results for multilevel MI, single-level MI, and LD are presented in Table 5. In general, a higher SES was associated with higher math achievement 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. Listwise deletion 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

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	
$Var(\epsilon_{ij})$	8318.936			8387.913			8299.000	

Note. Estimates were significant at $p < .001$; SE = standard error; MA = math achievement; SES = socioeconomic status; DPM = disciplinary problems in math classes; υ_{0j} = random intercepts; ϵ_{ij} = residuals at Level 1.

coefficient (Kreft et al., 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.
```

[3] The method by Li, Raghunathan, and Rubin (1991) requires that the 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.

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 math achievement 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 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., 2017; 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 cognitive ability (CA) and socioeconomic status (SES) on students' math achievement scores (MA). The effect of SES is assumed to be fixed, whereas the effect of cognitive ability is allowed to vary across groups.

$$\begin{aligned} MA_{ij} = {} & \beta_0 + \beta_1(CA_{ij} - \overline{CA}_j) + \beta_2\overline{CA}_j + \beta_3(SES_{ij} - \overline{SES}_j) + \beta_4\overline{SES}_j \\ & + \upsilon_{0j} + \upsilon_{1j}(CA_{ij} - \overline{CA}_j) + \epsilon_{ij} \end{aligned} \qquad \text{(2, revisited)}$$

As discussed before, the imputation model must consider the model of interest. In this example, the effect of cognitive ability 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 cognitive ability 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 cognitive ability 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 math achievement, SES, and the auxiliary variables on the left-hand side of the equation. In order to include the slope variance, cognitive ability is featured on the right-hand side, where `(1+CognAbility.STU|ID)` allows the intercept and the effect of the group-mean-centered cognitive ability 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 math achievement and SES on cognitive ability 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 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).

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

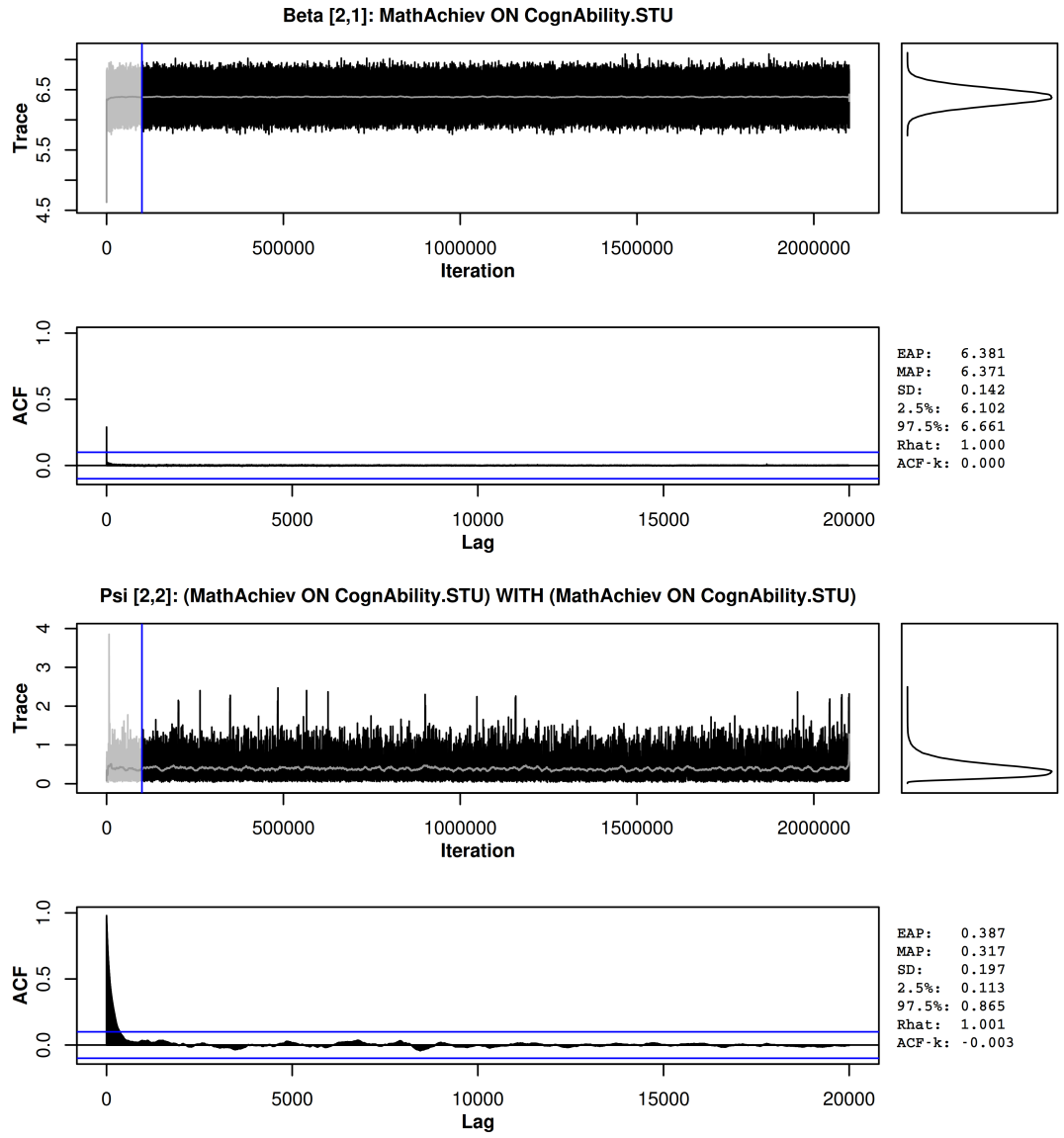

Figure 2: Diagnostic plots for the fixed effect (top) and the slope variance (bottom) for the regression of math achievement on cognitive ability in the full mixed-effects model. The trace plot includes all iterations from the burn-in and the imputation phase. The autocorrelation plot and the posterior summaries are calculated from the imputation phase.

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 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

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	
$Var(\epsilon_{ij})$	6452.506			6547.366			6287.600	

Note. Estimates for the fixed effects were significant at $p < .001$; SE = standard error; CA = cognitive ability; SES = socioeconomic status; υ_{0j} = random intercepts; υ_{1j} = random slopes; ϵ_{ij} = residuals at Level 1.

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

Model of interest. In order 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 cognitive ability (as compared with their class average) tended to score higher on the math achievement 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 cognitive ability. 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, researcher 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, 2012b). 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
```

[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.

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 cognitive ability may differ more or less strongly in their math achievement 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, listwise deletion, 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 continue 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) recommended 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, 2016a). 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 et al., 2014; L. Wu, 2010; see also Bartlett et al., 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., 2016). 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 rules (1987), the package implements the procedures commonly referred to as D_1 (Li, Raghunathan, & Rubin, 1991; Reiter, 2007), D_2 (Li, Meng, et al., 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, AIC or 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, 2016c).

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.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods*, *18*, 155-176. doi: 10.1177/1094428114563618

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.

Allison, P. D. (2012). Handling missing data by maximum likelihood. In *Proceedings of the SAS Global Forum.*

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*(278), 200-203. doi: 10.2307/2280845

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

Andridge, R. R., & Thompson, K. J. (2015). Using the Fraction of Missing Information to Identify Auxiliary Variables for Imputation Procedures via Proxy Pattern-mixture Models. *International Statistical Review*, *83*, 472-492. doi: 10.1111/insr.12091

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques.* Mahwah, NJ: Erlbaum.

Asparouhov, T., & Muthén, B. O. (2006). *Constructing covariates in multilevel regression* (Mplus Web Notes No. 11).

Asparouhov, T., & Muthén, B. O. (2008). *Chi-square statistics with multiple imputation* (Technical Appendix).

Asparouhov, T., & Muthén, B. O. (2010a). *Bayesian analysis using Mplus: Technical implementation* (Technical Appendix).

Asparouhov, T., & Muthén, B. O. (2010b). *Multiple imputation with Mplus* (Technical Appendix).

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.

Audigier, V., & Resche-Rigon, M. (2017). *Micemd: Multiple imputation by chained equations with multilevel data (Version 1.0.0).*

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412. doi: 10.1016/j.jml.2007.12.005

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1312.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7).*

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2016). *Lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-12).*

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606. doi: 10.1037/0033-2909.88.3.588

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, *26*, 1368-1382. doi: 10.1002/sim.2619

Black, A. C., Harel, O., & Matthews, G. (2013). Techniques for analyzing intensive longitudinal data with missing values. In M. R. Mehl & T. S. Connor (Eds.), *Handbook of Research Methods for Studying Daily Life* (p. 339-356). New York, NY: Guilford Press.

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

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. doi: 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. doi: 10.1177/0049124115585360

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (p. 3-90). San Francisco, CA: Jossey-Bass.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Struc-*

tural 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.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets* (Doctoral Dissertation). Erasmus Universiteit Rotterdam.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*, 103-124. doi: 10.1177/1471082X0100100202

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. doi: 10.18637/jss.v045.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

Chen, Q., & Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, *32*, 3646-3659. doi: 10.1002/sim.5783

Cheung, M. W.-L. (2007). Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random.

Organizational Research Methods, *10*, 609-634. doi: 10.1177/1094428106295499

Claeskens, G., & Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, *64*, 1062-1069. doi: 10.1111/j.1541-0420.2008.01003.x

Clayton, D., & Rasbash, J. (1999, January). Estimation in large cross random-effect models by data augmentation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(3), 425-436. doi: 10.1111/1467-985X.00146

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

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.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude × treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, *67*.

Croon, M. A., van Veldhoven, M., Peccei, R., & Wood, S. J. (2014). Researching individual well-being and performance in context. In R. Peccei & M. van Veldhoven (Eds.), *Well-being and performance at work: The role of context* (p. 129-154). New York, NY: Psychology Press.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model.

Psychological Methods, *12*, 45-57. doi: 10.1037/1082-989X.12.1.45

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

de Leeuw, J., & Kreft, I. G. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, *20*, 171-189. doi: 10.3102/10769986020002171

Demirtas, H. (2009). Rounding strategies for multiply imputed binary data. *Biometrical Journal*, *51*, 677-688. doi: 10.1002/bimj.200900018

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, *78*, 69-84. doi: 10.1080/10629360600903866

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)*, 1-38.

Diaz-Ordaz, K., Kenward, M. G., Cohen, A., Coleman, C. L., & Eldridge, S. (2014). Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*, *11*, 590-600. doi: 10.1177/1740774514537136

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*, 49-93. doi: 10.2307/2986113

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

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, *68*, 589-599. doi: 10.2307/2335441

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association*, *68*, 117-130. doi: 10.2307/2284155

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 128-141. doi: 10.1207/S15328007SEM0801_7

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 434-448. doi: 10.1080/10705510802154307

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

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1-16. doi: 10.1037/a0022640

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430-457. doi: 10.1207/S15328007SEM0803_5

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., Keller, B. T., & Levy, R. (in press). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*. doi: 10.1037/met0000148

Enders, C. K., & Mansolf, M. (in press). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*. doi: 10.1037/met0000102

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. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in*

Medicine, *35*, 2955–2974. doi: 10.1002/sim.6944

Fahrmeir, L., & Tutz, G. (2010). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York, NY: Springer.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, *43*, 557-572. doi: 10.2307/2094779

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Galati, J. C., & Seaton, K. A. (2016). MCAR is not necessary for the complete cases to constitute a simple random subsample of the target sample. *Statistical Methods in Medical Research*, *25*, 1527-1534. doi: 10.1177/0962280213490360

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

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

Geronimi, J., & Saporta, G. (2017). Variable selection for multiply-imputed data with penalized generalized estimating equations. *Computational Statistics & Data Analysis*, *110*, 103-114. doi: 10.1016/j.csda.2017.01.001

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

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, *63*, 204-238. doi: 10.1177/0013164402250987

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

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models:

Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*, 523–541.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.

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

Goldstein, H., Carpenter, J. R., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*, 173-197. doi: 10.1177/1471082X0800900301

Gottfredson, N. C., Sterba, S. K., & Jackson, K. M. (2017). Explicating the conditions under which multilevel multiple imputation mitigates bias resulting from random coefficient-dependent missing longitudinal data. *Prevention Science*, *18*, 12–19. doi: 10.1007/s11121-016-0735-3

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. doi: 10.1080/00273171.2012.640589

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80-100. doi: 10.1207/S15328007SEM1001_4

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

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology*, *7*, 121-133. doi: 10.1027/1614-2241/a000030

Grund, S., Lüdtke, O., & Robitzsch, A. (2016a). 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. (2016b). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open*, *6*(4), 1-17. doi: 10.1177/2158244016668220

Grund, S., Lüdtke, O., & Robitzsch, A. (2016c). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*, *12*, 75-88. doi: 10.1027/1614-2241/a000111

Grund, S., Lüdtke, O., & Robitzsch, A. (in press-a). Missing data in multilevel research. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook for multilevel theory, measurement, and analysis.* Washington, DC: American Psychological Association.

Grund, S., Lüdtke, O., & Robitzsch, A. (in press-b). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*. doi: 10.1177/1094428117703686

Grund, S., Robitzsch, A., & Lüdtke, O. (2016). *Mitml: Tools for multiple imputation in multilevel modeling (Version 0.3-2).*

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*, 75-89. doi: 10.1016/j.stamet.2006.03.002

Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, *36*, 1109-1118. doi: 10.1080/02664760802553000

He, Y., & Raghunathan, T. E. (2006). Tukey's *gh* distribution for multiple imputation. *The American Statistician*, *60*, 251-256. doi: 10.1198/000313006X126819

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151-179. doi: 10.3102/1076998606298040

Heymans, M. W., van Buuren, S., Knol, D. L., van Mechelen, W., & de Vet, H. C. W. (2007, July). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, *7*, 33. doi: 10.1186/1471-2288-7-33

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral statistics*, *23*, 117–128. doi: 10.3102/10769986023002117

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NY: Springer.

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

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*, 945-960. doi: 10.2307/2289064

Holman, R., & Glas, C. A. W. (2005, May). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17. doi: 10.1348/000711005X47168

Hox, J. J. (1994). *Applied multilevel analysis*. Amsterdam: TT-publikaties.

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

Hox, J. J., van Buuren, S., & Jolani, S. (2016). Incomplete multilevel data. In J. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 39–61). Charlotte, NC: Information Age Publishing.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, *14*(28), 1-10. doi: 10.1186/1471-2288-14-28

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, *55*, 591– 596. doi: 10.1111/j.0006-341X .1999.00591.x

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*, 551-564. doi: 10.1093/biomet/88.2.551

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, *30*, 55-78. doi: 10.2307/3315865

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. doi: 10.1198/016214504000001844

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

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford: Clarendon Press.

Jelicic, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*, 1195-1199. doi: 10.1037/a0015665

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. doi: 10.3102/1076998615622221

Keller, B. T. (2015). *Three-level multiple imputation: A fully conditional specification approach* (Unpublished doctoral dissertation). Arizona State University.

Keller, B. T., & Enders, C. K. (2016). *Blimp Software Manual (Version Beta 6.6)* (Tech. Rep.). Los Angeles, CA.

Kientoff, C. J. (2011). *Development of weighted model fit indexes for structural equation models using multiple imputation* (Doctoral Dissertation). Iowa State University.

Kim, J. K., Brick, J. M., Fuller, W. A., & Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 509-521. doi: 10.1111/j.1467-9868.2006.00546

.x

Kim, S., Belin, T. R., & Sugar, C. A. (in press). Multiple imputation with non-additively related variables: Joint-modeling and approximations. *Statistical Methods in Medical Research*. doi: 10.1177/0962280216667763

Kim, S., Sugar, C. A., & Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine*, *34*, 1876-1888. doi: 10.1002/sim.6435

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., . . . Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (p. 512-553). San Francisco, CA: Jossey-Bass.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (p. 3-90). San Francisco, CA: Jossey-Bass.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

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

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852. doi: 10.1177/1094428106296642

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.

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.

Little, R. J. A. (1992). Regression with missing *X*'s: A review. *Journal of the American Statistical Association*, *87*, 1227-1237.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121. doi: 10.1080/01621459.1995.10476615

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

Liu, H., Zhang, Z., & Grimm, K. J. (2015). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-14. doi: 10.1080/10705511.2015.1057285

Liu, Y., & Enders, C. K. (in press). Evaluation of multi-parameter test statistics for multiple imputation. *Multivariate Behavioral Research*. doi: 10.1080/00273171.2017.1298432

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. (2017). Multiple imputation of missing data in multilevel

designs: A comparison of different strategies. *Psychological Methods*, *22*, 141-165. doi: 10.1037/met0000096

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. doi: 10.1023/A:1008929526011

Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, *46*, 801-817. doi: 10.1016/j.csda.2003.10.005

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*, 280.

Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Medical Research Methodology*, *9*, 57. doi: 10.1186/1471-2288-9-57

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*, *50*, 504-519. doi: 10.1080/00273171.2015.1068157

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using M*plus*. *Journal of Educational and Behavioral Statistics*, *41*, 27-56. doi: 10.3102/1076998615621299

Mehta, P. D. (2013). *xxM (Version 0.6.0).*

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259-284. doi: 10.1037/1082-989X.10.3.259

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex

samples. *Psychometrika*, *56*, 177–196. doi: 10.1007/BF02294457

Mistler, S. A. (2013a). A SAS macro for applying multiple imputation to multilevel data. In *Proceedings of the SAS Global Forum.*

Mistler, S. A. (2013b). A SAS macro for computing pooled likelihood ratio tests with multiply imputed data. In *Proceedings of the SAS Global Forum.*

Mistler, S. A. (2015). *Multilevel multiple imputation: An examination of competing methods* (Doctoral Dissertation).

Mistler, S. A., & Enders, C. K. (in press). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*. doi: 10.3102/1076998617690869

Molenberghs, G., Kenward, M. G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, *84*, 33-44. doi: 10.1093/biomet/84.1.33

Morris, C. N. (1983, March). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*(381), 47. doi: 10.2307/2287098

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. doi: 10.1111/1467-985X.00177

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 scools in 35 countries*. Chestnut Hill, MA: Boston College.

Muthén, B. O. (1990). Mean and covariance structure analysis of hierarchical data. Princeton, NJ.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398. doi: 10.1177/0049124194022003006

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

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328-362. doi: 10.1177/1094428103254673

Newman, D. A. (2009). Missing data techniques and low response rates. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences.* New York, NY: Routledge.

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

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*, 143-153. doi: 10.1177/0165025415618275

OECD. (2014). *PISA 2012 technical report*. Paris, France: Author.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525-556. doi: 10.3102/00346543074004525

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Plummer, M. (2016). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling (Version 4.2.0).*

Pornprasertmanit, S. (2014). *semTools: Useful tools for structural equation modeling (Version 0.4-6).*

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209-233. doi: 10.1037/a0020141

Quartagno, M. (2016). *Multiple Imputation for Individual Patient Data Meta-Analyses.* (Doctoral Dissertation). London School of Hygiene & Tropical Medicine.

Quartagno, M., & Carpenter, J. R. (2016a). *Jomo: A package for multilevel joint modelling multiple imputation (Version 2.3-1).*

Quartagno, M., & Carpenter, J. R. (2016b). Multiple imputation for IPD meta-analysis: Allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, *35*, 2938–2954. doi: 10.1002/sim.6837

Quartagno, M., & Carpenter, J. R. (2017). *Jomo: A package for multilevel joint modelling multiple imputation (Version 2.4-0).*

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413-425. doi: 10.1016/j.jml.2008.02.002

R Core Team. (2014). *R: A language and environment for statistical computing (Version 3.1.2).*

R Core Team. (2015). *R: A language and environment for statistical computing (Version 3.2.1).*

R Core Team. (2016). *R: A language and environment for statistical computing (Version 3.3.0).*

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190. doi: 10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In *Handbook of structural equation modeling*. New York, NY: Guilford Press.

Raghunathan, T. E. (2015). *Missing data analysis in practice*. Boca Raton, FL: CRC Press.

Raghunathan, T. E., & Dong, Q. (2011). *Analysis of variance from multiply imputed data sets*. Unpublished manuscript, University of Michigan, Ann Arbor, MI.

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.

Rasbash, J., & Browne, W. J. (2008). Non-hierarchical multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (p. 301-334). New York, NY: Springer.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2015). *MLwiN (Version 2.34).*

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

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component 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

Resche-Rigon, M., & White, I. R. (in press). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*. doi: 10.1177/0962280216666564

Rhemtulla, M., Savalei, V., & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, *81*, 60-89. doi: 10.1007/s11336-014-9422-0

Robins, J., & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, *87*, 113-124. doi: 10.1093/biomet/87.1.113

Robitzsch, A., Grund, S., & Henke, T. (2016). *Miceadds: Some additional multiple imputation functions, especially for mice (Version 1.7-8).*

Robitzsch, A., Grund, S., & Henke, T. (2017). *Miceadds: Some additional multiple imputation functions, especiallly for mice (Version 2.3-0).*

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*, 795-819. doi: 10.1007/s11336-016-9544-7

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*, 227-241.

Royston, P. (2005). Multiple imputation of missing values: Update. *The Stata Journal*, *5*, 188-201.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*(4), 1-20.

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1-26. doi: 10.3102/10769986002001001

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

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC.

Statistica Neerlandica, *57*, 3–18. doi: 10.1111/1467-9574.00217

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322-331. doi: 10.1198/016214504000001880

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*, 366. doi: 10.2307/2289225

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (p. 117-153). Boca Raton, FL: CRC Press.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142-151. doi: 10.3102/0013189X10363170

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model Assisted Survey Sampling*. New York, NY: Springer.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 477-497. doi: 10.1080/10705510903008238

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, 150-170. doi: 10.1037/1082-989X.13.2.150

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. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3-15. doi: 10.1177/096228029900800102

Schafer, J. L. (2001). Multiple imputation with PAN. In L. M. Collins & A. G. Sayer

(Eds.), *New methods for the analysis of change* (p. 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., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545-571. doi: 10.1207/s15327906mbr3304_5

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).*

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*, 185-206. doi: 10.1080/00273171.2015.1065398

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC medical research methodology*, *12*(46), 1-13. doi: 10.1186/1471-2288-12-46

Searle, R., Casella, G., & McCulloch, C. (2009). *Variance components*. Hoboken, NJ: Wiley.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605-610. doi: 10.2307/2289471

Shen, Z. (2000). *Nested multiple imputation* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26-53. doi: 10.3102/1076998609345252

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (2012a). Components of variance. In *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (p. 114-117). Thousand Oaks, CA: Sage.

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

Stan Development Team. (2016). *Stan modeling language user's guide and reference manual (Version 2.15.0).*

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*, 481-520. doi: 10.3102/1076998616646200

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439-455. doi: 10.1037/1082-989X.11.4.439

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

Stubbendick, A. L., & Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, *59*, 1140–1150. doi: 10.1111/j.0006-341X.2003.00131.x

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

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-540. doi: 10.2307/2289457

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 631-642. doi: 10.1080/10705511.2014.937378

Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, *49*, 443-459. doi: 10.1080/00273171.2014.931799

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*, 267–288.

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

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

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. doi: 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. doi: 10.18637/jss.v045.i03

van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, *49*, 78-91. doi: 10.1080/00273171.2013.855890

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K., & Magidson, J. (2013). *Latent GOLD (Version 5.0).*

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. doi: 10.1111/j.1467-9531.2008.00202.x

Vink, G., & van Buuren, S. (2013). Multiple imputation of squared terms. *Sociological Methods & Research*, *42*, 598-607. doi: 10.1177/0049124113502943

von Hippel, P. T. (2007). Regression with missing *Y*s: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*, 83-117. doi: 10.1111/j.1467-9531.2007.00180.x

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265-291. doi: 10.1111/j.1467-9531.2009.01215.x

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute

skewed variables? *Sociological Methods & Research*, *42*, 105-138. doi: 10.1177/0049124112464866

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York, NY: Springer.

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*, 1-18. doi: 10.1186/s40536-014-0009-0

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.

Wu, W., Jia, F., & Enders, C. K. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*, 484-503. doi: 10.1080/00273171.2015.1022644

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183-201. doi: 10.1037/a0015858

Yang, J. S., & Seltzer, M. (2016). Handling measurement error in predictors using a multilevel latent variable plausible values approach. In J. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications.* Charlotte, NC: Information Age Publishing.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165-200. doi: 10.1111/0081-1750.00078

Yuan, K.-H., & Hayashi, K. (2005). On Muthén's maximum likelihood for two-level covariance structure models. *Psychometrika*, *70*, 147-167. doi: 10.1007/s11336-003-1070-8

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, *41*, 598-629. doi: 10.1177/0049124112460373

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*, 2389-2403. doi: 10.1098/rsta.2008.0038

Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, *11*, 351-370. doi: 10.1177/1471082X1001100404

Yucel, R. M., Ding, H., Uludag, A. K., & Tomaskovic-Devey, D. (2008). Multiple imputation in multiple classification and multiple-membership structures. In *Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association.*

Yucel, R. M., Schenker, N., & Raghunathan, T. E. (2007). *Sequential hierarchical regression imputation (SHRIMP).* Unpublished Manuscript, University of Massachusetts, Amherst, MA.

Zhang, Q., & Wang, L. (2016). Moderation analysis with missing data in the predictors. *Psychological Methods*. doi: 10.1037/met0000104

Zinn, S. (2013). *An imputation model for multilevel binary data* (NEPS Working Paper No. 31).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Lebenslauf

Persönliche Angaben:

Name:	Simon Grund
Geburtsdatum/-ort:	30. März 1989, Rüdersdorf
Staatsangehörigkeit:	deutsch

Schulbildung:

2008	Abitur (Katholisches Gymnasium Bernhardinum, Fürstenwalde/Spree)

Akademischer Werdegang:

2008-2013	Studium der Psychologie (Humboldt-Universität zu Berlin, Diplom)
2009-2011	studentische Hilfskraft (Max-Planck Institut für Bildungsforschung)
2011-2013	studentische Hilfskraft (Humboldt-Universität zu Berlin)
2013	Diplomarbeit (“Estimation of the random-coefficient model with missing covariate data: Issues with multiple imputation and random slopes”)
2013-2014	wissenschaftlicher Mitarbeiter (Humboldt-Universität zu Berlin, Institut für Psychologie)
2014-	wissenschaftlicher Mitarbeiter (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik)
2016-2017	Promotionsstudent (Christian-Albrechts-Universität zu Kiel)