Article

Journal of Educational and Behavioral Statistics
2018, Vol. 43, No. 3, pp. 316–353
DOI: 10.3102/1076998617738087

Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs

Simon Grund
Oliver Lüdtke
Alexander Robitzsch
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
Centre for International Student Assessment, Germany

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 their 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 Programme for International Student Assessment 2012 study.

Keywords: *multiple imputation; missing data; multilevel; Level 2.*

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, Mistler, & Keller, 2016; Hox, van Buuren, & Jolani, 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, Harel, & McCoach, 2011; Drechsler, 2015; Enders et al., 2016; Lüdtke, Robitzsch, & Grund, 2017; Taljaard, Donner, & Klar, 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 Levels 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, 2010; Carpenter & Kenward, 2013; Goldstein, Carpenter, Kenward, & Levin, 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; Organisation for Economic Co-operation and Development [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, Zyphur, & Zhang, 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

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

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 $\boldsymbol{\Sigma}$. Consequently, assuming latent cluster means, the joint distribution of $\mathbf{x}_{ij}$ and $\mathbf{z}_j$ can be expressed as

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

where $\boldsymbol{\Phi}$ 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 Levels 1 and 2 are equivalent in complete data, $\mathrm{Cov}(\bar{\mathbf{x}}_{\bullet j}, \mathbf{z}_j) = \mathrm{Cov}(\mathbf{u}_j, \mathbf{z}_j)$, the manifest means tend to have a larger variance across clusters, $\mathrm{Var}(\bar{\mathbf{x}}_{\bullet j}) = \mathrm{Var}(\mathbf{u}_j) + \mathrm{Var}(\bar{\mathbf{e}}_{\bullet j})$. 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, 2012). 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 A4 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,

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

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 (FCS) of MI. 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 Levels 1 and 2; nonnormal and categorical variables can also be addressed and will be considered in the Discussion section.

Joint Modeling

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

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

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

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 (FCS) of MI but sometimes also referred to as "chained equations" or sequential MI (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; Royston & White, 2011; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

Let y_{1ijp} denote observation i in cluster j for the pth variable at Level 1 ($p = 1, \ldots, P$), and let y_{2jq} denote the value of cluster j for the qth 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

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

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 pth imputation model. Similarly, for missing values at Level 2, imputations may be generated from

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

where the subscript $(-q)$ denotes the set of variables from which q is excluded and $\boldsymbol{\theta}_q$ denotes the parameters of the qth imputation model. For example, the imputation model at Level 1 may be a multilevel random intercept model (e.g., Schafer & Yucel, 2002; Snijders & Bosker, 2012; 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 qth variable, let y_{2jq}^{obs} and y_{2jq}^{mis} denote the observed and missing values and $\boldsymbol{\theta}_q = \{\beta_{0q}, \boldsymbol{\beta}_{1q}, \phi_q^2\}$ the parameters of the imputation model. For variable q at iteration t,

1. Calculate $\tilde{\mathbf{y}}_{1j}^{(t)}$ from $\mathbf{y}_{1ij}$ either as manifest or as latent cluster means (see below).
2. Draw $\boldsymbol{\theta}_q^{(t+1)} \sim P\left(\boldsymbol{\theta} | \tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j}^{(t)}\right)$ 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 $y_{2jq}^{\mathrm{imp},(t+1)} \sim P\left(y_{2jq}^{\mathrm{mis}} | \boldsymbol{\theta}_q^{(t+1)}, \tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j(-q)}^{(t)}\right)$ from a univariate normal distribution $N\left(\beta_{0q}^{(t+1)} + \tilde{\mathbf{x}}_{j(-q)}^{(t)} \boldsymbol{\beta}_{1q}^{(t+1)}, \phi_q^{2,(t+1)}\right)$, conditional on the predictor variables $\tilde{\mathbf{x}}_{j(-q)}^{(t)} = \left(\tilde{\mathbf{y}}_{1j}^{(t)}, \mathbf{y}_{2j(-q)}^{(t)}\right)$.

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 pth variable at Level 1,

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

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, Schenker, & Raghunathan, 2007). For the pth 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 $y_{1ijp} = \hat{\beta}_{0p}^{(t)} + \widehat{\boldsymbol{\beta}}_{1p}^{(t)} \tilde{\mathbf{x}}_{j(-p)}^{(t)} + \upsilon_j + \epsilon_{ij}$, obtaining estimates for the conditional mean $\hat{\mu}_{p|(-p)j}^{(t)} = \hat{\beta}_{0p}^{(t)} + \widehat{\boldsymbol{\beta}}_{1p}^{(t)} \tilde{\mathbf{x}}_{j(-p)}^{(t)}$ 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 $\tilde{\mathbf{x}}_{j(-p)}^{(t)} = \left(\tilde{\mathbf{y}}_{1j(-p)}^{(t)}, \mathbf{y}_{2j}^{(t)}\right)$.
2. Draw $u_{1jp}^{(t)} \sim P\left(u_{1jp} | \tilde{\mathbf{y}}_{1j(-p)}^{(t)}, \mathbf{y}_{2j}^{(t)}, \widehat{\boldsymbol{\theta}}_p^{*(t)}\right)$ from a univariate normal distribution $N\left(\tilde{u}_{1jp}^{(t)}, U_{1jp}^{(t)}\right)$, where the mean and variance are calculated as follows.
   a. $\tilde{u}_{1jp}^{(t)} = \mu_{p|(-p)j}^{(t)} + \lambda_{p|(-p)j}^{(t)} \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y_{1ijp} - \hat{\mu}_{p|(-p)j}^{(t)}\right)$, where $\lambda_{p|(-p)j}^{(t)} = \frac{\hat{\psi}_{p|(-p)j}^{2,(t)}}{\hat{\psi}_{p|(-p)j}^{2,(t)} + \hat{\sigma}_p^{2,(t)}/n_j}$ is the reliability of the conditional cluster means of y_{1ijp}, given $\tilde{\mathbf{y}}_{1j(-p)}$ and $\mathbf{y}_{2j}$.
   b. $U_{1jp}^{(t)} = \lambda_{p|(-p)j}^{(t)} \cdot \frac{\hat{\sigma}_p^{2,(t)}}{n_j}$, 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, Honaker, & King, 2017; 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:

$$\tilde{u}_{1jp} = \left(1 - \lambda_{p|(-p)j}\right) \cdot \hat{\mu}_{p|(-p)j} + \lambda_{p|(-p)j} \cdot \frac{1}{n_j}\sum_{i=1}^{n_j} y_{1ijp}. \tag{9}$$

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 $\hat{\mu}_{p|(-p)j}$ 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 toward 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

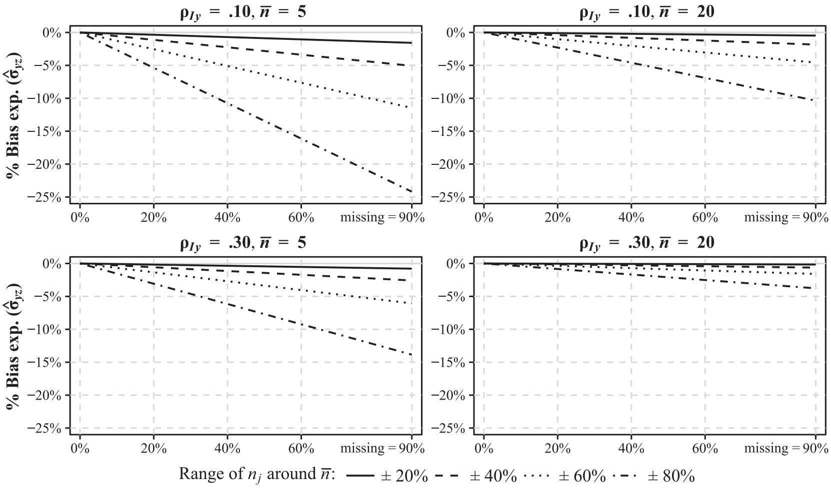

FIGURE 1. *Expected bias for the covariance of* y *with* z *under FCS-MAN, for varying amounts of missing data, different intraclass correlations of* y *(ρ_{Iy}), different average cluster sizes (*$\bar{\mathrm{n}}$*), and different ranges of cluster sizes (*$\mathrm{n_j}$*, assuming a uniform distribution). FCS-MAN = two-level FCS with manifest cluster means; FCS = fully conditional specification.*

sample. In the Appendix, the bias is derived under the assumption that the number of clusters goes to infinity and that the data are missing 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

$$\%\mathrm{Bias}(\hat{\sigma}_{yz}) = \alpha \left[\sum_{k \in \mathcal{S}} \left(\frac{k}{\bar{n}} - 1 \right) \pi_k \left(\tau_y^2 + \frac{\sigma_y^2}{k} \right) \right] \left[\sum_{k \in \mathcal{S}} \pi_k \left(\tau_y^2 + \frac{\sigma_y^2}{k} \right) \right]^{-1}, \qquad (10)$$

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 are the variance components of y at Levels 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 $\left(\frac{k}{\bar{n}} - 1\right)$ as opposed to larger clusters with less variability, the bias in $\hat{\sigma}_{yz}$ tends to be negative (i.e., toward 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 of the cluster sizes, different amounts of missing data, and different values for the intraclass correlation (ICC) of y (ρ_{Iy}). 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 A6. 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 Levels 1 and 2, we defined the ICC of y (ρ_{Iy}) and the correlation between the two

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, 1,000	50, 200, 1,000
Range in cluster size	—	uniform ($\pm$40%, 80%), bimodal ($\pm$40%, 80%)
Intraclass correlation 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.

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., $\alpha_0 = -0.842$ 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 $1 - \lambda^2 \mathrm{Var}(\bar{y}_{\bullet j})$. A value z_j was deleted if $r_j > 0$.

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, 2016). 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 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 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* Version 7.3 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 Online Supplemental Material.

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 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}),

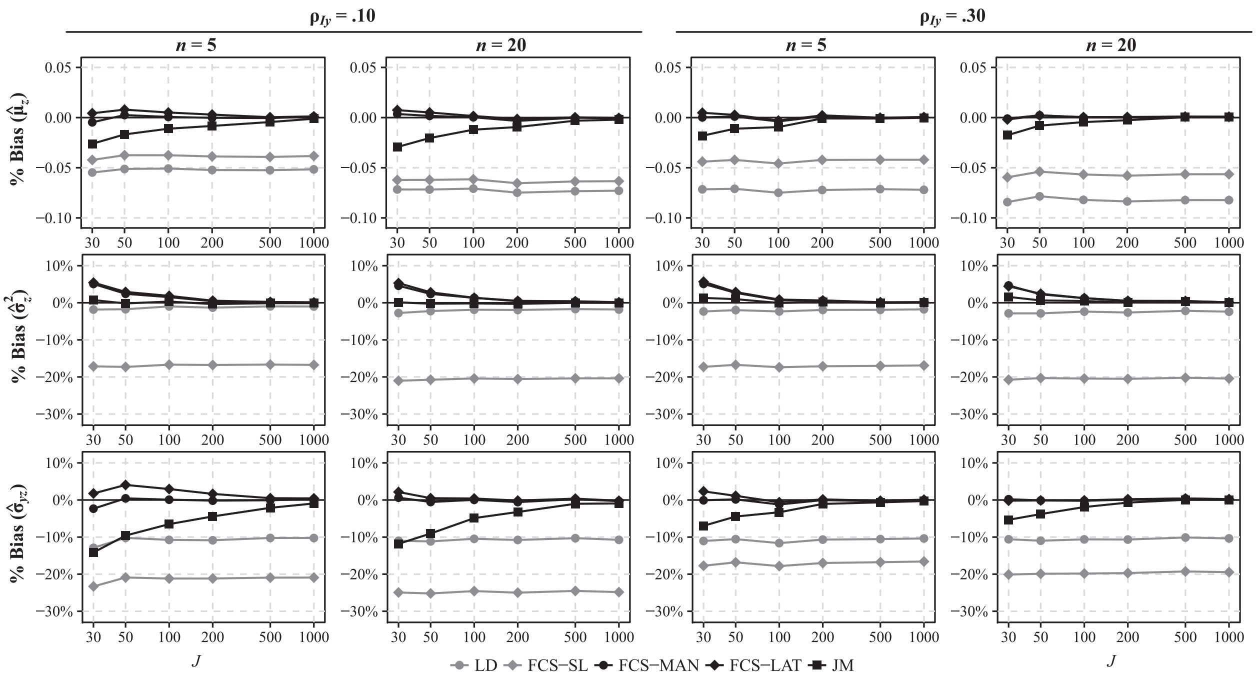

FIGURE 2. *Estimated bias for the mean and the variance of* z *(*$\hat{\mu}_z$ *and* $\hat{\sigma}_z^2$*) and the covariance of* y *with* z *(*$\hat{\sigma}_{yz}$*) for varying sample sizes at Level 1 (*n*) and Level 2 (*J*) and intraclass correlation of* y *(*ρ_{Iy}*) with 20% missing data (missing at random,* $\lambda = 0.5$*). 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; FCS = fully conditional specification.*

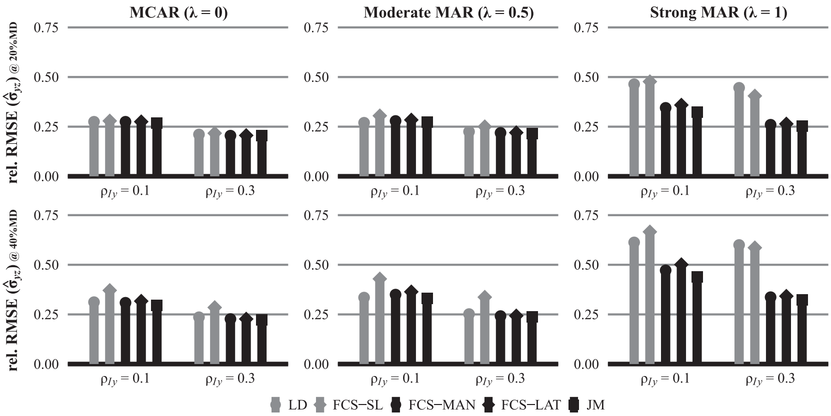

FIGURE 3. *Relative root mean squared error (RMSE) for the covariance of* y *with* z *($\hat{\sigma}_{yz}$) with* $n = 5$ *and* $J = 200$*, for varying intraclass correlations of* y *(ρ_{Iy}), different missing data mechanisms (λ), and different portions of missing data (MD). LD = listwise deletion; 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; FCS = fully conditional specification.*

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 Supplement D in the Online Supplemental Material). 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 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 $\rho_{Iy} = .10$) or in conditions with larger portions of missing data (see Supplement

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 a Small Intraclass Correlation of* y *(ρ_{Iy} = .10) and 20% Missing Data (Missing at Random, λ = 0.5)*

	Bias (%)				Relative 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
$J = 1{,}000$	−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
$J = 1{,}000$	−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 $y \sim z(\hat{\beta}_{yz})$												
$n = 5$												
$J = 30$	−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

(continued)

TABLE 2. *(continued)*

	Bias (%)				Relative RMSE				Coverage (%)			
	FCS-SL	FCS-MAN	FCS-LAT	JM	FCS-SL	FCS-MAN	FCS-LAT	JM	FCS-SL	FCS-MAN	FCS-LAT	JM
$J = 200$	−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
$J = 1{,}000$	−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
$J = 1{,}000$	−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; FCS-SL = single-level FCS; JM = joint modeling; RMSE = root mean squared error; FCS = fully conditional specification.

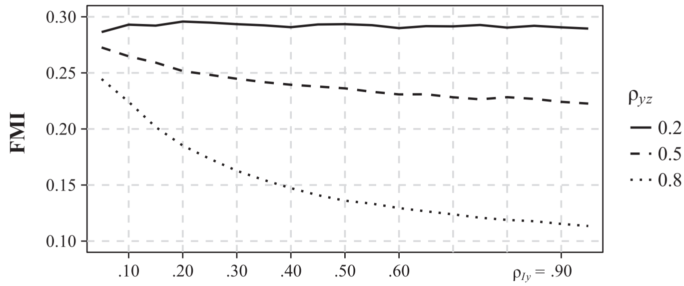

FIGURE 4. *Fraction of missing information (FMI) for estimates of the covariance of* y *with* z *(*$\hat{\sigma}_{yz}$*) with* $n = 5$ *and* $J = 200$*, and 20% missing data (missing at random,* $\lambda = 0.5$*), for varying intraclass correlations of* y *(*ρ_{Iy}*), and different correlations between* y *and* z *(*ρ_{yz}*).*

D in the Online Supplemental Material). 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 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_{Iy}) 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 $\pm 40\%$ or $\pm 80\%$ around the average cluster size $\bar{n}$ (e.g., for $\bar{n} = 5$ and range $\pm 80\%$, $n_j = 1, 2, \ldots, 9$) or from a bimodal distribution that included only the extreme points of this range (e.g., for $\bar{n} = 5$ and range $\pm 80\%$, $n_j = 1$ 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 in the Online Supplemental Material). These results are summarized in Table 3 for conditions with a low ICC of y ($\rho_{Iy} = .10$) 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 ($\bar{n} = 5$) and strongly unbalanced data ($\pm 80\%$). Biases larger than -10% were obtained under FCS-MAN only in conditions with 40% missing data and strongly unbalanced data ($\pm 80\%$, 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 the Online

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 a Small Intraclass Correlation of* y *($\rho_{Iy} = .10$) and 20% Missing Data (Missing at Random, $\lambda = 0.5$)*

	Bias (%)				Relative 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
$J = 1{,}000$	−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
$J = 1{,}000$	−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
$J = 1{,}000$	−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
$J = 1{,}000$	−0.9	−0.3	−0.1	−0.8	0.099	0.098	0.098	0.097	94.4	94.2	93.8	94.1

(continued)

TABLE 3. *(continued)*

	Bias (%)				Relative 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 (bimodal, $\pm 40\%$)												
$\bar{n} = 5$												
$J = 50$	−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
$J = 1{,}000$	−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
$J = 1{,}000$	−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, $\pm 80\%$)												
$\bar{n} = 5$												
$J = 50$	−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
$J = 1{,}000$	−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
$J = 1{,}000$	−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; RMSE = root mean squared error; FCS = fully conditional specification.

Supplemental Material). 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 upward under FCS-LAT ($J = 50$) and downward 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 in the Online Supplemental Material). 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 ($\bar{n} = 5, J = 50$). 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 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 Levels 1 and 2 using MI, we applied the procedures used in Study 1 to the German subsample of the PISA 2012 study (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 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 Online Supplemental Material.

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

TABLE 4.

Parameter Estimates Obtained From the Programme for International Student Assessment 2012 Data in the Empirical Example Using Different Multiple Imputation 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	1,140.753	(202.834)	1,257.663	(158.599)	1,280.906	(173.783)	1,265.579	(163.697)
Residual variance	4,044.420	(86.424)	4,058.046	(87.204)	4,055.189	(87.459)	4,060.389	(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; *SE* = standard error; ESCS = economic, social, and cultural status; FCS = fully conditional specification.

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 ($\rho_{Iy} = .10$), 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 Levels 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 Online Supplemental Material). 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, 2016; 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 Online Supplemental Material, 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, Olchowski, & Gilreath, 2007; see also the Empirical Example subsection). 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 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, Lüdtke, & Robitzsch, in press; 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 Levels 1 and 2 (Asparouhov & Muthén, 2010; Quartagno & Carpenter, 2016). 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 $\mathbf{T} \equiv \mathrm{Var}(\mathbf{u}_j)$, $\boldsymbol{\Sigma} \equiv \mathrm{Var}(\mathbf{e}_{ij})$, and $\phi^2 \equiv \mathrm{Var}(w_j)$ 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

$$\mathrm{Var}(\mathbf{y}_{ij}) = \begin{pmatrix} \mathbf{T} + \boldsymbol{\Sigma} & \boldsymbol{\sigma}^{\mathrm{T}} \\ \boldsymbol{\sigma} & \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{T}\boldsymbol{\gamma} + \phi^2 \end{pmatrix}. \tag{A2}$$

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 $\pi_k \equiv P(n_j = k)$ 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, $z = (z^{\mathrm{mis}}, z^{\mathrm{obs}})$, 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-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 $\mathbf{y}_{ij}^{\mathrm{com}} = (\mathbf{y}_{ij}^{\mathrm{imp}}, \mathbf{y}_{ij}^{\mathrm{obs}})$ 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:

$$\begin{aligned} u_{jp}^{\mathrm{imp}} &\sim P(u_{jp} | \mathbf{x}_{ij}, \mathbf{u}_{j(-p)}, z_j) \\ z_j^{\mathrm{imp}} &\sim P(z_j | \mathbf{x}_{ij}, \mathbf{u}_j), \end{aligned} \tag{A3}$$

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

$$z_j^{\mathrm{imp}} = \mathbf{u}_j^{\mathrm{imp}} \boldsymbol{\gamma} + w_j^{\mathrm{imp}}, \tag{A4}$$

where estimates of $\boldsymbol{\gamma}$ and ϕ^2 are obtained from the observed data, and posterior draws for $\mathbf{u}_j^{\mathrm{imp}}$ 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-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} is distributed normally with mean zero and variance $\text{Var}(\epsilon_j^{\text{imp}})$. 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

$$\widehat{\boldsymbol{\beta}} = \left[\frac{1}{J_1}\sum_{j=J_0+1}^{J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}\bar{\mathbf{x}}_{\bullet j}\right]^{-1}\left(\frac{1}{J_1}\sum_{j=J_0+1}^{J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}z_j^{\text{obs}}\right). \tag{A6}$$

Note that $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}\bar{\mathbf{x}}_{\bullet j}) = \mathbf{T} + \frac{1}{n_j}\Sigma$ and $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}z_j^{\text{obs}}) = \mathbf{T}\boldsymbol{\gamma}$. Then, as the number of clusters goes to infinity ($J \rightarrow \infty$), the expected value of $\widehat{\boldsymbol{\beta}}$ can then be expressed as

$$E(\widehat{\boldsymbol{\beta}}) \overset{J\rightarrow\infty}{=} \left[\sum_{k\in\mathcal{S}}\pi_k\left(\mathbf{T} + \frac{1}{k}\boldsymbol{\Sigma}\right)\right]^{-1}\mathbf{T}\boldsymbol{\gamma}. \tag{A7}$$

In the special case with balanced data with a constant cluster size $n_j = k_0$, it is further worth noting that $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}\bar{\mathbf{x}}_{\bullet j}) = \left(\mathbf{T} + \frac{1}{k_0}\boldsymbol{\Sigma}\right)$, in which case Equation A7 reduces to

$$E(\widehat{\boldsymbol{\beta}}) \overset{J\rightarrow\infty}{=} \left(\mathbf{T} + \frac{1}{k_0}\boldsymbol{\Sigma}\right)^{-1}\mathbf{T}\boldsymbol{\gamma}. \tag{A8}$$

Carpenter and Kenward (2013) showed 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 $\hat{z}_j = \bar{\mathbf{x}}_{\bullet j}\widehat{\boldsymbol{\beta}}$. Under the given assumptions, it holds that $\text{Var}(\hat{z}_j^{\text{mis}}) = \text{Var}(\hat{z}_j^{\text{obs}})$ and $\text{Var}(\epsilon_j^{\text{imp}}) = \text{Var}(\epsilon_j^{\text{obs}})$. As a result, $\text{Var}(z_j^{\text{imp}}) = \text{Var}(\hat{z}_j^{\text{mis}}) + \text{Var}(\epsilon_j^{\text{imp}}) = \text{Var}(z_j)$, 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 ($\widehat{\boldsymbol{\sigma}}$) allows estimating $\boldsymbol{\sigma}$ in closed form and can be expressed as

$$\widehat{\boldsymbol{\sigma}} = \frac{1}{J}\sum_{j=1}^{J}\frac{n_j}{c_J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j\,, \tag{A9}$$

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, $\widehat{\boldsymbol{\sigma}}$ 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

$$\widehat{\boldsymbol{\sigma}} = \frac{1}{J}\sum_{j=1}^{J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j\,. \tag{A10}$$

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-MAN.

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 $z_j^{\mathrm{com}} = (z_j^{\mathrm{obs}}, z_j^{\mathrm{imp}})$ as follows:

$$\widehat{\boldsymbol{\sigma}} = \frac{1}{J}\left(\sum_{j=1}^{J_0}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}} + \sum_{j=J_0+1}^{J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right). \tag{A11}$$

The expected value of $\widehat{\boldsymbol{\sigma}}$ can then be expressed as

$$E(\widehat{\boldsymbol{\sigma}}) = E\left[\frac{1}{J}\left(\sum_{j=1}^{J_0}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}} + \sum_{j=J_0+1}^{J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right)\right] = E\left(\frac{J_0}{J}\right)E\left(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}}\right) + E\left(\frac{J_1}{J}\right)E\left(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right). \tag{A12}$$

In the limit of $J \to \infty$, it holds that $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}}) \overset{J\to\infty}{=} E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}}\bar{\mathbf{x}}_{\bullet j})E(\widehat{\boldsymbol{\beta}})$. Then, by further noting that $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}) = \mathbf{T}\boldsymbol{\gamma}$ as before and by plugging in Equation A8, it can be shown that Equation A12 converges to

$$E(\widehat{\boldsymbol{\sigma}}) \overset{J\to\infty}{=} \alpha\left(\mathbf{T} + \frac{1}{k_0}\boldsymbol{\Sigma}\right)\left[\mathbf{T} + \frac{1}{k_0}\boldsymbol{\Sigma}\right]^{-1}\mathbf{T}\boldsymbol{\gamma} + (1-\alpha)\mathbf{T}\boldsymbol{\gamma} = \mathbf{T}\boldsymbol{\gamma} = \boldsymbol{\sigma}\,, \tag{A13}$$

which shows that $\widehat{\boldsymbol{\sigma}}$ 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 $\widehat{\boldsymbol{\sigma}}$ 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 $\widehat{\boldsymbol{\sigma}}_k$ denote the value of $\widehat{\boldsymbol{\sigma}}$ 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 Equation A9. Consequently, $\widehat{\boldsymbol{\sigma}}_k$ can be expressed as

$$\widehat{\boldsymbol{\sigma}}_k = \widehat{\boldsymbol{\sigma}}_{|n_j=k} = \frac{1}{J_{(k)}}\left(\sum_{j\in\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{imp}} + \sum_{j\in\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{obs}}\right), \tag{A14}$$

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 $J_{(k)} = J_{0(k)} + J_{1(k)}$ denotes the total number of clusters of size k. By noting that $E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{imp}}) \overset{J\to\infty}{=} E(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} \bar{\mathbf{x}}_{\bullet j})E(\widehat{\boldsymbol{\beta}})$ as before and by plugging in Equation A7, the expected value of $\widehat{\boldsymbol{\sigma}}_k$ can be written as

$$\begin{aligned} E(\widehat{\boldsymbol{\sigma}}_k) &= E\left[\frac{1}{J_{(k)}}\left(\sum_{j\in\mathcal{J}_{0(k)}} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{imp}} + \sum_{j\in\mathcal{J}_{1(k)}} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{obs}}\right)\right] \\ &= E\left(\frac{J_{0(k)}}{J_{(k)}}\right)E\left(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{imp}}\right) + E\left(\frac{J_{1(k)}}{J_{(k)}}\right)E\left(\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{obs}}\right) \\ &\overset{J\to\infty}{=} \alpha\left(\mathbf{T} + \frac{1}{k}\Sigma\right)\left[\sum_{k'\in\mathcal{S}} \pi_{k'}\left(\mathbf{T} + \frac{1}{k'}\Sigma\right)\right]^{-1} \mathbf{T}\gamma + (1-\alpha)\mathbf{T}\boldsymbol{\gamma}, \end{aligned} \tag{A15}$$

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, $\widehat{\boldsymbol{\sigma}}$ is again based on both the observed and imputed data, z_j^{obs} and z_j^{imp}, and can be written as

$$\widehat{\boldsymbol{\sigma}} = \frac{1}{J}\left(\sum_{j=1}^{J_0} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{imp}} + \sum_{j=J_0+1}^{J} \frac{n_j}{c_J} \bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\text{obs}}\right). \tag{A16}$$

The expected value of $\widehat{\boldsymbol{\sigma}}$ can then be expressed as

$$
\begin{aligned}
E(\widehat{\boldsymbol{\sigma}}) &= E\left[\frac{1}{J}\left(\sum_{j=1}^{J_0}\frac{n_j}{c_J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}} + \sum_{j=J_0+1}^{J}\frac{n_j}{c_J}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right)\right] \\
&= E\left(\frac{J_0}{J}\right)E\left(\sum_{k\in\mathcal{S}}\frac{J_{0(k)}}{J_0}\frac{k}{c_J}\frac{1}{J_{0(k)}}\sum_{\mathcal{J}_{0(k)}}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}}\right) + E\left(\frac{J_1}{J}\right)E\left(\sum_{k\in\mathcal{S}}\frac{J_{1(k)}}{J_1}\frac{k}{c_J}\frac{1}{J_{1(k)}}\sum_{\mathcal{J}_{1(k)}}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right) \\
&= E\left(\frac{J_0}{J}\right)\sum_{k\in\mathcal{S}} k\, E\left(\frac{J_{0(k)}}{J_0}\right)E\left(\frac{1}{c_J}\right)E\left(\frac{1}{J_{0(k)}}\sum_{j\in\mathcal{J}_{0(k)}}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{imp}}\right) \\
&\quad + E\left(\frac{J_1}{J}\right)\sum_{k\in\mathcal{S}} k\, E\left(\frac{J_{1(k)}}{J_1}\right)E\left(\frac{1}{c_J}\right)E\left(\frac{1}{J_{1(k)}}\sum_{j\in\mathcal{J}_{1(k)}}\bar{\mathbf{x}}_{\bullet j}^{\mathrm{T}} z_j^{\mathrm{obs}}\right),
\end{aligned}
\tag{A17}
$$

which illustrates the contribution of the conditional expectations given in Equation A15. In the limit as $J \to \infty$, this expression converges to

$$
\begin{aligned}
E(\widehat{\boldsymbol{\sigma}}) &\overset{J\to\infty}{=} \alpha\sum_{k\in\mathcal{S}}\frac{k}{\bar{n}_\infty}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\left[\sum_{k'\in\mathcal{S}}\pi_{k'}\left(\mathbf{T}+\frac{1}{k'}\boldsymbol{\Sigma}\right)\right]^{-1}\boldsymbol{\sigma} + (1-\alpha)\sum_{k\in\mathcal{S}}\frac{k}{\bar{n}_\infty}\pi_k\boldsymbol{\sigma} \\
&= \alpha\left(\sum_{k\in\mathcal{S}}\frac{k}{\bar{n}_\infty}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\right)\left[\sum_{k\in\mathcal{S}}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\right]^{-1}\boldsymbol{\sigma} + (1-\alpha)\boldsymbol{\sigma}\,.
\end{aligned}
\tag{A18}
$$

This expression is generally not equal to $\boldsymbol{\sigma}$. Consequently, the asymptotic bias of $\widehat{\boldsymbol{\sigma}}$ as an estimator of $\boldsymbol{\sigma}$ can be expressed as

$$
\begin{aligned}
\mathrm{Bias}(\widehat{\boldsymbol{\sigma}}) &= \alpha\left\{\sum_{k\in\mathcal{S}}\frac{k}{\bar{n}_\infty}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\left[\sum_{k\in\mathcal{S}}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\right]^{-1} - 1\right\}\boldsymbol{\sigma} \\
&= \alpha\left\{\sum_{k\in\mathcal{S}}\left(\frac{k}{\bar{n}_\infty}-1\right)\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\right\}\left[\sum_{k\in\mathcal{S}}\pi_k\left(\mathbf{T}+\frac{1}{k}\boldsymbol{\Sigma}\right)\right]^{-1}\boldsymbol{\sigma}\,,
\end{aligned}
\tag{A19}
$$

which is not generally zero in unbalanced data. Because the expected value of $\widehat{\boldsymbol{\sigma}}$ 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]

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

S. Grund https://orcid.org/0000-0002-1290-8986

Notes

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.
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 fully conditional specification (FCS) approach and will not be considered further (for a discussion, see Enders et al., 2016).
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 $\widehat{\boldsymbol{\theta}}^{*(t)}$ are replaced with Bayesian posterior draws $\boldsymbol{\theta}^{*(t)}$ (see the Discussion section).
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 intraclass correlation 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.
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).
6. It is interesting that unbiased estimates of $\boldsymbol{\sigma}$ might be obtained under FCS with manifest cluster means 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.

References

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

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

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

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

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

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. Retrieved from http://www.stat.sinica.edu.tw/statistica/

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. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, *46*, 303–341. doi:10.1177/0049124115585360

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

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

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

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis*. Stanford, CA: Stanford Evaluation Consortium. Retrieved from http://eric.ed.gov/

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

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., 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

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

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

Enders, C. K., 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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534. Retrieved from http://projecteuclid.org

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.

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

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

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

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

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

He, Y., & Raghunathan, T. E. (2006). Tukey's *gh* distribution for multiple imputation. *The American Statistician*, *60*, 251–256. doi:10.1198/000313006X126819

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*, 1–10. doi:10.1186/1471-2288-14-28

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

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

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

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using *Mplus*. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. doi:10.3102/1076998615621299

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. doi:10.1007/BF02294457

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

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47. doi:10.2307/2287098

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Paper presented at the meeting of the Psychometric Society, Princeton, NJ. Retrieved from https://www.statmodel.com

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

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. Paris, France: Author.

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., & Carpenter, J. R. (2016). *jomo: A package for multilevel joint modelling multiple imputation* [Version 2.3-1]. Retrieved from http://CRAN.R-project.org/package=jomo

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. Retrieved from http://www.statcan.gc.ca/

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

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

Robitzsch, A., Grund, S., & Henke, T. (2017). *miceadds: Some additional multiple imputation functions, especially for mice* [Version 2.3-0]. Retrieved from http://CRAN.R-project.org/package=miceadds

Royston, P. (2005). Multiple imputation of missing values: Update. *The Stata Journal*, *5*, 188–201. Retrieved from http://www.stata-journal.com

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*, 1–20.

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

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* (pp. 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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.

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

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

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. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

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

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

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

van Buuren, S., 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*, 1–67. doi:10.18637/jss.v045.i03

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

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

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., & 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

Yucel, R. M., Schenker, N., & Raghunathan, T. E. (2007). *Sequential hierarchical regression imputation (SHRIMP)*. Unpublished Manuscript, University of Massachusetts, Amherst, MA.

Authors

SIMON GRUND is a PhD student at the Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; email: grund@ipn.uni-kiel.de. His research interests include missing data, causality, and multilevel modeling.

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

ALEXANDER ROBITZSCH is a principal researcher at the Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; email: robitzsch@ipn.uni-kiel.de. His research interests include item response modeling, missing data, and multilevel modeling.

Manuscript received November 30, 2016
First revision received April 20, 2017
Second revision received July 14, 2017
Accepted September 25, 2017