



A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices

Craig K. Enders^{a,b}, Timothy Hayes^c, and Han Du^b

^aUniversity of California, Los Angeles, CA, USA; ^bUCLA Department of Psychology, University of California, Los Angeles, CA, USA;

^cFlorida International University, Miami, FL, USA

ABSTRACT

Literature addressing missing data handling for random coefficient models is particularly scant, and the few studies to date have focused on the fully conditional specification framework and “reverse random coefficient” imputation. Although it has not received much attention in the literature, a joint modeling strategy that uses random within-cluster covariance matrices to preserve cluster-specific associations is a promising alternative for random coefficient analyses. This study is apparently the first to directly compare these procedures. Analytic results suggest that both imputation procedures can introduce bias-inducing incompatibilities with a random coefficient analysis model. Problems with fully conditional specification result from an incorrect distributional assumption, whereas joint imputation uses an underparameterized model that assumes uncorrelated intercepts and slopes. Monte Carlo simulations suggest that biases from these issues are tolerable if the missing data rate is 10% or lower and the sample is composed of at least 30 clusters with 15 observations per group. Furthermore, fully conditional specification tends to be superior with intraclass correlations that are typical of cross-sectional data (e.g., $ICC = .10$), whereas the joint model is preferable with high values typical of longitudinal designs (e.g., $ICC = .50$).

KEYWORDS

Missing data; multiple imputation; multilevel models; joint model imputation; fully conditional specification

A substantial body of methodological research supports missing data handling techniques that assume a missing at random (MAR) mechanism by which the probability of nonresponse on a given variable is unrelated to that variable’s scores after conditioning on observed data (Little & Rubin, 2002; Rubin, 1976). Maximum likelihood estimation and multiple imputation are the principal MAR-based methods that enjoy widespread use in the behavioral sciences. Relative to single-level analysis problems, the application of these approaches to multilevel missing data has received somewhat less attention in the literature. Existing research has largely focused on random intercept models (DiazOrdaz, Kenward, Gomes, & Grieve, 2016; Drechsler, 2015; Goldstein, Carpenter, Kenward, & Levin, 2009; Larsen, 2011; Lüdtke, Robitzsch, & Grund, 2017; Mistler & Enders, 2017; Reiter, Raghunathan, & Kinney, 2006; Shin & Raudenbush, 2007, 2010; Taljaard, Donner, & Klar, 2008; van Buuren, 2011; Yucel & Demirtas, 2010), and literature

addressing random coefficient models is especially nascent (Enders, Keller, & Levy, 2018; Grund, Lüdtke, & Robitzsch, 2016).

Joint model imputation and fully conditional specification (also known in the literature as chained equations and sequential regression imputation) are the major multiple imputation frameworks for multilevel data (Carpenter, Goldstein, & Kenward, 2011; Enders, Mistler, & Keller, 2016; Enders et al., 2017; Goldstein, Bonnet, & Rocher, 2007; Goldstein et al., 2009; Schafer, 2001; Schafer & Yucel, 2002; van Buuren, 2011, 2012), and both are widely available in software packages. Consistent with their single-level counterparts, these procedures create several copies of the data, each containing different estimates of the missing values. Both employ an iterative algorithm that estimates multilevel model parameters and then samples’ missing values from a distribution that conditions on these quantities. Importantly, joint model imputation uses a multivariate multilevel regression

model that features incomplete variables regressed on complete variables, whereas fully conditional specification imputes variables one at a time from a sequence of univariate models with an incomplete variable regressed on complete and previously imputed variables. As explained later, univariate and multivariate imputation schemes offer very different approaches to random coefficient models.

The few studies to examine missing data handling for random coefficient models have focused on the fully conditional specification imputation framework (Enders et al., 2016, 2017; Grund et al., 2016) popularized by van Buuren and colleagues (van Buuren, 2011, 2012; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren et al., 2014). In the context of a random slope analysis, fully conditional specification employs a “reverse random coefficient” approach (Grund et al., 2016) that treats y as a random predictor of an incomplete covariate, and *vice versa*. Simulation results suggest that fully conditional specification can produce negatively biased slope variance estimates (Enders et al., 2017; Grund et al., 2016), and analytic work shows that an incorrect distributional assumption is responsible for this problem (Enders, Du, & Keller, 2017, October). As an alternative, Yucel (2011) proposed a joint (multivariate) imputation strategy that uses random within-cluster covariance matrices to model cluster-specific associations. This promising approach was recently implemented in the R package *jomo* (Quartagno & Carpenter, 2016a), and the only study to examine its performance did so in the context of individual patient data meta-analysis (Quartagno & Carpenter, 2016b). As we show later, using within-cluster covariance matrices to model random coefficients is potentially problematic because the method cannot preserve covariance between the random intercepts and slopes, but the practical magnitude of this shortcoming is unclear.

Random coefficient models pose particularly challenging missing data problems because existing imputation frameworks can introduce bias. Promising new Bayesian imputation procedures are currently under development but are not yet widely available in software (Enders et al., 2016, 2017, October). Although fully conditional specification and joint modeling with random covariance matrices are both imperfect, they will not necessarily produce comparable results because their underlying models are quite different. As such, understanding the relative strengths and weaknesses of these strategies is important for selecting an appropriate imputation procedure. As the methodology literature currently offers no guidance

on this issue, the primary goal of our manuscript is to compare fully conditional specification and joint model imputation with random level-1 covariance matrices. We begin by describing each approach and discussing its compatibility with a random coefficient analysis, after which we use Monte Carlo computer simulations to evaluate parameter recovery in a wide range of conditions typical of applied research settings. This article concludes with a real data analysis example.

To keep the ensuing discussion as simple as possible, we focus on missing data handling for a two-level random slope model with a single predictor at each level. The analysis model is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 w_j + b_{0j} + b_{1j} x_{ij} + e_{ij} \quad (1)$$

where y_{ij} is the outcome score for observation i in cluster j , x_{ij} and w_j are level-1 and level-2 predictors, respectively, β_0 is the intercept, and β_1 and β_2 are the fixed effect slope coefficients for the two predictors. Consistent with the standard multilevel model formulation, we assume that the intercept and slope residuals, b_{0j} and b_{1j} , respectively, are multivariate normal with zero means and an unstructured covariance matrix Σ_b . Similarly, the within-cluster residuals are normal with common variance σ_e^2 .

As an aside, readers may question our focus on multiple imputation, given that maximum likelihood is the default estimator in most multilevel software programs. Maximum likelihood is applicable to a wide range of single-level missing data problems, but it is arguably less flexible for multilevel missing data. Dedicated software programs that operate within the traditional mixed model framework allow incomplete outcomes (e.g., a longitudinal analysis where the number and configuration of measurements differs across respondents), but the absence of distributional assumptions for explanatory variables forces programs to exclude observations or clusters with missing predictor scores. The HLM program (Raudenbush, Bryk, & Congdon, 2013) appears to be the only dedicated multilevel software package that can accommodate incomplete explanatory variables, but this functionality is limited to random intercept models with normally distributed predictors (Shin & Raudenbush, 2007, 2010, 2013). The multilevel structural equation modeling framework (Mehta & Neale, 2005; Rabe-Hesketh, Skondral, & Zheng, 2012; Stapleton, 2013) is a second option for fitting random coefficient models with missing data. *Mplus* (Muthén & Muthén, 1998–2017) is the only package we know of that can accommodate random slope predictors with

missing data, but no technical documentation is available describing this functionality. Maximum likelihood estimators for incomplete multilevel data tend to invoke complicated transformations of the population joint distribution (Shin & Raudenbush, 2007, 2010), making it difficult to understand and explain their performance without also knowing their technical underpinnings.¹ As such, we focus on multiple imputation because it is arguably a more mature technology for incomplete multilevel data.

Joint model imputation with random covariance matrices

The joint imputation framework is based on a multivariate regression model that is typically parameterized by regressing the set of incomplete variables on the set of complete variables, or by treating all variables as outcomes regardless of missing data pattern (Asparouhov & Muthén, 2010; Carpenter & Kenward, 2013; Goldstein et al., 2009; Schafer & Yucel, 2002; Yucel, 2011). We describe the latter parameterization, focusing on a variant of the joint model that introduces heterogeneous within-cluster covariance matrices (Yucel, 2011). This approach holds promise for random coefficient models but has not received much attention in the literature. It has to be noted that we alter our notation in this section (e.g., replacing β s with γ s, etc.) to clearly differentiate the imputation and analysis models, and we use alphanumeric subscripts to associate parameters to specific variables (e.g., $\gamma_{(y)}$ and $\gamma_{(x)}$).

The joint imputation model for the random coefficient analysis defines each observation as the sum of a grand mean and one or more normally distributed deviation scores

$$\begin{aligned} y_{ij} &= \gamma_{(y)} + u_{j(y)} + \varepsilon_{ij(y)} \\ x_{ij} &= \gamma_{(x)} + u_{j(x)} + \varepsilon_{ij(x)} \\ w_j &= \gamma_{(w)} + u_{j(w)} \end{aligned} \tag{2}$$

$$\begin{pmatrix} u_{j(y)} \\ u_{j(x)} \\ u_{j(w)} \end{pmatrix} \sim \mathcal{N}(0, \Sigma_u) \quad \begin{pmatrix} \varepsilon_{ij(y)} \\ \varepsilon_{ij(x)} \end{pmatrix} \sim \mathcal{N}(0, \Sigma_{\varepsilon_j})$$

where $\gamma_{(\cdot)}$ is the regression intercept or grand mean, $u_{j(\cdot)}$ is a level-2 deviation score capturing the difference between a latent cluster mean and the grand mean (e.g., $u_{j(y)} = \mu_{j(y)} - \gamma_{(y)}$), and $\varepsilon_{ij(\cdot)}$ is a within-cluster deviation between an observation and its

corresponding latent group mean (e.g., $\varepsilon_{ij(y)} = y_{ij} - \mu_{j(y)}$). Furthermore, Σ_u is the level-2 covariance matrix of the between-cluster deviation scores, and Σ_{ε_j} is the within-cluster covariance matrix for cluster j , as follows.

$$\begin{aligned} \Sigma_u &= \begin{pmatrix} \sigma_{u_{(y)}}^2 & \sigma_{u_{(y,x)}} & \sigma_{u_{(y,w)}} \\ \sigma_{u_{(x,y)}} & \sigma_{u_{(x)}}^2 & \sigma_{u_{(x,w)}} \\ \sigma_{u_{(w,y)}} & \sigma_{u_{(w,x)}} & \sigma_{u_{(w)}}^2 \end{pmatrix} \quad \Sigma_{\varepsilon_j} = \begin{pmatrix} \sigma_{\varepsilon_{j(y)}}^2 & \sigma_{\varepsilon_{j(y,x)}} \\ \sigma_{\varepsilon_{j(x,y)}} & \sigma_{\varepsilon_{j(x)}}^2 \end{pmatrix} \\ \Sigma_{\varepsilon_j} &\sim \mathcal{IW}(v + n_j, \mathbf{S} + \mathbf{S}_{\varepsilon_j}) \end{aligned} \tag{3}$$

where \mathcal{IW} denotes an inverse Wishart probability distribution for cluster j 's covariance matrix, v and \mathbf{S} are the degrees of freedom and scale matrix (i.e., jointly, these quantities define the center and spread) of the hierarchically specified inverse Wishart distribution of all random covariance matrices, n_j is the number of observations in cluster j , and $\mathbf{S}_{\varepsilon_j}$ is the within-cluster residual sum of squares and crossproducts matrix for cluster j . Importantly, allowing the covariance matrices to vary across clusters differs from the classic joint model formulation (Schafer, 2001; Schafer & Yucel, 2002) and provides a mechanism for preserving random slope variation.

The computational machinery for joint model imputation uses an iterative Markov chain Monte Carlo (MCMC) algorithm consisting of two major steps. At each iteration, the algorithm first uses Bayesian estimation to generate model parameters and level-2 deviation scores, conditional on the filled-in data from the previous cycle, after which it updates imputations, conditional on the current estimates. The estimation sequence generates a coefficient vector, γ , between-cluster covariance matrix, Σ_u , level-2 deviation scores, $\mathbf{u} = (u_{(y)}, u_{(x)}, u_{(w)})$, and a set of J within-cluster covariance matrices, $\Sigma_{\varepsilon} = (\Sigma_{\varepsilon_1}, \dots, \Sigma_{\varepsilon_J})$. The distribution of the random covariance matrices across all clusters requires two additional parameters, the degrees of freedom and scale matrix, v and \mathbf{S} (i.e., parameters defining the distribution's center and spread, respectively). Roughly speaking, v and \mathbf{S} can be viewed as quantities that define a pooled level-1 covariance matrix.

The Bayesian estimation sequence views model parameters as random variables, and it uses Monte Carlo computer simulation to "sample" plausible estimates from a series of probability distributions. Specifically, each iteration t of the MCMC algorithm applies the following estimation steps to the filled-in data

¹We included the *Mplus* implementation of maximum likelihood estimation in a subset of our simulations, and the method was uniformly biased. We do not report these results here because there is currently no technical documentation that helps us understand and explain the results.

1. Sample a coefficient vector $\dot{\boldsymbol{\gamma}}^{(t)}$ from $\mathcal{N}\left(\boldsymbol{\gamma} | \dot{\mathbf{u}}^{(t-1)}, \dot{\boldsymbol{\Sigma}}_e^{(t-1)}, \text{data}\right)$
 2. Sample the level-2 covariance matrix $\dot{\boldsymbol{\Sigma}}_u^{(t)}$ from $IW\left(\boldsymbol{\Sigma}_u | \dot{\mathbf{u}}^{(t-1)}, \text{data}\right)$
 3. Sample the scale matrix $\dot{\mathbf{S}}^{(t)}$ from $\mathcal{JW}\left(\mathbf{S} | \dot{v}^{(t-1)}, \dot{\boldsymbol{\Sigma}}_e^{(t-1)}, \text{data}\right)$
 4. Sample the scalar degrees of freedom $\dot{v}^{(t)}$ from $f\left(v | \dot{\boldsymbol{\Sigma}}_e^{(t-1)}, \text{data}\right)$
 5. Sample level-2 residuals $\dot{\mathbf{u}}^{(t)}$ from $\mathcal{N}\left(\mathbf{u} | \dot{\boldsymbol{\gamma}}^{(t)}, \dot{\boldsymbol{\Sigma}}_u^{(t)}, \dot{\boldsymbol{\Sigma}}_e^{(t-1)}, \text{data}\right)$
 6. Sample level-1 covariance matrices $\dot{\boldsymbol{\Sigma}}_{\epsilon_j}^{(t)}$ from $\mathcal{JW}\left(\boldsymbol{\Sigma}_{\epsilon_j} | \dot{\boldsymbol{\gamma}}^{(t)}, \dot{\mathbf{S}}^{(t)}, \dot{v}^{(t)}, \dot{\mathbf{u}}^{(t)}, \text{data}\right)$
- (4)

where \mathcal{N} and \mathcal{JW} are normal and inverse Wishart distributions, respectively, and the dot accents denote synthetic estimates generated via Monte Carlo computer simulation (van Buuren, 2012). The full conditional distributions of $\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_u$, and \mathbf{u} are given in a variety of resources that describe the classic joint model formulation (Carpenter & Kenward, 2013; Goldstein et al., 2009; Schafer & Yucel, 2002; Yucel, 2011).² The conditional distributions of \dot{v} and \mathbf{S} are based on aggregate information from $\dot{\boldsymbol{\Sigma}}_e$ (Yucel, 2011, p. 359), and the final sampling step draws random covariance matrices from a distribution that leverages this pooled information as well as cluster-specific variation (Equation 3). Finally, we use the generic symbol f to denote the conditional distribution of v because this quantity does not follow a standard parametric form. In addition to Yucel (2011), interested readers can consult Carpenter and Kenward (2013) and Quartagno and Carpenter (2016b) for additional technical details, including a discussion of prior distributions.

After estimating the parameters and random effects, a final MCMC step updates missing values by sampling imputations from a multivariate normal distribution, the center, and spread of which depend on the multilevel model parameters and level-2 residual terms.

²Cluster-specific covariance matrices replace a common within-cluster matrix in the classic joint model formulation, but the probability distributions for $\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_u$, and \mathbf{u} are otherwise the same. The joint model uses improper uniform prior distributions for the coefficients and level-2 residual terms. Both the level-1 and level-2 covariance matrices use inverse Wishart prior distributions with identity scale matrices and degrees of freedom equal to the dimension of the respective matrices.

$$\begin{pmatrix} \dot{y}_{ij}^{(t)} \\ \dot{x}_{ij}^{(t)} \\ \dot{w}_j^{(t)} \end{pmatrix} \sim \mathcal{N}\left(\dot{\boldsymbol{\gamma}}^{(t)} + \dot{\mathbf{u}}_j^{(t)}, \dot{\boldsymbol{\Sigma}}_{\epsilon_j}^{(t)}\right) \tag{5}$$

Recall that w has no within-cluster variation, and hence each \dot{w}_j is simply the sum of a grand mean and between-cluster residual term (i.e., $\dot{w}_j = \dot{\gamma}_{(w)} + \dot{u}_{j(w)}$ because this variable has zero entries in the within-cluster covariance matrix). The updated version of the completed data set carries forward to the next iteration, where MCMC estimation steps provide new parameter estimates and residual terms for the next round of imputation. The desired number of imputed data sets is obtained by iterating the MCMC algorithm and culling data sets at specified intervals in the chain (Schafer, 1997; Schafer & Olsen, 1998).

Compatibility of imputation and analysis models

A vital consideration with multiple imputation is whether (and how) the imputation model preserves the features of the analysis; methodologists have described this issue as congeniality (Meng, 1994; Schafer, 2003) and more recently as compatibility (Bartlett, Seaman, White, & Carpenter, 2014; Carpenter & Kenward, 2013). Yucel (2011) explained that modeling heterogeneous covariance matrices is “mimicking the idea underlying random effects” (p. 352), but he did not specifically examine random coefficient models.³ The covariance matrix formulation makes it difficult to see which parts of the imputation model are preserving important features of the random slope analysis, but we can informally explore compatibility by using estimates of $\boldsymbol{\Sigma}_{\epsilon_j}$ and $\boldsymbol{\Sigma}_u$ to define within- and between-cluster regression parameters that are analogous to those of the analysis model. To be clear, we are not examining whether the random coefficient analysis and imputation model are formally compatible with the joint distribution of the population. Rather, we are deriving regression model parameters as functions of $\boldsymbol{\Sigma}_{\epsilon_j}$ and $\boldsymbol{\Sigma}_u$ to highlight major similarities and differences between the imputation and the analysis models. Quartagno and Carpenter (2016b) provide a similar assessment in the context of meta-analysis models.

To begin, consider the level-1 part of the imputation model from Equation 2. The $\boldsymbol{\Sigma}_{\epsilon_j}$ matrix contains variances and covariances of within-cluster deviation scores for cluster j , which we use to define a level-1

³Quartagno and Carpenter (2016b) considered the application of random covariance matrices to various meta-analysis models.

regression model with $\varepsilon_{ij(x)}$ predicting $\varepsilon_{ij(y)}$

$$\begin{aligned}
 (y_{ij} - \mu_{j(y)}) &= \frac{\sigma_{\varepsilon_{ij(x,y)}}}{\sigma_{\varepsilon_{ij(x)}}^2} (x_{ij} - \mu_{j(x)}) + e_{ij} = \pi_{1j}(x_{ij} - \mu_{j(x)}) + e_{ij} \\
 \sigma_{e_j}^2 &= \sigma_{\varepsilon_{ij(y)}}^2 - \pi_{1j}^2 \sigma_{\varepsilon_{ij(x)}}^2
 \end{aligned}
 \tag{6}$$

where π_{1j} is the slope coefficient for cluster j , e_{ij} is a within-cluster residual for observation i in cluster j , and $\sigma_{e_j}^2$ is the corresponding residual variance for that group. Equation 6 clearly shows that the random covariance matrices are responsible for preserving slope variation as these parameters imply a unique regression coefficient for each group, π_{1j} . Importantly, the level-1 model has no mean structure or intercept that varies across level-2 units because the within-cluster deviation scores have an expectation of zero in all groups. To facilitate our subsequent comparison with the analysis model, we express the cluster-specific slopes as a function of an average coefficient and a between-cluster residual term, $\pi_{1j} = \gamma_1 + g_{1j}$. Substituting the right-hand side of this expression into Equation 6 gives the following level-1 regression model.

$$(y_{ij} - \mu_{j(y)}) = \gamma_1(x_{ij} - \mu_{j(x)}) + g_{1j}(x_{ij} - \mu_{j(x)}) + e_{ij} \tag{7}$$

Next, the Σ_u matrix contains variances and covariances of between-cluster deviation scores, which we can use to define a level-2 multiple regression model with $u_{j(x)}$ and $u_{j(w)}$ predicting $u_{j(y)}$

$$\begin{aligned}
 (\mu_{j(y)} - \gamma(y)) &= \left[(\mu_{j(x)} - \gamma(x)) \quad (w_j - \gamma(w)) \right] \\
 &\quad \begin{bmatrix} \sigma_{u_{(x)}}^2 & \sigma_{u_{(x,w)}} \\ \sigma_{u_{(w,x)}} & \sigma_{u_{(w)}}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{u_{(x,y)}} \\ \sigma_{u_{(w,y)}} \end{bmatrix} + g_{0j} \\
 &= \left[(\mu_{j(x)} - \gamma(x)) \quad (w_j - \gamma(w)) \right] \begin{bmatrix} \gamma_2 \\ \gamma_3 \end{bmatrix} + g_{0j} \\
 \sigma_{g_0}^2 &= \sigma_{u_{(y)}}^2 - \begin{bmatrix} \gamma_2 \\ \gamma_3 \end{bmatrix}' \begin{bmatrix} \sigma_{u_{(x)}}^2 & \sigma_{u_{(x,w)}} \\ \sigma_{u_{(w,x)}} & \sigma_{u_{(w)}}^2 \end{bmatrix} \begin{bmatrix} \gamma_2 \\ \gamma_3 \end{bmatrix}
 \end{aligned}
 \tag{8}$$

where γ_2 and γ_3 are between-cluster slope coefficients, g_{0j} is a residual that captures unexplained variation in the cluster means, and $\sigma_{g_0}^2$ is the residual variance.

Finally, combining the right- and left-hand sides of the within- and between-cluster regression equations and adding the grand mean of y to each side of the resulting expression gives a model with the same form as the random coefficient analysis.

$$\begin{aligned}
 y_{ij} = \gamma(y) &+ \gamma_1(x_{ij} - \mu_{j(x)}) + \gamma_2(\mu_{j(x)} - \gamma(x)) + \gamma_3(w_j - \gamma(w)) \\
 &+ g_{0j} + g_{1j}(x_{ij} - \mu_{j(x)}) + e_{ij}
 \end{aligned}
 \tag{9}$$

Comparing Equations 1 and 9 reveals two important differences between the imputation and analysis

models, one of which is potentially problematic. First, Equation 9 includes an additional regression slope that partitions the relationship between x and y into distinct within- and between-cluster components (i.e., γ_1 and γ_2 , respectively), as in the classic contextual effects model (Enders, 2013; Kreft, de Leeuw, & Aiken, 1995; Longford, 1989; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). Second, the imputation model accommodates a heterogeneous within-cluster residual variance structure (e.g., $\sigma_{e_j}^2$ from Equation 6), whereas the analysis model assumes constant variance. Incompatibilities such as these that arise from a rich imputation model with ancillary parameters are usually not problematic and may even be beneficial (Collins, Schafer, & Kam, 2001; Meng, 1994; Schafer, 2003). Although Equations 6–9 clearly confirm that random covariance matrices mimic random effects, they highlight an important caveat. Specifically, because Σ_{e_j} and the corresponding within-cluster regression model do not include a mean structure or intercept that varies across level-2 units, between-cluster variation in π_{1j} (or equivalently, g_{1j}) is necessarily orthogonal to functions of Σ_u , notably the between-cluster residual term, g_{0j} . Consequently, an imputation model based on random covariance matrices could only be approximately compatible with a random coefficient analysis with uncorrelated intercepts and slopes (i.e., a block-diagonal Σ_b). This restriction could be detrimental to analyses where the level-2 covariance captures an interesting substantive phenomenon (e.g., growth models where initial status informs change rates).

Fully conditional specification

Although joint model approach uses a multivariate regression model to generate imputations, fully conditional specification cycles through incomplete variables one at a time, generating imputations from a sequence of univariate multilevel models that feature an incomplete variable regressed on all remaining variables, complete and previously imputed (Enders et al., 2016, 2017; van Buuren, 2011, 2012). Consistent with the previous section, we alter our notation (e.g., replacing β s with γ s, etc.) to clearly differentiate the imputation and analysis models, and we use alphanumeric subscripts to associate parameters to specific variables (e.g., $\gamma_{(y)}$ and $\gamma_{(x)}$).

Returning to the random slope analysis from Equation 1, fully conditional specification applies the following imputation models

$$y_{ij} = \gamma_{0(y)} + \gamma_{1(y)}x_{ij} + \gamma_{2(y)}w_j + \gamma_{3(y)}\bar{x}_j + u_{0j(y)} + u_{1j(y)}x_{ij} + \varepsilon_{ij(y)} \quad (10)$$

$$\begin{pmatrix} u_{0j(y)} \\ u_{1j(y)} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{u(y)}) \quad \varepsilon_{ij(y)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\varepsilon(y)}^2)$$

$$x_{ij} = \gamma_{0(x)} + \gamma_{1(x)}y_{ij} + \gamma_{2(x)}w_j + \gamma_{3(x)}\bar{y}_j + u_{0j(x)} + u_{1j(x)}y_{ij} + \varepsilon_{ij(x)} \quad (11)$$

$$\begin{pmatrix} u_{0j(x)} \\ u_{1j(x)} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{u(x)}) \quad \varepsilon_{ij(x)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\varepsilon(x)}^2)$$

$$w_j = \gamma_{0(w)} + \gamma_{1(w)}\bar{y}_j + \gamma_{2(w)}\bar{x}_j + u_{0j(w)} \quad (12)$$

$$u_{0j(w)} \sim \mathcal{N}(\mathbf{0}, \sigma_{u(w)}^2)$$

where \bar{x}_j and \bar{y}_j are the level-2 cluster means computed from the imputed data for cluster j , $u_{0j(\cdot)}$ and $u_{1j(\cdot)}$ are now between-cluster residual terms, and $\varepsilon_{ij(\cdot)}$ is a within-cluster residual with constant variance $\sigma_{\varepsilon(\cdot)}^2$. Notice that the round-robin imputation scheme features the outcome from one equation (the target of imputation) as a predictor in all other equations. Importantly, the models use a “reverse random coefficient” approach (Grund et al., 2016) preserving random slope variation by which x is a random predictor in the y imputation model, and *vice versa*. Also, note that latent rather than manifest variable group means can be used to preserve between-cluster variation among the incomplete variables (Grund, Lüdtke, & Robitzsch, 2017). Analytic and simulation results suggest that manifest cluster means provide good performance in most situations but can introduce slight biases with low intraclass correlations, few observations per cluster, and extremely unbalanced cluster sizes (Grund et al., 2017).

Like joint model imputation, fully conditional specification employs an iterative MCMC algorithm that generates Bayesian estimates of model parameters and level-2 residual terms, conditional on the filled-in data, after which it updates the imputations, conditional on current estimates and residuals. Each iteration applies these estimation and imputation steps to the incomplete variables in a sequence. The model parameters for an incomplete level-1 variable q include a coefficient vector, $\boldsymbol{\gamma}_{(q)}$, level-2 residuals, $\mathbf{u}_{(q)}$, within-cluster residual variance $\sigma_{\varepsilon(q)}^2$, and the level-2 residual covariance matrix $\boldsymbol{\Sigma}_{u(q)}$, whereas level-2 imputation requires only the coefficients and residual variance (Equation 12). Consistent with the joint model, Monte Carlo computer simulation samples plausible estimates from the series of probability distributions given below.

1. Sample a coefficient vector $\dot{\boldsymbol{\gamma}}_{(q)}^{(t)}$ from $\mathcal{N}(\boldsymbol{\gamma}_{(q)} | \dot{\mathbf{u}}_{(q)}^{(t-1)}, \dot{\sigma}_{\varepsilon(q)}^{2(t-1)}, \text{data})$
 2. Sample level-2 residuals $\dot{\mathbf{u}}_{(q)}^{(t)}$ from $\mathcal{N}(\mathbf{u}_{(q)} | \dot{\boldsymbol{\gamma}}_{(q)}^{(t)}, \dot{\sigma}_{\varepsilon(q)}^{2(t-1)}, \dot{\boldsymbol{\Sigma}}_{u(q)}^{(t-1)}, \text{data})$
 3. Sample level-1 residual variance $\dot{\sigma}_{\varepsilon(q)}^{2(t)}$ from $\mathcal{JW}(\sigma_{\varepsilon(q)}^2 | \dot{\boldsymbol{\gamma}}_{(q)}^{(t)}, \dot{\mathbf{u}}_{(q)}^{(t)}, \text{data})$
 4. Sample the level-2 covariance matrix $\dot{\boldsymbol{\Sigma}}_{u(q)}^{(t)}$ from $\mathcal{JW}(\boldsymbol{\Sigma}_{u(q)} | \dot{\mathbf{u}}_{(q)}^{(t)}, \text{data})$
- (13)

As mentioned before, the dot accents denote synthetic values generated via Monte Carlo simulation (van Buuren, 2012). A variety of resources provide the technical details about these distributions (Browne & Draper, 2000; Enders et al., 2017; van Buuren, 2012).⁴

After estimating the parameters and random effects for incomplete variable q , a final MCMC step updates missing values by sampling imputations from a univariate normal distribution, the center and spread of which depend on the model parameters. The imputation steps for our example are as follows.

$$y_{ij}^{(t)} \sim N(\dot{\gamma}_{0(y)} + \dot{\gamma}_{1(y)}\dot{x}_{ij}^{(t-1)} + \dot{\gamma}_{2(y)}\dot{w}_j^{(t-1)} + \dot{\gamma}_{3(y)}\bar{x}_j^{(t-1)} + \dot{u}_{0j(y)} + \dot{u}_{1j(y)}\dot{x}_{ij}^{(t-1)}, \dot{\sigma}_{\varepsilon(y)}^2) \quad (14)$$

$$\dot{x}_{ij}^{(t)} \sim N(\dot{\gamma}_{0(x)} + \dot{\gamma}_{1(x)}\dot{y}_{ij}^{(t)} + \dot{\gamma}_{2(x)}\dot{w}_j^{(t-1)} + \dot{\gamma}_{3(x)}\bar{y}_j^{(t)} + \dot{u}_{0j(x)} + \dot{u}_{1j(x)}\dot{y}_{ij}^{(t)}, \dot{\sigma}_{\varepsilon(x)}^2) \quad (15)$$

$$\dot{w}_j^{(t)} \sim N(\dot{\gamma}_{0(w)} + \dot{\gamma}_{1(w)}\bar{y}_j^{(t)} + \dot{\gamma}_{2(w)}\bar{x}_j^{(t)}, \dot{\sigma}_{u(w)}^2) \quad (16)$$

To simplify notation, we omit the iteration superscript from the parameters and residual terms because these quantities always originate from iteration t .

Compatibility of imputation and analysis models

Consistent with the joint model, the imputation models in Equations 10 and 11 are more flexible than necessary because they include a contextual

⁴As described in the technical appendix of Enders et al. (2017) improper uniform prior distributions are adopted for the coefficients and level-2 residual terms. The inverse Wishart distribution for the level-2 covariance matrices $\boldsymbol{\Sigma}_{u(q)}$ implemented an identity prior with $p+1$ degrees of freedom, which is a common recommendation in the literature. Similarly, the inverse Wishart prior for the residual variance uses a scale parameter of unity with two degrees of freedom.

effects-type specification that partitions the relationship between x and y into unique within- and between-cluster components (Carpenter & Kenward, 2013; Enders et al., 2016, 2017).⁵ The primary problem with fully conditional specification is its use of reverse regression to preserve random slope variation (e.g., y is a random predictor of x , and *vice versa*). As noted previously, simulation studies suggest that this strategy can produce negatively biased slope variance estimates (Enders et al., 2016, 2017; Grund et al., 2016), and an incorrect distributional assumption appears to be responsible for this problem (Enders et al., 2017, October). Specifically, when a random predictor is missing, assuming a normal distribution for the level-2 residuals in the analysis or y imputation model precludes the possibility that the corresponding residuals from the reverse regression are also normal. The estimation of MCMC violates this assumption when it samples level-2 residuals for x 's imputation model in Equation 11. A related issue occurs in the single-level context with reverse regressions that include a product term (Bartlett et al., 2014; Gelman & Raghunathan, 2001).

To provide additional insight into the incompatibility problem, we can express the random effects from the reverse regression as a function of the corresponding terms from the analysis model. To begin, we solve for x in the analysis model as follows.

$$x_{ij} = \frac{y_{ij} - (\beta_0 + \beta_2 w_j + b_{0j} + e_{ij})}{\beta_1 + b_{1j}} \quad (17)$$

Next, the right-hand side of this expression replaces x_{ij} in the left-hand side of the reverse random coefficient model in Equation 11. For simplicity, we omit the cluster means from the right-hand side of the imputation model as they do not appear in the analysis. Finally, solving for the imputation model's random effects gives the following expressions.

$$\begin{aligned} \gamma_{0(x)} + u_{0j(x)} &= \frac{-(\beta_0 + b_{0j})}{\beta_1 + b_{1j}} \\ \gamma_{1(x)} + u_{1j(x)} &= \frac{1}{\beta_1 + b_{1j}} \\ \varepsilon_{ij(x)} &= \frac{-e_{ij}}{\beta_1 + b_{1j}} \end{aligned} \quad (18)$$

Equation 18 shows that assuming normality for b_{0j} and b_{1j} precludes the possibility that $u_{0j(x)}$, $u_{1j(x)}$, and $\varepsilon_{ij(x)}$ are normal because neither the inverse of a normal variate nor the ratio of two Gaussians are members of the normal distribution family. In fact, the

above expressions can be quite skewed and kurtotic. Consequently, applying the MCMC estimation steps from Equation 13 to the reverse random coefficient model violates a key distributional assumption that ends up impacting imputation quality. As an aside, this issue does not occur when x is complete, nor does it occur in random intercept models.

Computer simulation study

To date, the only study to examine imputation with random covariance matrices did so in the context of individual patient data meta-analysis (Quartagno & Carpenter, 2016b), and it did not examine fully conditional specification. Given the compatibility issues highlighted previously, it is important to understand the relative strengths and weaknesses of these two approaches as they will not necessarily produce comparable results. To address this gap in the literature, we designed a Monte Carlo simulation study with two within-participants factors (missing data handling method and cause of missingness) and five between-participants factors: intraclass correlation ($ICC = .10$ and $.50$), residual correlation between the random intercepts and slopes (0 and $.50$), number of clusters ($J = 15, 30,$ and 100), within-cluster sample size ($n_j = 5, 15, 30,$ and 50), and MAR missing data rate (10, 20, and 30%). We generated 2000 artificial data sets within each of the 144 between-participants design cells, resulting in 288,000 replications.

The previous studies of fully conditional specification (Enders et al., 2017; Grund et al., 2016) and recommendations from the literature guided our design choices. For example, the ICC values of $.10$ and $.50$ are common in published applications and are consistent with those from cross-sectional and repeated measures designs, respectively (Gulliford, Ukoumunne, & Chinn, 1999; Hedges & Hedberg, 2007; Murray & Blitstein, 2003; Spybrook et al., 2011). In choosing the level-1 and level-2 sample sizes, we attended to recommendations from the literature (Kreft & de Leeuw, 1998; Maas & Hox, 2005), but we also selected small values capable of introducing bias (McNeish & Stapleton, 2016). It is difficult to identify typical missing data rates because published manuscripts do not routinely report this information, and hence we selected the values large enough to expose small-sample biases in variance estimates (Enders et al., 2017).

Data generation

The random coefficient analysis from Equation 1 served as the true population model for the

⁵Carpenter and Kenward (2013, p. 220) attribute the idea of including cluster means as predictors to a personal communication from Ian White.

Table 1. Population correlations for computer simulation.

	<i>y</i>	<i>X</i>	<i>w</i>	<i>av</i> ₁	<i>av</i> ₂
1. <i>y</i>	1.0	.4	.3	.4	0
2. <i>x</i>	.4	1.0	.3	0.4	0
3. <i>w</i>	.3	.3	1.0	.0	.4
4. <i>av</i> ₁	.4	.4	.0	1.0	0
5. <i>av</i> ₂	0	0	.4	0	1.0

simulations. The data generation procedure also included level-1 and level-2 auxiliary variables, *a*₁ and *a*₂, respectively. As described later, *a*₁ determined missingness probabilities for *x* in one of the simulations (*y* was the cause of missingness in the second), and *a*₂ always served as the cause of missingness on *w*. To specify the effect sizes on a convenient metric, we used the values listed in Table 1 to define the total correlations among the variables. We chose correlations of .30 to align with Cohen’s (1988) medium effect size benchmark, and we set the correlations between the analysis and the auxiliary variables to .40 to ensure that bias would result if these additional variables are omitted from imputation (Collins et al., 2001). We also set the correlation between *x* and *y* to .40 and hence we could manipulate the cause of missingness (i.e., *a*₁ or *y*) without changing the strength of the MAR selection mechanism.

We solved for the population parameters using the following steps. First, we fixed the total variance of all variables to unity. The level-2 predictor *w* and the auxiliary variables had all variation assigned to a single level, whereas *x* and *y* had variation at both levels, with between-cluster variance set to either .10 or .50 (ICC = .10 and .50, respectively). Second, we computed the level-1 and the level-2 covariance matrices, $\Sigma^{(1)}$ and $\Sigma^{(2)}$, by pre- and postmultiplying the total correlation matrix by a diagonal matrix containing the level-1 or level-2 standard deviations (e.g., in the ICC = .10 condition the level-1 and level-2 standard deviations of *x* (or *y*) were $\sqrt{.90}$ and $\sqrt{.10}$,

respectively). The total covariance matrix was the sum of the within- and between-cluster matrices (i.e., $\Sigma = \Sigma^{(1)} + \Sigma^{(2)}$). Third, we used the following expressions to solve for the model parameters

$$\begin{aligned}
 \beta &= \Sigma_{(xw)}^{-1} \Sigma_{(y,xw)} \\
 \sigma_{b_1}^2 &= .40 \left(\sigma_y^{2(2)} \right) \\
 \sigma_e^2 &= \sigma_y^{2(1)} - \beta' \Sigma_{(xw)}^{(1)} \beta - \sigma_{b_1}^2 \sigma_x^{2(1)} \\
 \sigma_{b_0}^2 &= \sigma_y^{2(2)} - \beta' \Sigma_{(xw)}^{(2)} \beta - \sigma_{b_1}^2 \sigma_x^{2(2)} \\
 \sigma_{b_0 b_1} &= r_{b_0 b_1} (\sigma_{b_0}^2 \sigma_{b_1}^2)^{-2} \\
 \beta_0 &= 5
 \end{aligned}
 \tag{19}$$

where $\Sigma_{(xw)}$, $\Sigma_{(xw)}^{(1)}$, and $\Sigma_{(xw)}^{(2)}$ are submatrices containing the variances and covariances of the predictors, $\Sigma_{(y,xw)}$ is a vector containing covariances between the outcome and the predictors, $\sigma_y^{2(1)}$ and $\sigma_y^{2(2)}$ (or $\sigma_x^{2(1)}$ and $\sigma_x^{2(2)}$) are within- and between-cluster variances, respectively, $r_{b_0 b_1}$ is the correlation between the intercepts and the slopes. We somewhat arbitrarily set the slope variance equal to 40% of the between-cluster variance, reasoning that this parameter should be large enough to reveal important differences between the imputation methods.

Applying the previous expressions produced the parameter values in Equation 20. We then used the following steps to generate the data: (a) sample the within- and between-cluster components of the predictor variables from a normal distribution with zero means and covariance matrices of $\Sigma_{(xw)}^{(1)}$ and $\Sigma_{(xw)}^{(2)}$, respectively, (b) compute *x* as the sum of its within- and between-cluster deviation scores, (c) sample *b*_{0*j*} and *b*_{1*j*} from a normal distribution with zero means and a covariance matrix Σ_b , (d) sample *e*_{*ij*} from a normal distribution with variance σ_e^2 , and (e) compute *y*_{*ij*} as a weighted sum of *x*, *w*, *b*_{0*j*}, *b*_{1*j*}, and *e*_{*ij*}, as in Equation 20. The R data generation script is available upon request from the first author.

$$\begin{aligned}
 &y_{ij} = 5.00 + .395(x_{ij}) + .057(w_j) + b_{0j} + b_{1j}x_{ij} + e_{ij} \\
 \text{ICC} = .10 &\quad \Sigma_b = \begin{bmatrix} .073 & .027 \\ .027 & .040 \end{bmatrix} \text{ or } \begin{bmatrix} .073 & 0 \\ 0 & .040 \end{bmatrix}, \sigma_e^2 = .734 \\
 &y_{ij} = 5.00 + .372(x_{ij}) + .133(w_j) + b_{0j} + b_{1j}x_{ij} + e_{ij} \\
 \text{ICC} = .50 &\quad \Sigma_b = \begin{bmatrix} .292 & .121 \\ .121 & .200 \end{bmatrix} \text{ or } \begin{bmatrix} .292 & 0 \\ 0 & .200 \end{bmatrix}, \sigma_e^2 = .331
 \end{aligned}
 \tag{20}$$

As noted previously, either the level-1 auxiliary variable *a*₁ or the outcome *y* determined missing values on *x*, and *a*₂ served the same role for *w*. In all cases, higher scores on the cause of missingness produced higher rates of missing data. Using a logistic

regression model, we defined the level-1 missingness probability for each observation as

$$P_{ij} = \frac{e^{\lambda_0 + \lambda_1 z_{ij}}}{1 + e^{\lambda_0 + \lambda_1 z_{ij}}}
 \tag{21}$$

where λ_0 and λ_1 are intercept and slope coefficients, respectively, and z_{ij} is the standardized version of a_1 or y . Using the latent variable formulation for logistic regression (Agresti, 2012; Johnson & Albert, 1999), we derived coefficients that produced the desired missing data rate while maintaining a .50 correlation between the cause of missingness and the underlying normal propensity for missing data. The slope coefficient was $\lambda_1 = 1.815$, and the intercept coefficients for the 10, 20, and 30% missing data rates were approximately $\lambda_0 = -3.22$, -2.10 , and -1.31 . Substituting z into Equation 21 produced an N -row vector of level-1 missingness probabilities, and we subsequently created an N -row vector of missing data indicators (0 = observed, 1 = missing) by sampling binary values from a binomial distribution with success rate equal to each observation's missingness probability. An identical procedure produced missing values on w , such that p_j and z_j (the standardized version of a_2) replaced p_{ij} and z_{ij} in the logistic function. As different auxiliary variables determined level-1 and level-2 missingness, this deletion process induced a general missing data pattern where x or w (or both) could be missing for a particular observation.

Imputation, estimation, and outcomes

We used the R package *jomo* (Quartagno & Carpenter, 2016a) for joint model imputation and the *Blimp* application (Enders et al., 2017) for fully conditional specification. To diagnose convergence of the MCMC algorithm, we used *Blimp* to generate potential scale reduction factors (Gelman & Rubin, 1992) for a single replication in every design cell. These diagnostic values generally reached acceptable levels (e.g., <1.05) in fewer than 1500 iterations. Based on these convergence diagnostics, we generated 20 imputations⁶ (Graham, Olchowski, & Gilreath, 2007) from MCMC chains with 50,000 iterations, such that the first data set was saved after the 2500th iteration and the remaining data sets were saved every 2500 iterations thereafter (i.e., in *Blimp*, we set the burn-in and thinning intervals to 2500). We relied on the default prior distributions of each software package under the assumption that most researchers would do the same. The imputation models for both procedures included the three analysis variables and two auxiliary variables. We used *Mplus 8* (Muthén & Muthén, 1998–2017) to fit the analysis

model to the multiply imputed data sets, and we wrote a custom R program to pool the resulting estimates and standard errors. All computational tasks were executed on UCLA's Shared Hoffman2 Cluster, and we used a Linux shell script to coordinate the simulation components. The simulation code is available upon request from the first author.

We examined three outcomes, relative bias, mean squared error, and confidence interval coverage. Relative bias was computed as the difference between an average estimate and the true population value divided by the true value and multiplied by 100 (i.e., bias as a percentage of the true value). In the conditions with orthogonal random effects, we used the nonzero covariances from Equation 20 as the divisor of the relative bias expression to avoid dividing by zero. The published simulation studies often define relative bias $<10\%$ in absolute value as acceptable (Finch, West, & MacKinnon, 1997; Kaplan, 1988). Mean squared error (MSE) is the average squared difference between an estimate and its true population value. This outcome is an overall measure of accuracy that captures the sum of squared bias and sampling variance. To facilitate interpretation, we defined an MSE ratio as MSE_{JM}/MSE_{FCS} , such that values greater than unity favor fully conditional specification (i.e., the joint model had larger squared errors), whereas values <1 favor the joint model. Finally, 95% confidence interval coverage was computed as the proportion of estimates where the 95% symmetric confidence interval (i.e., the estimate plus or minus 1.96 standard error units) included the true parameter value. Values lower than the nominal 95% rate reflect Type I error inflation (e.g., a coverage value of 90% suggests a twofold increase in Type I errors), whereas values $>95\%$ reflect conservative inference. Confidence interval coverage is an indicator of standard error quality when estimates are unbiased. We consider coverage for only the slope coefficients because the literature argues that symmetric confidence intervals are inappropriate for variance estimates (Maas & Hox, 2005; Snijders & Bosker, 2012).

Results

The results are organized as follows. We begin by discussing the scenario where auxiliary variables determine missingness and the random intercepts and slopes are uncorrelated in the population. This situation is a useful baseline because the random covariance matrix approach should be approximately compatible with the data-generating model. Next, we

⁶When considering the design of the study, we examined a subset of results with 5 and 100 imputations. As bias values were effectively the same in both situations, we adopted Graham et al.'s (2007) recommendation to use 20 imputations.

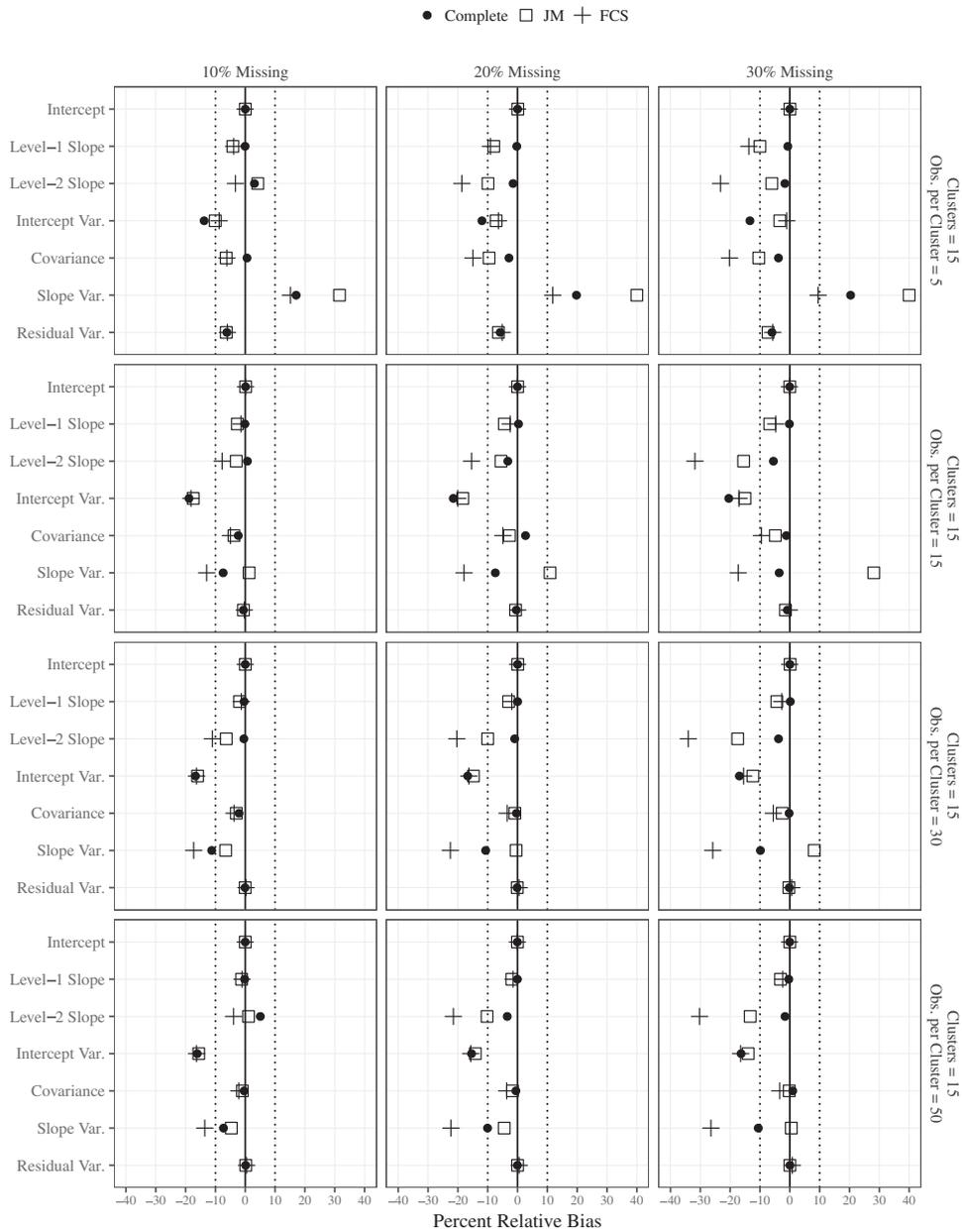


Figure 1. Average relative bias values for design cells with ICC = .10 and 15 clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ± 10 .

examine the same missingness mechanism with correlated random effects. As described previously, this correlation should be detrimental to the joint model because imputation uses fewer parameters than the data-generating process. We did not expect this correlation to meaningfully influence fully conditional specification. Finally, we discuss any differences that result from an MAR mechanism where the outcome variable causes missingness in x . In the interest of space, we use a limited set of figures to highlight the main findings, and we present the results for all

combinations of conditions in the online supplemental material.

Missingness owing to auxiliary variables and uncorrelated random effects

Figures 1 and 2 are trellis plots showing relative bias for the ICC = .10 simulation with 15 and 30 clusters, respectively. In the interest of space, we relegate the 100-cluster plot to the online supplemental material because it was comparable to Figure 2. The figures

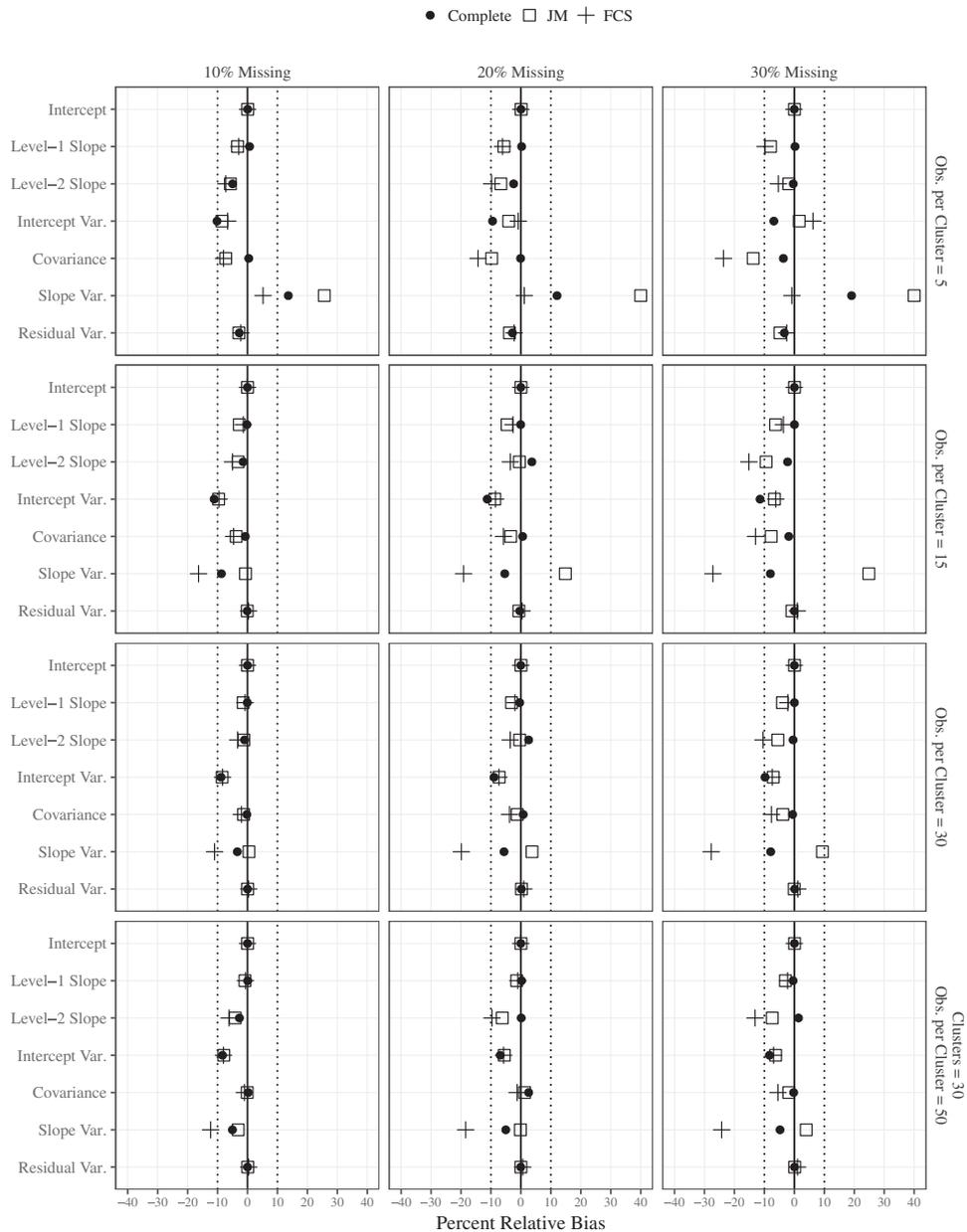


Figure 2. Average relative bias values for design cells with ICC = .10 and 30 clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ± 10 .

highlight a few general trends. Not surprisingly, increasing the number of clusters from 15 to 30 reduced bias, particularly for the β_2 coefficient and the intercept variance, estimates that rely heavily on the number of level-2 units. Similarly, increasing the within-cluster sample size from 5 to 15 reduced bias, but further increasing the group sizes did not impact most parameters. With the exception of the slope variance, using 30 clusters with at least 15 observations per group produced imputation estimates that exhibited roughly the same bias as complete-data estimates, even with 30% of missing data.

Figures 1 and 2 suggest that both imputation approaches had difficulty preserving random slope variation although they erred in different directions and by different amounts. Specifically, the joint model overestimated slope variance, with bias decreasing to reasonable levels (e.g., <10% in absolute value) when the within-cluster sample size was 30 or larger. The reliance on the within-cluster sample size is not surprising, given that joint model uses random level-1 covariance matrices to preserve random slope variation (Equation 6). On the other hand, fully conditional specification consistently underestimated slope

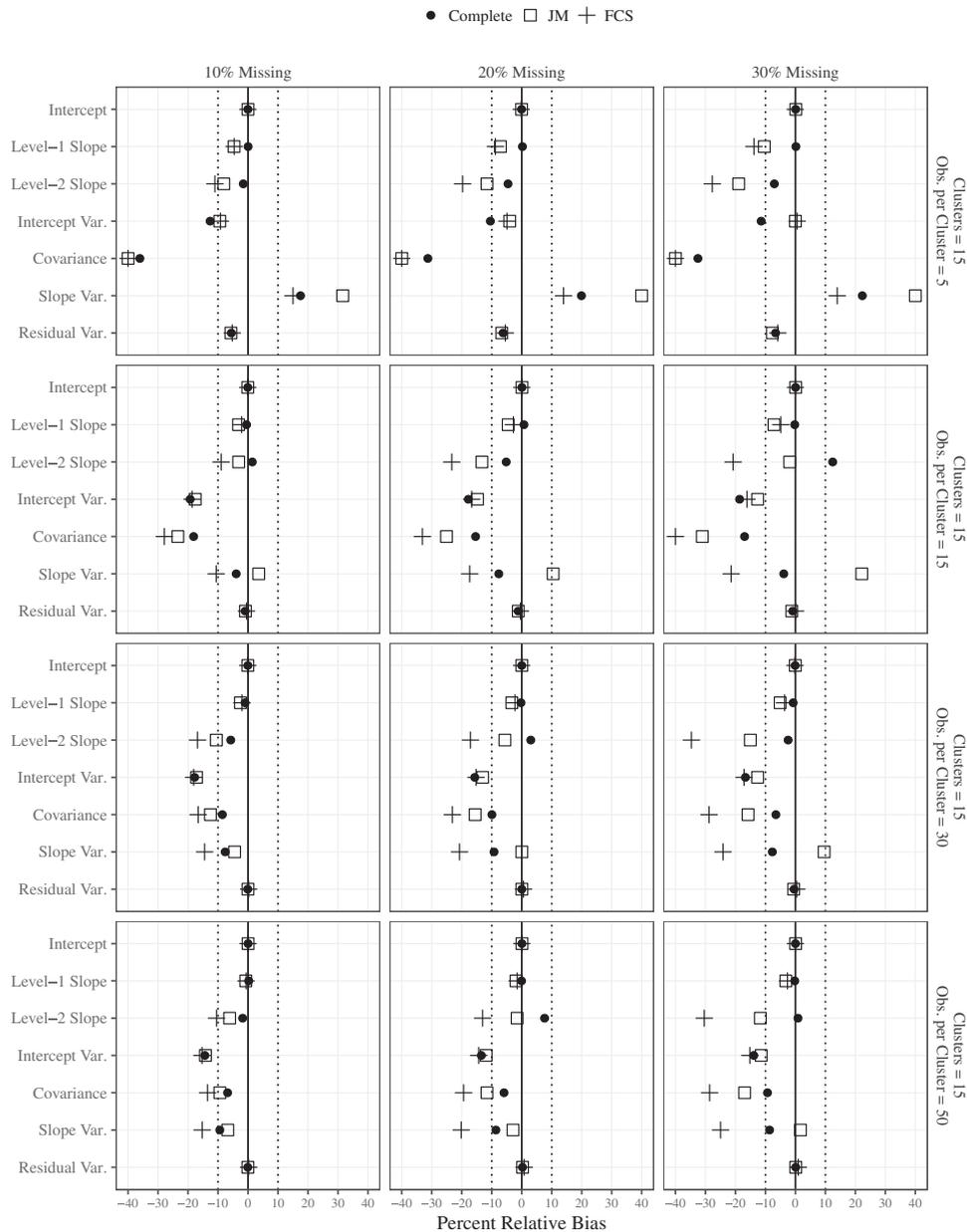


Figure 3. Average relative bias values for design cells with $ICC = .50$ and 15 clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ± 10 .

variance, and the magnitude of this bias did not decrease in larger samples. Rather, bias was largely a function of missingness, with values exceeding 10% in absolute value at a 20% missing data rate.

The trellis plots in Figures 3 and 4 show relative bias for the $ICC = .50$ simulation with 15 and 30 clusters, respectively. The major trends from the $ICC = .10$ simulations were also evident here, and hence we do not reiterate these results. The primary difference in Figures 3 and 4 is that the joint model slope estimates were negatively rather than positively biased. We suspect that the prior distribution (an inverse Wishart with identity scale matrix) is

responsible for this change, perhaps because its mass better approximates that of the likelihood function. As described before, the bias decreased to reasonable levels with group sizes of 30 or larger.

Mean-squared error ratios provide additional information about the relative performance of the two imputation approaches. Recall that mean squared error captures the overall accuracy of an estimate, with ratios greater than unity favoring fully conditional specification (i.e., joint model estimates were further from their true values, on average). For brevity, we report average ratios across the fixed regression coefficients and the level-2 covariance matrix

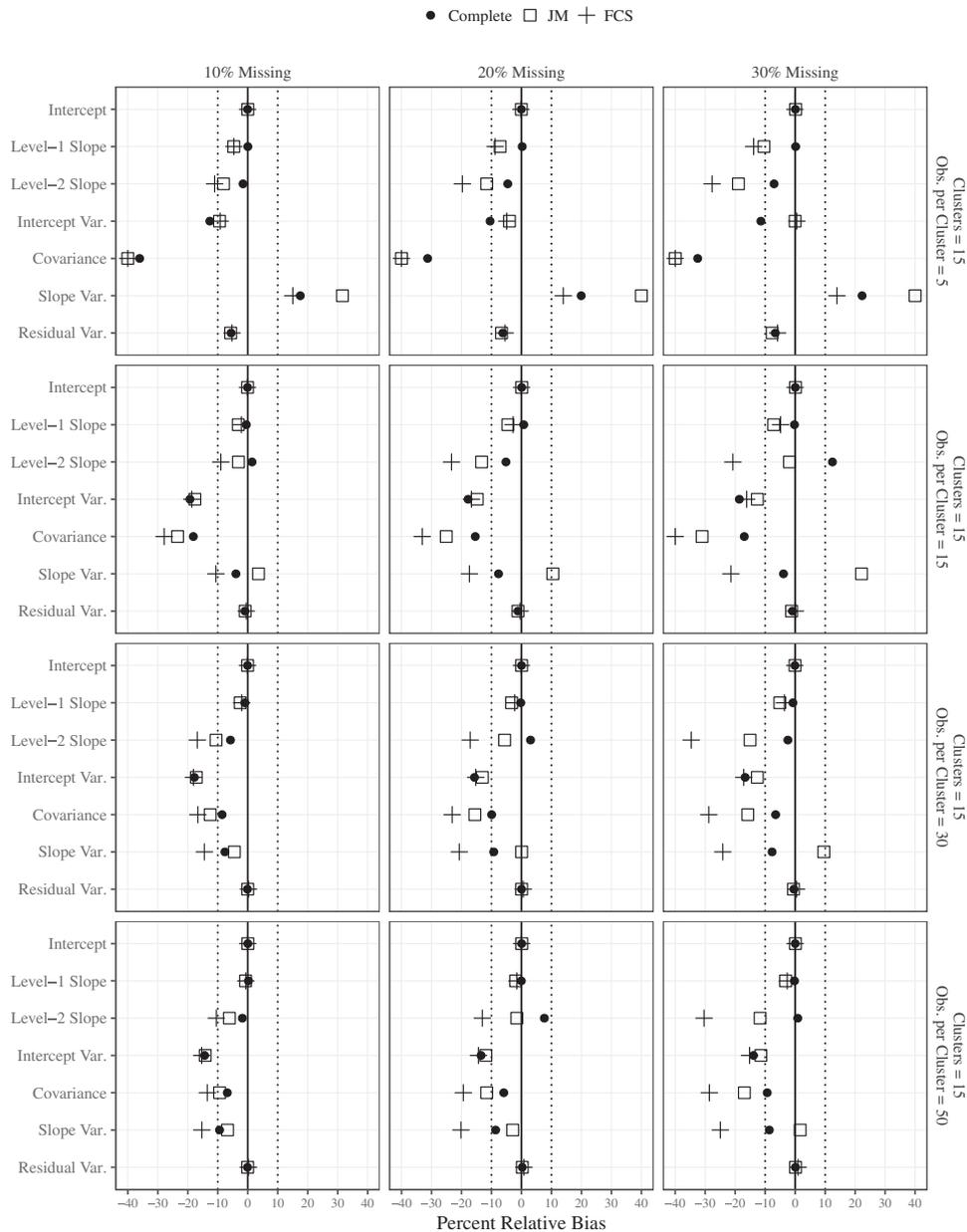


Figure 4. Average relative bias values for design cells with ICC = .50 and 30 clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ± 10 .

elements and do not consider ratios for individual parameters. However, the online supplemental material includes trellis plots of the mean squared error ratios for all conditions. Mean-squared error ratios for the regression coefficients favored fully conditional specification, albeit slightly, with average ratio values of approximately 1.02. The ratios for the level-2 covariance matrix differed by intraclass correlation. Fully conditional specification was superior in the ICC=.10 conditions, with average ratio values of approximately 1.10, whereas the joint model was superior in the ICC=.50 conditions, with average ratios of about .96.

No additional figures are needed to convey the confidence interval coverage results. Collapsing across other design cells, the average coverage value for the joint model regression slopes was .94 (min=.88, max=.99), and the corresponding average for fully conditional specification was .93 (min=.89, max=.99). To further convey the results, we examined the percentage of design cells with coverage values between .925 and .975, an interval defining the so-called liberal criterion from Bradley (1978). These results varied by intraclass correlation but consistently favored joint model imputation. Specifically, in the ICC=.10 simulations, approximately 81.9% of the

joint model coverage values fell in Bradley's interval, as compared to 75% for fully conditional specification. In the ICC = .50 conditions, the corresponding values were 75 and 59.7%.

Missingness owing to auxiliary variables and correlated random effects

Next, we consider the results from the simulation conditions where the residual correlation between the random intercepts and the slopes equaled .50. As described previously, a nonzero covariance should be detrimental to the joint model because imputation uses fewer parameters than the data-generating process. To illustrate these results, the trellis plot in Figure 4 shows relative bias values from the ICC = .50 condition with 30 clusters. As expected, the joint model underestimated the level-2 covariance although the magnitude of this bias was not as large as one might expect. In particular, bias was generally in a tolerable range (e.g., <10% in absolute value) if the missing data rate was 20% or less. Consistent with the previous results, joint model imputation improved as the within-cluster sample size increased. Mean-squared error ratios were largely similar to those from the simulation with uncorrelated random effects. For example, ratios for the level-2 covariance matrix elements again differed by intraclass correlation, with average ratio values of about 1.06 and .93 in the ICC = .10 and .50 conditions, respectively.

Missingness owing to the outcome variable

To explore whether our simulation results can generalize to other MAR processes, we repeated the previous simulations, this time treating y as the cause of missingness on x . The level-2 auxiliary a_2 again determined missingness for w , and the imputation routines always included both auxiliary variables. Manipulating the cause of missingness produced no meaningful changes to the results. For completeness, the online supplemental material includes a full set of trellis plots showing relative bias, but no additional plots are needed to convey these results here. Similarly, the mean-squared error ratios and confidence interval coverage results were virtually identical to those from the previous simulation, and hence no further discussion of these results is warranted.

Real data example

In this section, we use the R package *jomo* (Quartagno & Carpenter, 2016a) and the *Blimp*

application (Enders et al., 2017) to illustrate joint model imputation and fully conditional specification, respectively. The *Blimp* application for macOS, Windows, and Linux is a free download available at www.appliedmissingdata.com/multilevel-imputation.

The data set for the analysis mimics a cluster-randomized design from an educational study of math achievement where 50 schools with 25 students each are randomly assigned to a novel math curriculum condition (the intervention) or a standard curriculum condition (the control) (Montague, Krawec, Enders, & Dietz, 2014). The analysis model examines the influence of the intervention dummy code (0 = control, 1 = intervention), controlling for math self-efficacy and teacher's experience in years.

$$\begin{aligned} \text{math}_{ij} = & \beta_0 + \beta_1(\text{mathse}_{ij}) \\ & + \beta_2(\text{teach exp}_j) + \beta_3(\text{condition}_j) \\ & + b_{0j} + b_{1j}(\text{mathse}_{ij}) + e_{ij} \end{aligned} \quad (22)$$

The intervention dummy code is complete but all other analysis variables have missing values. For the purposes of the analysis example, we treat math self-efficacy (a 6-point rating scale) as continuous, but *jomo* and *Blimp* have facilities for imputing categorical variables (Carpenter & Kenward, 2013; Enders et al., 2017).

The online supplemental material gives the *jomo* and *Blimp* syntax files and R code for the analysis and pooling steps. The *Blimp* commands can also be implemented via pull-down menus from the graphical user interface (macOS and Windows only). In the *jomo* package, random level-1 covariance matrices are invoked by specifying the imputation method as "random," whereas *Blimp* denotes a random association by joining two variables with a colon (e.g., "math:mathse") on the MODEL line. We refer readers to the software documentation files for additional details on syntax conventions (Enders et al., 2017; Quartagno & Carpenter, 2016a). Table 2 lists the estimates from the two approaches. Consistent with the simulation results, the imputation methods produced similar fixed effect estimates but different variance components. In particular, fully conditional specification produced level-2 covariance matrix elements that were uniformly smaller than those of joint model imputation. Pooled likelihood ratio tests (Meng & Rubin, 1992) obtained from the *mitml* package (Grund, Robitzsch, & Lüdtke, 2016) indicated that the slope variance (and covariance) were significant although the joint model test statistic was somewhat larger in value at $F(2,1176.28) = 13.60$, $p < .001$ versus $F(2,567.04) = 6.51$, $p = .002$. Of course, it is difficult to form judgments from a single sample, but the results

Table 2. Real-data analysis results.

<i>Fully conditional specification</i>					
Effect	Est.	SE	<i>t</i>	df	<i>p</i>
Intercept	34.08	2.30	14.79	6324.43	0.00
Selfefficacy	1.86	0.34	5.49	751.07	0.00
Teacher exp.	0.58	0.14	4.11	8483.01	0.00
Condition	3.00	1.33	2.25	3188.41	0.02
Intercept var.	29.05				
Covariance	-5.75				
Slope var.	2.40				
Residual var.	89.92				
<i>Joint model with random covariance matrices</i>					
Effect	Est.	SE	<i>t</i>	df	<i>p</i> Value
Effect	33.97	2.38	14.26	1924.85	.00
Selfefficacy	1.86	.37	5.01	1818.39	.00
Teacher exp.	.60	.14	4.30	5138.86	.00
Condition	2.78	1.30	2.15	11,586.23	.03
Intercept var.	40.29				
Covariance	-9.74				
Slope var.	3.74				
Residual var.	89.00				

of the analysis highlight that the two imputation approaches can produce divergent estimates.

Discussion

The procedures for handling multilevel missing data have received less attention in the literature relative to single-level analysis problems. Existing research has largely focused on random intercept models, and the literature addressing random slopes is particularly scant. To date, the few studies to examine missing data handling for these models have focused on the fully conditional specification framework and “reverse random coefficient” imputation (Enders et al., 2017; Grund et al., 2016). Although it has not received much attention in the literature, a joint modeling strategy that uses random within-cluster covariance matrices to preserve cluster-specific associations (Yucel, 2011) is a promising alternative for random coefficient analyses. This approach was recently implemented in the R package *jomo* (Quartagno & Carpenter, 2016a), and the only study to examine its performance did so in the context of individual patient data meta-analysis (Quartagno & Carpenter, 2016b).

The primary goal of our manuscript was to compare fully conditional specification to joint model imputation with random level-1 covariance matrices. A vital consideration with multiple imputation is whether (and how) the imputation model preserves features of the analysis, a notion described as congeniality (Meng, 1994; Schafer, 2003) or compatibility (Bartlett et al., 2014; Carpenter & Kenward, 2013). Simulation results suggest that fully conditional specification can produce negatively biased slope variance estimates (Enders et al., 2017; Grund et al., 2016), and

we show that an incorrect distributional assumption is responsible for this issue; the reverse regression model’s random effects cannot be normally distributed if the analysis model’s random effects are also normal (a standard assumption in multilevel modeling applications). In a similar vein, we show that joint model imputation with random covariance matrices is potentially problematic because it fails to preserve covariation between the intercepts and the slopes.

Our simulation results convey a number of take-home messages to substantive researchers. First, both imputation procedures generally produced tolerable biases when the missing data rate was 10% or less and the sample was composed of at least 30 clusters with 15 observations per group. These favorable conditions may encompass a wide range of multilevel modeling applications. Second, in terms of overall accuracy, fully conditional specification appears to be superior when the intraclass correlation is low (e.g., $ICC = .10$), whereas joint model imputation is somewhat better when the intraclass correlation is high (e.g., $ICC = .50$). Roughly speaking, this suggests that fully conditional specification is preferable for cross-sectional data, and the joint model is better for within-participants designs. Third, the missing data pattern is an important determinant of imputation quality. Problems with fully conditional specification are generally restricted to incomplete predictors with random slopes, and the incompatibilities that we described earlier in the article would not occur with incomplete outcomes or fixed predictors. On the other hand, there is no reason to expect joint model imputation to improve if y is missing and x is complete.

Perhaps the most important take-home message from our study is that neither fully conditional specification nor joint model imputation are optimal for random slope analyses with incomplete predictors. Fully Bayesian (i.e., model based) imputation (Erlor et al., 2016; Ibrahim, Chen, & Lipsitz, 2002; Zhang & Wang, 2017) and its close relative substantive model-compatible imputation (Bartlett et al., 2014; Enders, Du, & Keller, 2018; Goldstein, Carpenter, & Browne, 2014) are promising alternatives that are starting to receive attention in the literature. Erlor et al. (2016) and Grund, Lüdtke, and Robitzsch (2018) illustrate model-based Bayesian imputation with the JAGS software package (Plummer, 2016), and a substantive model-compatible variant of fully conditional specification was recently implemented in the *Blimp* application (Enders et al., 2018; Keller & Enders, 2018). Both procedures avoid bias by respecifying the conditional distribution of the incomplete covariates (e.g., the

problematic reverse random coefficient model from Equation 11) into a composite function, the components of which are compatible with the analysis model. Limited simulation results suggest that these procedures can provide a rather dramatic improvement over conventional imputation routines when the analysis model includes random coefficients and other types of interactive effects (Enders et al., 2018; Grund et al., 2018). For interested readers, the online supplemental material gives the Blimp code for applying substantive model-compatible imputation to a random slope analysis, and this material also gives plots showing relative bias values from a subset of our simulation conditions. Additional analysis examples (e.g., for models with interactive terms) are available in the Blimp user guide (Keller & Enders, 2018).

Though our study addresses an important gap in the methodological literature, our simulations lack generalizability because we necessarily investigated a limited set of conditions. Future studies could investigate different combinations of level-1 and level-2 sample sizes, different effect sizes, and more complex models. The interaction of model complexity and sample size could be particularly important for understanding joint model imputation, given the influence of within-cluster sample size on imputation quality. Also, it could be important to investigate the impact of unequal group sizes on fully conditional specification as the imputation model for level-2 variables (Equation 12) does not directly adjust for the fact that cluster means are potentially derived from different numbers of level-1 units. Analytic and simulation results suggest that manifest cluster means provide good performance in most situations but can introduce slight biases with low intraclass correlations, few observations per cluster, and extremely unbalanced cluster sizes (Grund et al., 2017). A future release of Blimp will feature the option to invoke the latent cluster mean procedure described by Grund et al. (2017). Next, we limited this preliminary comparison to normally distributed variables. Nonnormal continuous variables are potentially problematic for multiple imputation (Yuan, Yang-Wallentin, & Bentler, 2012; Yucel & Demirtas, 2010), and hence future studies should examine whether the two imputation frameworks are differentially impacted by this common problem. Additionally, both approaches readily accommodate incomplete categorical variables; Blimp uses a probit regression framework (i.e., latent variable imputation) for ordinal or nominal variables, and jomo offers the same functionality for nominal variables (imputation for nominal variables can also

accommodate ordered categories, albeit it with a richly parameterized model). Although both software programs use the same underlying latent variable definition of categorical variables (Albert & Chib, 1993; Carpenter & Kenward, 2013; Johnson & Albert, 1999), implementing this approach in the joint model framework could greatly accentuate model complexity-related issues, particularly with incomplete level-1 variables and random slopes. Finally, it is important to emphasize that we considered a rather conventional multilevel regression model. Further studies could examine the use of multiple imputation with other covariance structures or analytic frameworks (e.g., random intercept models, multilevel structural equation models, generalized estimating equations).

Conclusion

In sum, relatively few studies have examined missing data handling for random coefficient models, and ours was the first to compare fully conditional specification to joint model imputation with random covariance matrices (Yucel, 2011). Although both approaches can introduce bias-inducing incompatibilities, they tend to do so in somewhat different situations. However, the potential for problematic biases underscores the importance of promising new Bayesian imputation procedures (Enders et al., 2016, 2017, October).

Article Information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant R305D150056 from the Institute of Educational Sciences.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management,

analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the Institute of Educational Sciences is not intended and should not be inferred.

References

- Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. doi:10.2307/2290350
- Asparouhov, T., & Muthén, B. (2010). *Multiple imputation with Mplus*. Retrieved from <http://www.statmodel.com/download/Imputations7.pdf>.
- Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods and Research*, *24*(4), 462–487. doi:10.1177/0962280214521348.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x
- Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*(3), 391–420. doi:doi.org/10.1007/s001800000041
- Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*(5), 1–14. doi:10.18637/jss.v045.i05
- Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. West Sussex, UK: Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. doi:dx.doi.org/10.1037/1082-989X.6.4.330
- DiazOrdaz, K., Kenward, M. G., Gomes, M., & Grieve, R. (2016). Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Statistics in Medicine*, *35*(20), 3482–3496. doi:10.1002/sim.6935
- Drechsler, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69–95. doi:10.3102/1076998614563393
- Enders, C. K. (2013). Centering predictors and contextual effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 89–107). Los Angeles: Sage.
- Enders, C. K., Du, H., & Keller, B. T. (2017, October). A fully Bayesian imputation procedure for random coefficient models (and other pesky product terms). Paper presented at the Society for Multivariate Experimental Psychology, Minneapolis, MN.
- Enders, C. K., Du, H., & Keller, B. T. (2018). Substantive model-compatible imputation for multilevel regression models with random coefficients, interaction effects, and other nonlinear terms. *Manuscript submitted for publication*.
- Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, *Advance online publication*, *23*(2), 298–317. doi:dx.doi.org/10.1037/met0000148
- 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*(2), 222–240. doi:dx.doi.org/10.1037/met0000063
- Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974. doi:10.1002/sim.6944
- Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87–107. doi:10.1080/10705519709540063
- Gelman, A., & Raghunathan, T. E. (2001). Comment on conditionally specified distributions: An introduction. *Statistical Science*, *16*, 268–269. doi:10.1214/ss/1009213728
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi:www.jstor.org/stable/2246093
- Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252–286. doi:10.3102/1076998606298042
- Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(2), 553–564.
- Goldstein, H., Carpenter, J. R., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling: An International Journal*, *9*(3), 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*(3), 206–213. doi:dx.doi.org/10.1007/s11121-007-0070-9
- 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*(2), 640–649. doi:10.3758/s13428-015-0590-3
- Grund, S., Lüdtke, O., & Robitzsch, A. (2017). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs.

- Journal of Educational and Behavioral Statistics*, 43(3), 316–353.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, 21(1), 111–149. doi:10.1177/1094428117703686
- Grund, S., Robitzsch, A., & Lüdtke, O. (2016). Package 'mitml'. Retrieved from cran.r-project.org/web/packages/mitml/website.
- Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology*, 149(9), 876–883. doi:academic.oup.com/aje/article-abstract/149/9/876/113681
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. doi:doi.org/10.3102/0162373707299706
- Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, 30(1), 55–78.
- Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.
- Kaplan, D. (1988). The impact of specification error on the estimation, testing and improvement of structural equation models. *Multivariate Behavioral Research*, 23(1), 69–86. doi:10.1207/s15327906mbr2301_4
- Keller, B. T., & Enders, C. K. (2018). *Blimp user's guide (Version 1.1)*. Los Angeles, CA.
- Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.
- Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30(1), 1–21. doi:10.1207/s15327906mbr3001_1
- Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(4), 649–662. doi:dx.doi.org/10.1080/10705511.2011.607721
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.
- Longford, N. (1989). Contextual effects and group means. *Multilevel Modelling Newsletter*, 1(3), 5.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, 16(4), 444–467. doi:10.1037/a0024376
- Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1), 141–165. doi:dx.doi.org/10.1037/met0000096
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86–92. doi:10.1027/1614-1881.1.3.86
- McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28(2), 295–314. doi:10.1007/s10648-014-9287-x
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, 10(3), 259–284. doi:10.1037/1082-989X.10.3.259
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.
- Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1), 103–111. doi:www.jstor.org/stable/2337151
- Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, 42(4), 432–466. doi:10.3102/1076998617690869
- Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle-school students of varying ability. *Journal of Educational Psychology*, 106(2), 469–481. doi:10.1037/a0035176
- Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, 27(1), 79–103. doi:10.1177/0193841X02239019
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Plummer, M. (2016). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling (Version 4.2.0). Retrieved from <http://sourceforge.net/projects/mcmc-jags/>.
- Quartagno, M., & Carpenter, J. R. (2016a). Package 'jomo'. Retrieved from cran.r-project.org/web/packages/jomo/website.
- Quartagno, M., & Carpenter, J. R. (2016b). Multiple imputation for IPD meta-analysis: Allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17), 2938–2954. doi:10.1002/sim.6837
- Rabe-Hesketh, S., Skondral, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 512–531). New York: Guilford Press.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2013). HLM 7.01 for Windows. Skokie, IL.
- Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology*, 32, 143–150. doi:www.statcan.gc.ca/pub/12-001-x/2006002/article/9548-eng.pdf
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. doi:doi.org/10.1093/biomet/63.3.581
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall.
- Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355–377). Washington, DC: American Psychological Association.
- Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models

- differ. *Statistica Neerlandica*, 57(1), 19–35. doi:10.1111/1467-9574.00218
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 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(2), 437–457. doi:10.1198/106186002760180608
- Shin, Y., & Raudenbush, S. W. (2007). Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics*, 63(4), 1262–1268. doi:10.1111/j.1541-0420.2007.00818.x
- 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(1), 26–53. doi:10.3102/1076998609345252
- Shin, Y., & Raudenbush, S. W. (2013). Efficient analysis of Q-level nested hierarchical general linear models given ignorable missing data. *International Journal of Biostatistics*, 9, 109–133. doi:10.1515/ijb-2012-0048
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.
- Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). Optimal Design Plus Empirical Evidence: Documentation for the “Optimal Design” Software Version 3.0.
- Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 521–562). Charlotte, NC: Information Age.
- Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, 50(3), 329–345. doi:10.1002/bimj.200710423
- van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York: Routledge.
- van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.
- 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(12), 1049–1064. doi:doi.org/10.1080/10629360600810434
- van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2014). Package ‘mice’. Retrieved from cran.r-project.org/web/packages/mice/mice.pdf.
- Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distributional conditions. *Sociological Methods & Research*, 41(4), 598–629. doi:10.1177/0049124112460373
- Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling: An International Journal*, 11(4), 351–370. doi:10.1177/1471082X1001100404
- Yucel, R. M., & Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics and Data Analysis*, 54(3), 790–801. doi:10.1016/j.csda.2009.01.016
- Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, 22(4), 649–666.