Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

An Investigation of Factored Regression Missing Data Methods for Multilevel Models with Cross-Level Interactions

Brian T. Keller & Craig K. Enders

To cite this article: Brian T. Keller & Craig K. Enders (2023): An Investigation of Factored Regression Missing Data Methods for Multilevel Models with Cross-Level Interactions, Multivariate Behavioral Research, DOI: 10.1080/00273171.2022.2147049

To link to this article: https://doi.org/10.1080/00273171.2022.2147049

Published online: 05 Jan 2023.

Submit your article to this journal

Article views: 90

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2022.2147049

Routledge
Taylor & Francis Group

Check for updates

An Investigation of Factored Regression Missing Data Methods for Multilevel Models with Cross-Level Interactions

Brian T. Keller[a] and Craig K. Enders[b]

[a]The University of Texas at Austin; [b]The University of California, Los Angeles

ABSTRACT

A growing body of literature has focused on missing data methods that factorize the joint distribution into a part representing the analysis model of interest and a part representing the distributions of the incomplete predictors. Relatively little is known about the utility of this method for multilevel models with interactive effects. This study presents a series of Monte Carlo computer simulations that investigates Bayesian and multiple imputation strategies based on factored regressions. When the model's distributional assumptions are satisfied, these methods generally produce nearly unbiased estimates and good coverage, with few exceptions. Severe misspecifications that arise from substantially non-normal distributions can introduce biased estimates and poor coverage. Follow-up simulations suggest that a Yeo–Johnson transformation can mitigate these biases. A real data example illustrates the methodology, and the paper suggests several avenues for future research.

KEYWORDS

Missing data; multiple imputation; factored regression; multilevel modeling; moderation

Over the last few decades, a large body of research has accumulated supporting the use of multiple imputation as an appropriate method to handle missing data. Despite being introduced over thirty years ago, methodologists have only investigated and fine-tuned multiple imputation's ability to handle interactions and other nonlinear terms in the last decade. An early recommendation was to treat incomplete product terms as if they were "just another variable" to be imputed using standard methods for normally distributed data (Allison, 2002; von Hippel, 2009). However, subsequent studies have shown that this approach requires the same missing completely at random assumption as older deletion methods, and it will introduce bias under the less restrictive missing at random assumption required by modern missing data estimators (Bartlett et al., 2015; Enders et al., 2014; Lüdtke et al., 2020b; Seaman et al., 2012; Zhang & Wang, 2017).

More recently, advances have been made in generating imputations for nonlinear effects by factorizing the joint distribution into a part representing the analysis model of interest and a part representing the distributions of the incomplete predictors. Various names have referred to this approach in the literature (e.g., substantive-compatible imputation, sequential specification, model-based imputation), but we will refer to it simply as factored regression imputation. While the idea of using factored regression models dates back to early literature (Ibrahim, 1990; Ibrahim et al., 2002; Lipsitz & Ibrahim, 1996), it wasn't until more recently that this method was applied to incomplete interactive effects (Bartlett et al., 2015; Du et al., 2022; Enders et al., 2020; Erler et al., 2016; Goldstein et al., 2014; Lüdtke et al., 2020b; Zhang & Wang, 2017) and subsequently became accessible to researchers via software packages (Bartlett et al., 2021; Keller & Enders, 2021; Quartagno & Carpenter, 2020; Robitzsch & Lüdke, 2021). Virtually across the board, the literature shows that factored regression methods produce approximately unbiased estimates while requiring less strict assumptions than the earlier "just another variable" approach. This study attempts to extend our knowledge about these important models.

Relative to other topics in the missing data literature, much less is known about the behavior of the factored regression approach to fitting multilevel models with interactive effects. In the biostatistics literature, Erler et al. (2016) investigated cross-level interactions with a small simulation that didn't vary important features like sample size and intraclass correlation. Several studies considered factored-regression specification for random coefficient models (Enders et al., 2020; Enders et al., 2018; Erler et al., 2019;

CONTACT Brian T. Keller bk@utexas.edu The University of Texas at Austin, Austin, TX, USA.
© 2022 Taylor & Francis Group, LLC

Grund et al., 2018), and the most comprehensive study of interactive effects to date is a recent paper by Grund et al. (2021). These authors investigated multilevel multiple imputation strategies for moderated regression models where the interacting predictors were linearly or non-linearly related to one another. Following the main findings from the single-level regression literature (e.g., Bartlett et al., 2015; Lüdtke et al., 2020b; Zhang & Wang, 2017), their simulation studies showed that, when correctly specified, multiple imputation models based on factored regressions hold great promise for multilevel models with incomplete interactive effects.

Despite a growing body of literature supporting their use, our current knowledge about factored regression models is deficient on multiple fronts. Existing studies have primarily focused on ideal conditions where the multilevel model's distributional assumptions are strictly satisfied. For example, no research to date has considered the impact of skewed and kurtotic data on factored regression's multilevel parameter recovery. Two studies have investigated this issue in the context of single-level moderated regression models (Lüdtke et al., 2020a, 2020b), but distributional violations are considerably more complex with multilevel models, which could have non-normally distributed residuals in either the within- or between-cluster part of the model. Our study is the first to investigate this issue, and we use simulations to examine the possibility of modeling skewed predictors with a Yeo–Johnson transformation (Yeo & Johnson, 2000), a potential solution for non-normal data proposed in the single-level imputation literature (Lüdtke et al., 2020b).

A second major limitation of the existing literature is that it focuses almost entirely on multilevel multiple imputation, almost at the exclusion of Bayesian inference. In the complete-data literature there has been a growing interest in Bayesian multilevel models, especially with applications involving smaller sample size conditions (Asparouhov & Muthén, 2021a; Liu et al., 2016; McNeish, 2016a, 2016b; McNeish & Stapleton, 2016). From a practical perspective, Bayesian inference is convenient because the researcher simply fits a multilevel model to the observed data and summarizes the resulting posterior distributions without having to perform the extra steps of reanalyzing the data and pooling estimates and standard errors. However, Bayesian estimation with incomplete data is quite different from its complete-data counterpart because incomplete predictor variables require additional regression models with their own distributional assumptions and requisite priors. As such, evaluating Bayesian inference and comparing it to the newer "gold standard" multiple imputation models is a major thrust and contribution of this paper. An important issue here is the choice of prior distribution for the level-2 covariance matrix, which can substantially impact estimates and inference. We investigate several off-the-shelf Wishart prior distributions, and we describe a new separation strategy that introduces distinct prior distributions on each level-2 variance component and the random effect correlations. The parameterization we outline here is similar to other separation strategies proposed in the literature (Barnard et al., 2000; Merkle & Rosseel, 2018), and a comparable method has shown promise in simulation studies with single-level latent curve models (Liu et al., 2016). A major contribution of this paper is to describe and evaluate this new separation prior for multilevel analyses.

The structure of this paper is as follows. We begin by discussing factored regression specifications and their application to a cross-level interaction with random slopes. Next, we provide a conceptual overview of constructing the Markov chain Monte Carlo (MCMC) sampler for drawing parameters and generating imputations. Next, we describe five different Monte Carlo simulations that investigate factored regression specifications. Finally, we include a substantive example applying this method to real data.

Conventional imputation schemes

Historically, multiple imputation was developed to generate complete data sets for large-scale survey data (Rubin, 1987; Rubin, 2004). In part due to computational restrictions of its era, the goal was often to provide data sets that could then be used for a wide range of secondary analyses across multiple research teams. We will refer to this traditional approach to multiple imputation as *agnostic* imputation, as the imputations need to be flexible enough to support a range of different analyses, but they are not generated to be mathematically consistent with one particular analysis model.

Generally, the agnostic imputation approach generates samples from the posterior predictive density of the unobserved variables conditional on the observed variables. Symbolically, this idea is

$$f\left(\mathbf{V}_{(\mathrm{M})}|\mathbf{V}_{(\mathrm{O})}\right) \tag{1}$$

where $f(\cdot)$ represents a generic density (e.g., a normal curve), $\mathbf{V}_{(\mathrm{M})}$ represents the set of variables with

incomplete observations for a particular individual, $\mathbf{V}_{(O)}$ represents the set of fully observed variables, and $\mathbf{V}$ is the set of all variables. Note that this representation makes no distinction between analysis variables, predictor variables, and auxiliary variables. For example, if an outcome measure is incomplete, it would be part of the set; conversely, if the outcome is complete, it would be in the set.

As a heuristic, one can think of Equation 1 as analogous to a regression problem involving the prediction of the missing variables given the fully observed variables, with imputations sampled from a distribution based upon the regression model parameters. Importantly, the imputation regression model need not be the same as the analysis model, and it usually isn't under an agnostic scheme; this is an important point of contrast with factored regression specifications, which tailor missing data handling around a particular analysis. Joint model imputation (Asparouhov & Muthén, 2010b; Carpenter et al., 2011; Goldstein et al., 2009; Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008, 2011) and fully conditional specification (Enders et al., 2018; van Buuren, 2011) are the two major frameworks for implementing Equation 1 with multilevel data. The former uses a mean vector and covariance matrix as the basis for imputation, whereas the latter uses round robin regression equations where each incomplete variable is regressed on all other variables. Enders et al. (2016) provide an overview of these methods.

Factored regression specifications for two-level models

As mentioned previously, the factored regression specification factorizes a joint distribution into a part representing the analysis model of interest and a part representing the distributions of the incomplete predictors. To illustrate, consider an analysis involving an outcome variable Y and two predictors, X and Z. The trivariate joint distribution factors into the product of multiple distributions, each corresponding to a regression model. There are two predominant factorizations suggested in the literature. The first factorizes the joint distribution into the product of a univariate distribution for the analysis model and a joint distribution for the predictors (Enders et al., 2020; Keller & Enders, 2021). Applied to the trivariate example, the factorization is

$$f(Y,X,Z) = f(Y|X,Z) \times f(X,Z) \tag{2}$$

where the leftmost term is the multivariate distribution of all three variables, the first term to the right of the equals sign is a density that corresponds to the analysis model, and the rightmost term is the joint distribution of the predictors (e.g., a bivariate normal distribution). Continuing to factorize the predictor distribution leads to the so-called sequential specification that instead uses a set of univariate distributions to represent the distributions of the incomplete predictors (Erler et al., 2019; Erler et al., 2016; Lüdtke et al., 2020b).

$$f(Y,X,Z) = f(Y|X,Z) \times f(X|Z) \times f(Z) \tag{3}$$

The two specifications are equivalent in some situations but different in others.

To illustrate factorized specifications more concretely, we focus on their application to a two-level hierarchical linear model with a cross-level interaction (i.e., the product of a level-1 and level-2 predictor). The analysis model is given as follows

$$\begin{gathered} y_{ij} = (\beta_0 + u_{0_j}) + (\beta_1 + u_{1_j})x_{ij} + \beta_2 z_j + \beta_3 x_{ij} z_j + e_{ij} = \hat{y}_{ij} + e_{ij} \\ \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} & \sigma^2_{u_1} \end{bmatrix} \right) \\ e_{ij} \sim N\left(0, \sigma^2_e\right) \end{gathered} \tag{4}$$

where y_{ij} is the dependent variable with observation i nested in cluster j, x_{ij} is the level-1 predictor value for that observation, and z_j is the level-2 moderator score for cluster j. The u_{0j} and u_{1j} terms are level-2 random effects that are multivariate normal with means of zero and an unstructured covariance matrix, $\boldsymbol{\Sigma}_u$. For our discussion, we assume that the observed data satisfy the missing at random mechanism and the parameters associated with the cause of missingness are distinct from the analysis model parameters (i.e., ignorability; Rubin, 1976). However, this assumption is unnecessary as the factorization scheme readily extends to nonignorable mechanisms (Du et al., 2022).

As discussed previously, the conditional density $f(Y|X,Z)$ in Equations 2 and 3 corresponds to the focal analysis model and the corresponding distributional assumptions for the outcome (i.e., Y is conditionally normal given the predictors and random effects). More concretely, the density is given by the conditional normal distribution implied by the cross-level analysis model from Equation 4

$$f(Y|X,Z) = \frac{1}{\sqrt{2\pi\sigma^2_e}} \exp\left[-\frac{1}{2}\left(\frac{y_{ij} - \hat{y}_{ij}}{\sigma^2_e} \right)^2 \right] = N_1\left(\hat{y}_{ij}, \sigma^2_e\right) \tag{5}$$

where $\hat{y}_{ij}$ is the predicted value from Equation 4, and N_1 denotes a normal distribution function with a

conditional mean and variance as its arguments. The outcome variable's distribution is common to Equations 2 and 3, but the two factorizations apply different models to the predictors.

Joint model specification for predictors

Returning to Equation 2, the $f(X, Z)$ term corresponds to a within- and between-level multivariate normal distribution for the predictors (Enders et al., 2020; Keller & Du, 2019). The within-cluster model for X describes the distribution of the level-1 scores around their level-2 *latent* group means, and the between-cluster model for X and Z is a bivariate normal distribution for the means and level-2 scores. These densities are

$$\begin{aligned} x_{ij} &\sim N_1\left(\mu_{Xj}, \sigma^2_{wX}\right) \\ \begin{bmatrix} \mu_{Xj} \\ z_j \end{bmatrix} &\sim N_2\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma^2_{bX} & \sigma_{bXZ} \\ \sigma_{bZX} & \sigma^2_{bZ} \end{bmatrix}\right) \end{aligned} \tag{6}$$

where μ_{Xj} is the latent group mean of X for cluster j, σ^2_{wX} is the within-cluster variance, $\boldsymbol{\mu}$ is a vector of grand means, and $\boldsymbol{\Sigma}_b$ is an unstructured between-cluster covariance matrix that defines the relationship between the latent cluster means and the level-2 predictor. Importantly, the expressions in Equation 6 are not posterior predictive distributions that generates imputations, as those distributions also rely on the model-implied distribution of Y.

The joint model parameterization for predictor variables has several advantages, chief among them is computational efficiency because the multivariate distribution can be recast as a fully conditional specification problem where each predictor is regressed on all others (Bartlett et al., 2015; Enders et al., 2020; Keller & Enders, 2021). Second, in the context of Bayesian estimation and posterior inference, the joint model specification facilitates both grand mean centering and latent group mean centering—a vital step in most multilevel analyses—because the Markov chain Monte Carlo (MCMC) algorithm yields estimates of these quantities at every iteration (Enders & Keller, 2019). Finally, this model readily accommodates binary, ordinal, and nominal predictors using a latent response variable (probit regression) formulation. The primary downside of this specification is that it does not allow nonlinear relations among predictors or non-normal continuous distributions.

Sequential specification for predictors

Alternatively, we can also model the density $f(X, Z)$ by using the chain rule to factor the joint distribution as a sequence of univariate conditional models, as in Equation 3. Applied to the cross-level interaction model, the conditional distributions are both normal, as follows.

$$\begin{aligned} f(X|Z) &= N_1\left(\gamma_0 + \gamma_1 z_j + g_{0j}, \sigma^2_{wX}\right) \\ f(Z) &= N_1\left(\mu_Z, \sigma^2_{bZ}\right) \end{aligned} \tag{7}$$

Considering the conditional distribution of the level-1 predictor, $\gamma_0 + \gamma_1 z_j + g_{0j}$ is the predicted value from a random intercept regression model X regressed on Z, and σ^2_{wX} is the within-cluster residual variance. The $f(Z)$ term translates into an intercept only regression model with a mean and between-cluster variance, μ_Z and σ^2_{bZ}, respectively. It is important to reiterate that these expressions are not the posterior predictive distributions that generate imputations, as those distributions also condition on the model-implied distribution of Y.

The main advantages of the sequential approach is that it can offer more flexibility in modeling the association between X and Z (Grund et al., 2021b; Lüdtke et al., 2020b). For example, a researcher might posit that a nonlinear relationship between X and Z exists and would like to model this explicitly. By using the sequential approach, we could easily model said nonlinearity by adding a polynomial term to the random intercept model that defines the $f(X|Z)$ density (Lüdtke et al., 2020b). Additionally, the sequential specification allows non-normal continuous variables to be modeled using a Yeo–Johnson transformation (Yeo & Johnson, 2000), which is an extension of the popular Box–Cox transformation. In context of Bayesian estimation and posterior inference, the sequential specification may be more limiting because the group and grand means needed for centering are not always explicit model parameters (a feature that is not necessary for conducting multiple imputation).

Categorical predictors

Either factored regression specification allow for categorical predictor variables. We describe the procedure for a binary covariate using the latent response formulation (Agresti, 2012; Albert & Chib, 1993; Johnson & Albert, 1999), and this framework also accommodates ordinal and multicategorical nominal predictors (Enders et al., 2018; Goldstein et al., 2009). The binary probit model imagines discrete responses as originating from an underlying normally

distributed latent variable with a single fixed threshold, $\tau = 0$, dividing the latent variable distribution into the discrete observations; participants with latent scores below the threshold have a discrete value of 0, and cases with latent responses above the threshold have values of 1. To illustrate how the probit model applies to the joint and sequential models, imagine that Z is now a level-2 dummy code with scores of 0 and 1. We denote the underlying latent response variable as Z^*. The link between Z and Z^* is as follows.

$$z_j = \begin{cases} 1 \text{ if } z_j^* > 0 \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

Using a probit model with normally distributed latent scores yields models that maintain the same general structure as those for continuous predictors. For example, when Z is binary, the between-cluster model from Equation 6 is

$$\begin{bmatrix} \mu_{Xj} \\ z_j^* \end{bmatrix} \sim N_2\left(\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma_{bX}^2 & \sigma_{bXZ} \\ \sigma_{bZX} & 1 \end{bmatrix}\right) \tag{9}$$

where μ_{Xj} and z_j^* are both level-2 latent variables, and $\sigma_{bZ}^2 = 1$ in the lower diagonal of $\boldsymbol{\Sigma}_b$ indicates that the variance is fixed for identification, and the fixed threshold identifies the mean structure. Similarly, the sequential specification's between-cluster model from Equation 7 is an empty regression model for Z^* with the variance again fixed at 1 for identification, and the within-cluster model is unchanged.

$$f(Z) = N_1(\mu_{Z^*}, 1) \tag{10}$$

In both parameterizations, the Z dummy code appears on the right side of the focal regression and thus the conditional mean of the normal distribution in Equation 5. Finally, the sequential specification also accommodates logistic covariate models using the Pólya-Gamma specification described by Polson et al. (2013) and Asparouhov and Muthén (2021b), but we focus on the latent response formulation because of its computational efficiency.

Distributions of missing values

In a factored regression specification, the distributions of missing values are a function of every model in which an incomplete variable appears. To replace missing outcome scores, MCMC samples imputations from the normal distribution in Equation 5, conditional on the current parameter values and random effects. In the sequential specification in Equation 3, the level-1 predictor X appears in the conditional mean of Y's normal distribution from Equation 5, and it functions as an outcome in its regression model and density in Equation 7. Symbolically, the conditional distribution of X given all other analysis variables is the product of two normal distributions.

$$\begin{aligned} f(X|Y,Z) &\propto f(Y|X,Z) \times f(X|Z) \\ &= N_1\left(\hat{y}_{ij}, \sigma_e^2\right) \times N_1\left(\gamma_0 + \gamma_1 z_j + g_{0j}, \sigma_{wX}^2\right) \end{aligned} \tag{11}$$

The resulting function is also normal distribution, the mean and variance of which depend on both the analysis and predictor model parameters (Enders et al., 2020, Eq. 20). Importantly, the presence of a product term in the imputation model introduces a heteroscedastic variance parameter, such that the normal distribution's spread depends on the product term, cluster j's random slope residual and its value on the moderator variable (Enders et al., 2020, Eq. 8).

Turning to the level-2 predictor Z, this variable appears in the conditional mean of the Y and X densities, and it is the outcome in the empty level-2 model on bottom line of Equation 7. Symbolically, the conditional distribution of Z given all other analysis variables is the product of three normal distributions

$$\begin{aligned} f(\boldsymbol{\Sigma}_u) &\propto |\boldsymbol{\Sigma}_u|^{-(df_0+V+1)/2} \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{S}_0\boldsymbol{\Sigma}_u{}^{-1}\right)\right) \\ &= N_1(\mu_Z, \sigma_{bZ}^2) \times \textstyle\prod_{i=1}^{n_j} N_1(\hat{y}_{ij}, \sigma_e^2) \times N_1(\gamma_0 + \gamma_1 z_j + g_{0j}, \sigma_{wX}^2) \end{aligned} \tag{12}$$

where the densities for the Y and X models are a product over all observations within a cluster, which are assumed independent after conditioning on the random effects.

Finally, when Z is a binary predictor, the latent responses are missing data that require imputation. For cases with observed Z responses, MCMC samples Z^* values from a truncated normal distribution; when $Z = 0$, the algorithm samples negative imputations for Z^* (i.e., latent scores fall below the fixed threshold), and it otherwise samples positive imputations if $Z = 1$. For cases where Z is missing, the Z^* are drawn from an unconstrained normal distribution. Like its continuous counterpart, the posterior predictive distribution of missing Z^* values is proportional to the triple product in Equation 10, where σ_{bZ}^2 is fixed at 1 to identify the latent response variable's metric. The procedure naturally produces a discrete impute by comparing each Z^* to the fixed threshold.

Prior distributions for the level-2 covariance matrix

As noted previously, the choice of prior specification for the level-2 covariance matrix can impact one's results. The inverse Wishart distribution is a popular choice of conjugate prior, the form of which is

$$f(\boldsymbol{\Sigma}_u) \propto |\boldsymbol{\Sigma}_u|^{-(df_0+V+1)/2} \exp\left(-\frac{1}{2}\text{tr}\left(\mathbf{S}_0\boldsymbol{\Sigma}_u^{-1}\right)\right) \quad (13)$$

where the hyperparameters $\mathbf{S}_0$ and df_0 are a sum of squares and cross-products matrix and degrees of freedom, respectively. Roughly speaking, the hyperparameters encode a prior guess about the random effect covariance matrix, and the degrees of freedom parameter is essentially the number of imaginary level-2 units assigned to that matrix. Assigning $\mathbf{S}_0 = \mathbf{0}$ and $df_0 = V - 1$ gives the multivariate sibling of the Jeffreys prior (Gelman et al., 2014), and other common choices include $\mathbf{S}_0 = \mathbf{I}$ and $df_0 = V$, $\mathbf{S}_0 = \mathbf{I}$ and $df_0 = V + 1$, and an improper prior with $df_0 = -V - 1$ (Asparouhov & Muthén, 2010a; Lunn et al., 2013). The hyperparameters can also be estimated from the data (Casella, 2001; Darnieder, 2011; McNeish, 2016b); however, data driven priors typically do not account for incomplete data. A limitation of the inverse Wishart prior is that it can't disentangle variances and covariances (Gelman, 2006; Grimm et al., 2016; Muthén & Asparouhov, 2012), and this one-size-fits-all characteristic of the prior can introduce bias in small samples (McNeish, 2016a, 2016b).

A major thrust of this paper is to describe and evaluate a separation-type strategy that treats each element of the level-2 covariance matrix as a distinct parameter with its own prior. Extending ideas from Merkle and Rosseel (2018) to the multilevel context, we define the u_{0_j} and u_{1_j} from Equation 4 as level-2 latent variables, and we use a phantom latent variable to induce their correlation. To illustrate the main ideas, Figure 1 shows a pair of path diagrams. The inferential model in panel (A) corresponds to the focal analysis from Equation 4, and the working model in panel (B) reflects a specification that decouples the random effect correlation from the variance components. In panel (B), the level-2 phantom variable, D, can be viewed as a proxy for the correlation, and the additional augmented parameters in the working model—λ_1, λ_2, $\sigma^{2*}_{u_0}$, and $\sigma^{2*}_{u_1}$—are deterministic functions that preserve the random effect correlation.

Essentially, the procedure works as follows. First, MCMC samples random effects that condition the working model parameters, including the phantom variable correlation. In turn, the inferential model's regression coefficients, level-1 residual variance, and level-2 variances all condition on these random effects. Importantly, MCMC treats each level-2 variance component as a distinct parameter with its own prior, and conditioning on correlated random effects produces variance estimates that reflect the working model's phantom variable-induced correlation. Finally, the level-2 variances enter the working model via the augmented parameters, and MCMC updates the residual correlation and phantom variable scores. Following Merkle and Rosseel (2018), we use a positively skewed inverse gamma prior with shape and scale parameters $a = 1$ and $b = 0.5$ (or equivalently, hyperparameters $df_p = 2$ and $S_p = 1$), and we use a beta prior distribution for the correlation with both

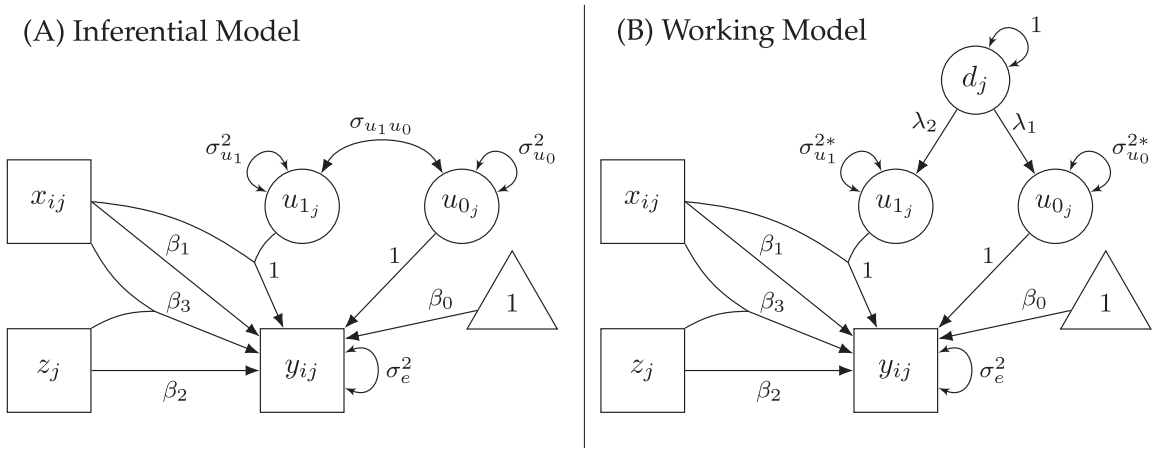

Figure 1. Path diagram of a multilevel model with a separation prior specification.
Note: Panel A is the path diagram of the inferential model corresponding to Equation 4. Panel B is the path diagram of the working model with a level-2 latent phantom variable (D). Parameters λ_1, λ_2, $\sigma^{2*}_{u_0}$, and $\sigma^{2*}_{u_1}$ are deterministically computed based on the equations given in the online supplemental material. Conjoining paths represent the deterministic multiplication of the variables.

shape parameters set to 1 (i.e., a flat prior over most of the correlation's −1 to +1 range). Interested readers can find a more technical summary of the MCMC estimation steps and conditional distributions in Section A of the online supplemental material.

Non-normal continuous predictors

In addition to incomplete categorical predictors, researchers often are faced with non-normally distributed continuous incomplete predictors. The multilevel complete data literature suggests that fixed and random effect parameters tend to be robust to violations of the distributional assumptions (Grilli & Rampichini, 2015; Jacqmin-Gadda et al., 2007; Maas & Hox, 2004a, 2004b; McCulloch & Neuhaus, 2011; Schielzeth et al., 2020; Verbeke & Lesaffre, 1996). However, these studies have focused on misspecification in the outcome model's distributional assumptions, not the predictor models, where parameter estimates are not of inferential interest. Instead, these predictor models aim to provide predictions when simulation imputations from the posterior predictive distribution. In the context of complete data outcomes, Schielzeth et al. (2020) found that violations of the random effect distribution resulted in poor estimates of specific random effects (i.e., the cluster residual), which are required for multilevel imputation. Thus, despite multilevel model's robustness to distributional violations for parameter inferences, it may be important to attempt to handle the non-normality in predictor's distributions.

The literature suggests several solutions to accounting for non-normally distributed variables, including copula-based (Gomes et al., 2019; Hollenbach et al., 2021; Käärik, 2006; Käärik & Käärik, 2009; Robbins et al., 2013) and transformation-based approaches (Goldstein et al., 2009; Lüdtke et al., 2020b). Thus far, copula-based approaches have focused on agnostic imputation schemes because they generally focus on modeling and imputing the unobserved variables conditional on the observed variables. In the context of multilevel data, Goldstein et al. (2009) discussed (but not studied in detail) the imputation of continuous non-normal responses using the Box–Cox transformation (Box & Cox, 1964) and argued that their approach is a multilevel extension of a Gaussian copula approach (p. 194, Section 5.3). A significant limitation of the Box-Cox transformation is that values must be positive. Because of this limitation, Yeo and Johnson (2000) proposed a transformation with similar properties to a Box–Cox transformation for positive values and accommodates negative values. The Yeo–Johnson transformation has shown promise when paired with single-level factored regression specifications (Lüdtke et al., 2020b). We focus on the Yeo–Johnson transformation because it is available in software programs (e.g., Blimp and the R package mdmb; Keller & Enders, 2021; Robitzsch & Lüdke, 2021) and readily extends to multilevel models.

The Yeo–Johnson procedure models a predictor, X, using a normalized proxy, X^*, with a mean and variance, μ_{X^*} and $\sigma^2_{X^*}$, respectively. The procedure uses the Yeo–Johnson transformation to link the proxy to the observed variable:

$$h_\lambda(X) = \begin{cases} \left((X+1)^\lambda - 1\right)/\lambda & \text{if} X \geq 0 \text{ and } \lambda \neq 0 \\ \ln(X+1) & \text{if} X \geq 0 \text{ and } \lambda = 0 \\ -\left((-X+1)^{2-\lambda} - 1\right)/(2-\lambda) & \text{if} X < 0 \text{ and } \lambda \neq 2 \\ -\ln(-X+1) & \text{if} X < 0 \text{ and } \lambda = 2 \end{cases} \tag{14}$$

where λ is a shape parameter and $h_\lambda(X) = X^*$. Importantly, the above power transformations subsume several common functions, including inverse, logarithmic, square root, identity function (i.e., $h_{\lambda=1}(X) = X$.), and Box-Cox transformations, the latter of which was discussed but not studied in detail by Goldstein et al. (2009).

The additional λ parameters must also be sampled, and details of this procedure are supplied in the Technical Appendix provided as part of the online supplemental material (Section A). Finally, imputations for missing observations are sampled from their appropriate factored conditional distributions (see equations 11 and 12) with the substitution of the models in Equation 14.

Returning to the sequential specification for predictor variables, the untransformed covariates appear in the focal regression from Equation 4, and the $f(X|Z)$ and $f(Z)$ terms in Equation 7 change to reflect the conditional distributions of the normalized scores.

$$\begin{aligned} f(X^*|Z) &= N_1\left(\gamma_0 + \gamma_1 z_j + g_{0j}, \sigma^2_{wX^*}\right) \\ f(Z^*) &= N_1\left(\mu_{Z^*}, \sigma^2_{bZ^*}\right) \end{aligned} \tag{15}$$

The additional λ parameters must also be sampled, and details of this procedure are supplied in the Technical Appendix provided as part of the online supplemental material (Section A). Finally, imputations for missing observations are sampled from their appropriate factored conditional distributions (see equations 11 and 12) with the substitution of the models in Equation 15.

MCMC estimation

This section briefly summarizes the MCMC estimation steps for the factored regression. While MCMC methods provide several ways of sampling imputations and parameters, we use a Gibbs sampler to break the complex multivariate distribution of the unknowns into separate conditional parts. For some quantities with nonstandard or intractable distributions, we use a Metropolis-Hastings step within the Gibbs sampler (Gelman et al., 2014; Hastings, 1970; Lynch, 2007). The following steps provide a conceptual recipe for the sampling procedure, and the online supplement (Section A) includes a detailed technical appendix that gives a complete summary of the estimation steps, including the full conditional distributions for each parameter.

1. Sample the focal regression model's slope coefficients, conditional on the current data, random effects, and level-1 residual variance.
2. Sample the level-2 random effects, conditional on the current data and model parameters. With an inverse Wishart prior for the level-2 covariance matrix, sampling follows standard multivariate normal expressions from the literature (Browne, 1998; Draper, 2008). With a separation-type prior, the random effect imputations condition on the inferential (focal) and working models in Figure 1.
3. Sample the level-1 residual variance, conditional on the current data, coefficients, and random effects.
4. Sample the level-2 covariance matrix, conditional on the current random effects. With an inverse Wishart prior for the level-2 covariance matrix, $\boldsymbol{\Sigma}_u$ is sampled in a single step. With a separation-type prior specification, the level-2 variance components and the random effect correlation are sampled in separate steps, as described previously.
5. For each incomplete predictor, sample the slope coefficients, random effects (if measured at level-1), between-cluster variance (if measured at level-1), and residual variance. These sampling steps rely on the same conditional distributions used to estimate the focal regression model. When applying the Yeo–Johnson procedure, the predictor's model also includes a shape parameter.
6. For each incomplete variable, sample imputations from the posterior predictive distribution. The distribution's center and spread depends on every model that features the incomplete variable, as described previously.

The MCMC algorithm simultaneously generates model parameters for Bayesian inference and multiple imputations. When using the former, the collection of parameter values obtained from the T iterations following the burn-in period estimate marginal posterior distributions, the center and spread of which are analogous to frequentist point estimates and standard errors. In this scenario, the imputations from each iteration t are simply a means to an end, which is to fill-in the data to simplify the estimation of parameters. In contrast, a multiple imputation analysis requires a small collection M imputed data sets (e.g., 20 or more; Graham et al., 2007) from the MCMC process. The researcher then fits the multilevel model of interest—the same model specified in $f(Y|X, Z)$—to each data set and then uses Rubin's (1987) pooling rules to obtain point estimates and standard errors. As noted previously, the literature to date has focused almost entirely on multiple imputation, and a major thrust of this study is to evaluate Bayesian inference and compare it to multiple imputation.

Computer simulation studies

This section outlines the five computer simulation studies that investigate factored regression specifications for multilevel models with cross-level interactions. As a worst-case baseline, we also include the multilevel extension of the just-another-variable imputation method that treats the cross-level product as a normally distributed level-1 variable. All five simulations used the analysis model given in Equation 4. The dependent variable was complete, and missing values on both predictors followed a missing at random mechanism with Y as the cause of missingness.

Simulations 1 and 2 investigated conditions where factored regression specifications described earlier correctly model the predictor distributions. More specifically, Simulation 1 investigated a continuous level-2 predictor, and Simulation 2 investigated a binary level-2 predictor. The multiple imputation aspects of Simulation 1 is comparable to Grund et al. (2021), but we additionally investigate Bayesian estimation and inference, we explore the effect of different prior distributions for the random effect covariance matrix. For the remaining three simulations, we systematically investigated the impact of misspecifying a distribution in one of the predictor models. Simulation 3 investigated a misspecification where the normal distribution for μ_{X_j} in Equation 6 is applied to a highly skewed and kurtotic set of latent group means that follows a

one degree of freedom chi-square distribution. Simulation 4 investigated a different between-cluster misspecification where the normal distribution for Z in Equation 6 or 7 is applied to a highly skewed and kurtotic level-2 predictor that follows a one degree of freedom chi-square distribution. Finally, Simulation 5 investigated a misspecification where the within-cluster normal distribution in the top expression of Equation 6 is applied to a level-1 predictor with highly skewed and kurtotic within-cluster residuals that follows a one degree of freedom chi-square distribution. We chose this chi-square distribution to induce extreme distributional misspecifications that likely represent an upper bound for what researchers might encounter in applied practice. Previous studies have not examined the robustness of factored regression specifications to such conditions in this context. For Simulations 4 and 5, we also investigate the possibility of applying a Yeo–Johnson transformation (Yeo & Johnson, 2000) proposed in the single-level imputation literature (Lüdtke et al., 2020b).

Overview of simulation conditions

Each simulation had four between-subjects factors: within-cluster sample size (n_j = 5, 15, and 25), number of clusters (J = 25, 50, and 200), intraclass correlation (ICC = 0.10 and 0.50), and missing data rate (15%, and 30%). Each study used 2,000 replications for each of the 36 conditions. For Simulation 2, we also included a condition where we investigated two sets of category proportions for the binary level-2 predictor: a 80/20 split and 50/50 split of 0 s and 1 s.

The conditions were chosen based on a review of guidelines from the literature, published Monte Carlo studies, generalizability to typical social science data, and considerations to methodological interests (e.g., evaluating behavior with small and large sample sizes). For example, Maas and Hox (2005) suggested that within cluster sample of n_j = 30 is the norm for level-1 in educational research; thus, we chose to vary the within cluster sample size at varying amounts under n_j = 30. For number of level-2 clusters, Maas and Hox (2005) suggested that 50 clusters are common in educational research, and Kreft and de Leeuw (1998) recommend a minimum of 30 clusters. Finally, the 200 clusters condition allowed us to investigate the large sample properties of the factored regression specification and how it might differ between imputation and Bayes posterior medians. Turning to the ICC condition, we chose ICC's that are representative of published research (Gulliford et al., 1999; Hedges & Hedberg, 2007; Murray & Blistein, 2003); thus, we chose an ICC of 0.10 to constitute a value for cross-sectional data, and an ICC of 0.50 to constitute a value for repeated measures data. The missing data rates we used represent a moderate and large amount of missing data (i.e., 15% and 30% on both predictors, respectively). The 15% condition is a threshold where traditional fully conditional specification begins to show biases (Enders et al., 2018; Enders et al., 2018; Grund et al., 2018), and the 30% condition should reveal distortions.

For all simulations, we set the population value of β_0 to equal to 10 and set the population means of X and Z to equal 0. In addition, we fixed the total variance of X and Z to 1 (when Z was continuous) and set the population variance of Y equal to 100. For the analysis model's random effects, we set the covariance between the random intercepts and slopes (i.e., $\sigma_{u_0u_1}$) to a value corresponding to a correlation equal to 0.30. We specified the random slope variance (i.e., $\sigma^2_{u_1}$) to account for 5% of the within-cluster outcome variance (Rights & Sterba, 2019). For the correlation structure, the slope coefficients were derived after setting the population correlations among all variables equal to 0.40, and the cross-level product term's slope was derived by setting its bivariate association with Y equal to 0.15. In the online supplemental material (Section F), we show the true multilevel model parameters for each condition, and provide a detailed description of their derivation.

Data generation

We generated the data using R programming language for statistical computing (R Core Team, 2020). To generate the predictors, we used the following equations, where $\mu_X = \mu_Z = 0$.

$$\begin{aligned} x_{ij} &= \mu_{Xj} + r_{ij(X)} \\ \mu_{Xj} &= \mu_X + r_{j(X)} \\ z^*_j &= \mu_Z + r_{j(Z)} \end{aligned} \tag{16}$$

The distributions of the variables depend on the simulation condition. For Simulation 1, we use the distributions given in Equation 8, where $z_j = z^*_j$. For Simulation 2, the $r_{j(Z)}$ term is on the latent response metric, and z^*_j is converted to a binary variable using the function given in Equation 10. For Simulations 3, the $r_{j(X)}$ residuals were generated based on a transformed one degree of freedom chi-square variable. To illustrate, the data generation for Simulation 3 was created with the following expressions

$$\begin{gathered} r_{ij(X)} \sim N\left(0, \sigma^2_{wX}\right) \\ \begin{bmatrix} \mu_{Xj} \\ z_j \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix} + \left(\begin{bmatrix} d^*_{j(X)} & \frac{d^*_{j(Z)}-1}{\sqrt{2}} \end{bmatrix} \mathbf{L} \right)^T \\ d^*_{j(X)} \sim N(0,1) \\ d^*_{j(Z)} = \chi^2(1) \end{gathered} \tag{17}$$

where $\mathbf{L}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_b$, the between-cluster covariance matrix from Equation 8. To generate the data, we drew pairs of normal and chi-square variates for each cluster, $d^*_{j(X)}$ and the $d^*_{j(Z)}$, respectively. The row vector to the left of $\mathbf{L}$ scales the variables to have a mean of 0 and a variance of 1, then multiplying scores by the Cholesky decomposition of the $\boldsymbol{\Sigma}_b$ matrix induces the desired correlation and scaling. In Simulation 4, $d^*_{j(X)}$ and $d^*_{j(Z)}$ switched roles, such that $d^*_{j(X)}$ became the rescaled chi-square variate. Finally, in Simulation 5, both $d^*_{j(X)}$ and $d^*_{j(Z)}$ were normally distributed variables, and $r_{ij(X)}$ was a rescaled chi-square variate with a mean of 0 and variance equal to σ^2_{wX}.

After generating the predictor scores and random effects, we generated Y scores by substituting these quantities into the analysis model from Equation 4 along with the derived population parameters. Finally, to generate data that are missing at random, we generated predictive probabilities of missingness using a logistic model with Y predicting a binary missing data indicator. The intercept parameter was specified to create the desired amount of missing data, and the logistic slope was derived as a positive value (i.e., missingness rates increased with the values of Y) that produced a pseudo-R^2 (McKelvey & Zavoina, 1975) equal to 0.50 (i.e., a very strong selection mechanism that would certainly induce bias when using listwise deletion). For X and Z, binary missing data indicators were first sampled from a binomial distribution with success rates equal to the predicted probabilities, and missing values were induced if the indicator score equaled 1.

Model fitting

Multiple imputations (i.e., both for factored regression imputation and just-another-variable imputation based on fully conditional specification) were generated using version 3 of Blimp software package (Keller & Enders, 2021).[1] For each replication, we generated 10,000 posterior draws after a 10,000 burn-in interval from a single chain. These values were chosen after assessing potential scale reduction factor diagnostics (Gelman & Rubin, 1992) from a sample of artificial data sets from various conditions, and we include simulation-specific convergence information in the online supplemental material (Section D). Ten imputations were generated by saving a data set after every 1,000th iteration following the burn-in period. We then used restricted maximum likelihood estimation (McNeish, 2017) with R and the *lme4* package (Bates et al., 2021) to fit the analysis model from Equation 4 to each filled-in data set. All estimates and standard errors were pooled using the *mitml* R package (Grund et al., 2021).

As discussed previously, the factored regression specification also generates Bayesian posterior summaries, and a major goal of this study is to evaluate Bayesian inference and compare it to gold standard factored imputation methods. Our simulation investigated both the joint model and sequential specification for the predictors. There is no reason to expect these methods to differ in this context, and our simulations confirmed that expectation. In the interest of space, we report the Bayesian results for the sequential specification because this approach accommodates the Yeo–Johnson (2000) transformation for skewed variables. The real data analysis example illustrates the joint model specification for predictors.

As a consequence of adopting a Bayesian approach to multiple imputation, we must also invoke prior distributions for the parameters in the analysis model and the supporting predictors models. As noted earlier, we adopt noninformative (flat) priors for regression coefficients and grand means, and we use inverse gamma priors for the focal model's residual variance and the residual variances in the supporting predictor models. These choices are not nearly as consequential as the prior distribution for the level-2 covariance matrix of the random effects (the $\boldsymbol{\Sigma}_u$ matrix in Equation 4); it is well documented that $\boldsymbol{\Sigma}_u$ is sensitive to prior specifications (see Gelman & Hill, 2007, Chapters 13 and 17 for discussion), and this sensitivity is exasperated by a small number of level-2 sampling units. We investigated two common noninformative inverse Wishart priors. The first is more informative because it adds the number of matrix dimensions plus 1 to the degrees of freedom, and it adds an identity matrix to the sum of squares and cross-products. The second is less informative because it subtracts the same value from the degrees of freedom and adds nothing to the sum of squares and cross-products. The Blimp software refers to these options as "prior1" and "prior2," respectively. We also

[1]Similar model-fitting procedures are available in the R package mdmb (Robitzsch & Lüdtke, 2021).

investigated the separation-type strategy described earlier that specifies unique priors on the individual random effect variances and their correlation.

Outcome measures

For all five simulations, we focused on two outcome measurements: percent bias and confidence interval coverage. The latter was only calculated for the fixed effects estimates because the literature suggests that symmetric, normal-theory confidence intervals are inappropriate for variance parameters (Maas & Hox, 2005; Snijders & Bosker, 2012). We defined percent bias as

$$\text{percent bias} = \frac{(\text{average estimate}) - (\text{true parameter})}{(\text{true parameter})} \times 100 \tag{18}$$

and confidence interval coverage as

$$95\%\ \text{C. I. coverage} = \frac{(\text{number of replications with population parameter in C. I.})}{(\text{total number of replications})}. \tag{19}$$

For Bayesian inference, the average estimate in the bias expression is the average posterior median, and coverage values were based on 95% credible intervals rather than confidence intervals. The posterior mean and mode could also be used, but we chose the median because it is the default point estimate in common software programs (Asparouhov & Muthén, 2010a; Gelman et al., 2015), and thus would be the summary most researchers would use in practice." These measures were calculated for all simulations and all condition within each simulation. As mentioned previously, complete-data estimates served as a baseline against which to compare other methods, as multilevel model parameters are not automatically free of bias. To facilitate interpretation, we used rules of thumb to assess the practical impact of bias and coverage; for percent bias the literature sometimes recommends that the bias should not exceed 10% (Kaplan, 1988), and we adopted Bradley's (1978) "liberal" criterion for coverage rates between 0.925 and 0.975.

Results

For all five simulations, we present a subset of results that illustrates the main findings. We focus on the 30% missing data condition because it would reveal each method's bias-inducing tendencies. For all simulations, the absolute values of the listwise deletion biases were consistently above 20%, suggesting that the missing at random selection mechanism was strong enough to induce nonresponse bias if not handled appropriately. We include the complete set of simulation conditions, including the severely biased listwise deletion and just-another-variable results, in the online supplemental material (Section K through Section V).

In the interest of space, we have presented the main findings of the simulation in tables, but we have also included trellis plots that we found helpful in deciphering the results. These plots are in Section B of the online supplemental material and provide a graphical depiction of the findings discussed throughout the results section.

Simulation 1: normally distributed predictors

To streamline the presentation, we first identified the prior distribution that provided the best parameter recovery with Bayesian inference, and we then compared that "winner" to its multiple imputations estimates. The effect of prior distributions tends to be much less pronounced when analyzing filled-in data sets, so we sought to keep that aspect of the comparison constant.

To begin, Table 1 gives the percentage bias values for the three different covariance matrix priors. The top half of the table illustrates the effects of varying the level-1 sample size while holding the level-2 sample size constant at $J = 50$, and the bottom half illustrates the effects of varying the level-2 sample size while holding the within-cluster sample size constant at $n_j = 15$. Not surprisingly, the choice of prior was most impactful on the variance–covariance matrix estimates. The graphs show that the separation prior was consistently superior to the inverse Wishart priors, as the random intercept and slope variance estimates generally exhibited minimal bias less than 10% in absolute value. The notable exception occurred in the design cell with $n_j = 5$ observations per cluster and $J = 50$ level-2 units, where the random slope variance for this prior was positively biased. The apparent downside of the separation prior is that covariance parameter estimates exhibited large negative bias values of about 30% (i.e., estimates closer to 0).[2] We suspect that most researchers would view correlation bias as a reasonable tradeoff for getting good estimates of the variance components. As such, we focus on the separation prior going forward. As mentioned previously, the prior choice was

[2]The separation prior parameterizes the random effect structure as a correlation and two variances. To provide comparability with the other priors, the covariance was computed by multiplying the estimated correlation at each iteration by the square root of the product of the variances.

Table 1. Comparing the posterior medians based on priors with 30% missing data rate for simulation 1.

		L1 = 5; L2 = 50			L1 = 15; L2 = 50			L1 = 25; L2 = 50		
ICC		W1	W2	SEP	W1	W2	SEP	W1	W2	SEP
0.1	β_0	−1.71	−0.57	−1.29	−1.19	−0.60	−0.97	−0.91	−0.47	−0.74
	β_1	−5.74	−7.23	−5.60	−0.50	−0.98	−0.96	0.01	−0.47	−0.63
	β_2	6.14	**−10.06**	−2.50	−5.21	**−13.89**	−7.26	−5.40	**−10.04**	−3.56
	β_3	−7.17	**−15.15**	−9.96	−1.99	−6.88	−3.20	−1.44	−5.27	−2.64
	$\sigma^2_{u_0}$	**−55.16**	**73.83**	1.80	**−22.98**	**28.61**	1.66	**−13.35**	**22.08**	2.06
	$\sigma^2_{u_1}$	**−46.34**	**150.54**	**31.81**	**−34.72**	**45.68**	3.67	**−24.17**	**32.10**	4.12
	σ_{u_0,u_1}	**−70.03**	**−26.29**	**−79.73**	−7.72	**11.40**	**−46.06**	5.08	**19.40**	**−30.46**
	σ^2_e	5.96	−2.55	−0.30	2.71	0.28	0.46	1.39	0.28	0.19
0.5	β_0	−1.31	−0.17	−0.89	−0.99	−0.34	−0.80	−0.95	−0.43	−0.83
	β_1	−4.77	−6.31	−5.59	−0.85	−1.22	−2.02	−0.25	−0.42	−1.06
	β_2	−7.49	**−14.93**	−9.78	−5.16	**−10.60**	−4.66	−5.52	−9.25	−3.45
	β_3	−8.15	**−14.49**	−7.78	−2.89	−7.97	−3.88	−2.64	−7.27	−3.89
	$\sigma^2_{u_0}$	−5.54	**18.72**	2.70	−1.12	**16.82**	3.34	−0.58	**16.86**	3.89
	$\sigma^2_{u_1}$	**−43.29**	**145.05**	**28.64**	**−44.37**	**46.06**	0.75	**−26.78**	**32.53**	2.06
	σ_{u_0,u_1}	**−30.03**	−8.50	**−60.09**	−6.90	**14.38**	**−34.67**	−6.18	**13.81**	**−26.67**
	σ^2_e	4.50	−0.46	0.07	2.28	0.24	0.49	0.99	0.10	0.13
		L1 = 15; L2 = 25			L1 = 15; L2 = 50			L1 = 15; L2 = 200		
ICC		W1	W2	SEP	W1	W2	SEP	W1	W2	SEP
0.1	β_0	−1.93	−0.82	−1.53	−1.19	−0.60	−0.97	−0.11	0.03	−0.07
	β_1	−1.09	−1.69	−1.12	−0.50	−0.98	−0.96	0.27	−0.24	−0.27
	β_2	−6.63	**−20.61**	**−11.32**	−5.21	**−13.89**	−7.26	−2.85	−3.99	−1.84
	β_3	−5.43	**−12.29**	−6.67	−1.99	−6.88	−3.20	−0.02	−1.46	−0.55
	$\sigma^2_{u_0}$	**−35.20**	**73.85**	4.25	**−22.98**	**28.61**	1.66	−5.78	6.06	0.49
	$\sigma^2_{u_1}$	**−38.79**	**117.92**	**11.26**	**−34.72**	**45.68**	3.67	**−15.42**	8.72	0.29
	σ_{u_0,u_1}	**−32.83**	**28.12**	**−63.38**	−7.72	**11.40**	**−46.06**	8.15	5.08	**−14.57**
	σ^2_e	2.88	−0.01	0.33	2.71	0.28	0.46	0.86	0.10	0.09
0.5	β_0	−1.92	−0.57	−1.37	−0.99	−0.34	−0.80	−0.07	0.11	−0.04
	β_1	−2.43	−2.86	−3.28	−0.85	−1.22	−2.02	0.10	−0.03	−0.37
	β_2	−4.55	**−12.80**	−4.76	−5.16	**−10.60**	−4.66	−0.43	−1.74	0.01
	β_3	−7.47	**−13.93**	−7.38	−2.89	−7.97	−3.88	0.96	−0.64	0.48
	$\sigma^2_{u_0}$	−5.23	**41.92**	7.63	−1.12	**16.82**	3.34	−0.05	3.84	0.85
	$\sigma^2_{u_1}$	**−44.45**	**117.75**	6.93	**−44.37**	**46.06**	0.75	**−14.15**	9.71	0.10
	σ_{u_0,u_1}	**−18.46**	**24.99**	**−53.71**	−6.90	**14.38**	**−34.67**	−1.62	3.55	**−10.93**
	σ^2_e	2.74	0.49	0.71	2.28	0.24	0.49	0.66	0.13	0.18

Note: The values are percent bias (Equation 18) for all parameters in Simulation 1 and compare the inverse Wishart with $df = d + 1$ prior (W1), the inverse Wishart with $df = d - 1$ prior (W2), and the separation prior (SEP). The top half of the table varies level-1 sample size holding level-2 sample size at 50, and the bottom half of the table varies level-2 sample size, holding level-1 at 15. Percent bias values with a magnitude greater than 10% are bolded.

much less noticeable when analyzing multiple imputations, so we sought to keep this aspect of the Bayes–imputation comparison constant.

Turning to the comparison of Bayesian estimation (BAY; based on the separation prior) and factored regression multiple imputation (IMP; also based on the separation prior), the Simulation 1 rows of Tables 2 and 3 give the percentage bias values for the regression slopes holding level-1 and level-2 constant, respectively. Complete-data estimates (CMP) are given as a comparison, as these represent a best-case scenario for restricted maximum likelihood estimation. The Bayesian and multiple imputation estimates generally tracked closely with the complete-data estimates. Turning to the level-2 covariance matrix, the Simulation 1 rows of Tables 4 and 5 give the percentage bias values for the regression slopes holding level-1 and level-2 constant, respectively. The largest differences occurred in the design cell with $n_j = 5$ observations per cluster and $J = 50$ level-2 units, where the complete-data estimates were more accurate. In most cases, there was very little difference between Bayesian estimation and multiple imputation. The notable exception was the covariance between the random intercepts and slopes, where the Bayes estimates exhibited greater bias. However, the tables do not tell the whole story, as Bayesian estimates of the covariance had dramatically smaller mean squared errors (multiple imputation mean squared errors were generally 25% to 50% larger). From a practical perspective, this means that the biased Bayesian estimates were closer to the true value, on average, despite their negative bias.

Although not displayed but available in the online supplemental material (Section B), the coverage results were somewhat unremarkable. Coverage values generally fell between 92.5% and 97.5% (Bradley's liberal criterion). When the ICC = 0.50, Bayesian coverage values improved slightly as the number of level-2 units increased, but this trend was very subtle and likely wouldn't be considered practically significant.

Table 2. Percent bias by level-1 sample sizes for regression slopes with 30% missing data rate.

			L1 = 5; L2 = 50			L1 = 15; L2 = 50			L1 = 25; L2 = 50		
	ICC		CMP	BAY	IMP	CMP	BAY	IMP	CMP	BAY	IMP
Sim 1	0.1	β_1	−0.93	−5.65	−5.60	0.06	−0.95	−0.96	−0.19	−0.69	−0.63
		β_2	4.08	−2.11	−2.50	−3.24	−7.43	−7.26	−0.97	−3.74	−3.56
		β_3	−1.91	−9.73	−9.96	−0.52	−2.84	−3.20	0.27	−3.02	−2.64
	0.5	β_1	−0.03	−5.63	−5.59	−0.20	−2.04	−2.02	−0.22	−1.11	−1.06
		β_2	−1.25	−9.10	−9.78	−0.85	−4.75	−4.66	0.22	−3.74	−3.45
		β_3	0.82	−8.26	−7.78	−0.90	−4.10	−3.88	−1.72	−3.74	−3.89
Sim 2 50/50	0.1	β_1	0.50	−3.23	−3.27	0.00	−1.93	−1.94	0.38	−0.98	−0.95
		β_2	8.81	**−12.89**	**−11.70**	0.67	**−19.34**	**−19.26**	−0.78	**−18.82**	**−18.83**
		β_3	−0.04	0.25	−0.22	−0.36	−1.17	−1.10	−0.71	−1.29	−1.24
	0.5	β_1	−0.31	−4.47	−4.50	0.15	−1.78	−1.78	0.08	−1.62	−1.58
		β_2	−0.51	**−12.61**	**−13.48**	−1.24	**−14.12**	**−13.69**	1.03	**−10.53**	**−11.13**
		β_3	0.26	−0.32	−0.15	−0.35	−1.25	−0.98	0.08	0.34	0.25
Sim 2 80/20	0.1	β_1	1.16	−2.55	−2.53	−0.31	−1.56	−1.56	0.35	−0.56	−0.57
		β_2	8.79	**−23.20**	**−21.21**	−3.47	**−34.49**	**−33.82**	−0.51	**−29.41**	**−29.33**
		β_3	−7.13	**−39.49**	**−38.88**	2.65	**−15.10**	**−15.50**	−2.70	**−15.83**	**−13.80**
	0.5	β_1	−0.30	−5.33	−5.45	0.09	−1.78	−1.77	0.36	−1.24	−1.25
		β_2	1.17	**−20.03**	**−20.12**	0.22	**−21.94**	**−21.66**	0.39	**−18.87**	**−18.64**
		β_3	−1.44	**−17.07**	**−15.60**	−0.97	−7.70	−7.66	−1.42	−3.74	−3.55
Sim 3	0.1	β_1	−0.11	−3.45	−3.13	−0.06	−1.06	−0.93	0.03	0.29	0.38
		β_2	0.86	−9.91	−9.41	0.53	0.12	−0.37	−1.45	−3.40	−3.20
		β_3	1.01	**−11.62**	**−11.74**	0.71	−5.22	−5.33	0.37	−3.61	−3.57
	0.5	β_1	0.42	−5.88	−5.73	−0.15	−2.42	−2.49	−0.14	−1.53	−1.59
		β_2	−1.05	−3.70	−3.42	2.03	−4.05	−3.79	−1.99	−7.19	−6.88
		β_3	−2.07	**−52.24**	**−51.65**	2.85	**−15.97**	**−17.67**	−3.31	**−24.72**	**−24.20**
Sim 4	0.1	β_1	−0.50	−4.77	−4.73	0.10	−0.58	−0.52	−0.08	0.22	0.29
		β_2	2.51	**−14.01**	**−15.00**	−1.31	−9.07	−8.94	−0.40	−5.40	−4.95
		β_3	−1.84	**−12.85**	**−11.70**	0.18	−3.19	−3.16	−0.45	−2.99	−2.98
	0.5	β_1	0.05	−4.92	−4.97	−0.56	−1.96	−1.96	0.26	0.42	0.46
		β_2	0.87	−5.52	−5.87	1.82	−1.37	−1.62	−0.78	−2.88	−3.09
		β_3	0.11	−4.49	−5.00	−0.88	−1.73	−1.72	0.08	1.32	0.96
Sim 5	0.1	β_1	0.47	6.89	6.74	−0.27	**12.21**	**12.12**	−0.41	**13.09**	**13.02**
		β_2	0.13	**−14.32**	**−14.81**	0.25	−3.83	−3.67	−1.00	−3.88	−4.54
		β_3	−3.41	3.33	3.64	2.17	**31.36**	**31.30**	0.02	**29.34**	**29.31**
	0.5	β_1	0.01	−1.13	−1.23	−0.07	4.78	4.54	0.19	7.34	7.22
		β_2	−1.09	**−10.40**	**−10.45**	−0.35	−9.53	−9.40	0.26	−8.96	−8.63
		β_3	−1.76	2.08	2.27	−0.21	**17.52**	**17.89**	−0.43	**18.80**	**18.52**
Sim 5 Reversed	0.1	β_1	−0.30	1.00	1.63	0.23	−0.75	−0.54	0.15	−0.75	−0.61
		β_2	1.54	−1.73	−1.74	1.54	2.81	2.68	0.68	4.04	4.06
		β_3	−0.53	**−13.69**	**−13.41**	0.29	**−12.41**	**−12.19**	−0.19	**−12.84**	**−12.82**
	0.5	β_1	0.31	3.80	4.09	0.26	3.07	3.17	0.27	3.08	3.15
		β_2	−0.03	−4.35	−4.34	0.48	−1.92	−2.03	−0.63	−2.76	−2.93
		β_3	0.02	−7.26	−7.49	−0.32	−9.90	−9.88	0.14	**−10.36**	**−10.24**

Note: The values are percent bias (Equation 18) for the regression coefficients as a function of level-1 sample size while holding level-2 sample size constant for complete data estimates (CMP), bayesian estimation (BAY), and factored regression imputation (IMP). Percent bias values with a magnitude greater than 10% are bolded.

Simulation 2: binary level-2 predictor

Turning to Simulation 2, recall that the level-2 moderator was a binary dummy code with either a 50/50 or 80/20 split of 0 s and 1 s, respectively. The Simulation 2 rows of Tables 2 to 5 provide the results for the two split conditions for the three estimators (complete-data, Bayesian, and factored multiple imputation). Tables 2 and 3 provide the percentage bias for the regression slopes, and Tables 4 and 5 provide percentage bias for the level-2 covariance matrix parameters.

The main new finding of this study is that the level-2 dummy code's lower-order effect was sensitive to the number of level-2 units, although the accuracy of β_2 was also dependent on the category proportions. When there was a 50/50 split between categories, factored regression imputation had acceptable estimates (approximately 10% bias) with 50 level-2 clusters and 25 level-1 units; however, larger level-2 sample sizes were needed to achieve the same reduction in bias when the categorical predictor had an 80/20 split of 0 s and 1 s. Importantly, the cross-level interaction effect was effectively unbiased in most cases. One notable finding on this cross-level interaction effect was the ICC, where the cross-level interaction was particularly biased in the 80/20 split with ICC = 0.1 conditions. This effect was only unbiased in the largest level-2 sample size condition; however, to reiterate that this is for a 30% missing data rate at level-2 units with a rather extreme binary split. Random effect

Table 3. Percent bias by level-2 sample sizes for regression slopes with 30% missing data rate.

			L1 = 15; L2 = 25			L1 = 15; L2 = 50			L1 = 15; L2 = 200		
	ICC		CMP	BAY	IMP	CMP	BAY	IMP	CMP	BAY	IMP
Sim 1	0.1	β_1	0.28	−1.15	−1.12	0.06	−0.95	−0.96	0.06	−0.25	−0.27
		β_2	−0.12	**−10.67**	**−11.32**	−3.24	−7.43	−7.26	−1.13	−1.80	−1.84
		β_3	0.31	−6.65	−6.67	−0.52	−2.84	−3.20	0.21	−0.53	−0.55
	0.5	β_1	0.03	−3.27	−3.28	−0.20	−2.04	−2.02	0.35	−0.39	−0.37
		β_2	1.55	−4.42	−4.76	−0.85	−4.75	−4.66	0.12	0.01	0.01
		β_3	−0.70	−6.86	−7.38	−0.90	−4.10	−3.88	0.78	0.55	0.48
Sim 2 50/50	0.1	β_1	0.09	−2.85	−2.82	0.00	−1.93	−1.94	0.02	−0.50	−0.49
		β_2	−2.45	**−34.46**	**−34.55**	0.67	**−19.34**	**−19.26**	0.22	−7.98	−8.06
		β_3	0.32	0.26	0.06	−0.36	−1.17	−1.10	0.00	−0.70	−0.74
	0.5	β_1	0.05	−3.35	−3.53	0.15	−1.78	−1.78	0.01	−0.28	−0.30
		β_2	−1.93	**−21.54**	**−22.93**	−1.24	**−14.12**	**−13.69**	1.23	−2.87	−2.91
		β_3	−0.01	0.27	0.57	−0.35	−1.25	−0.98	0.25	−0.45	−0.39
Sim 2 80/20	0.1	β_1	0.07	−2.02	−2.04	−0.31	−1.56	−1.56	0.05	−0.24	−0.24
		β_2	−5.24	**−46.66**	**−45.80**	−3.47	**−34.49**	**−33.82**	−0.11	**−11.15**	**−10.83**
		β_3	2.64	**−22.89**	**−23.14**	2.65	**−15.10**	**−15.50**	−0.27	−8.50	−8.01
	0.5	β_1	0.48	−3.41	−3.34	0.09	−1.78	−1.77	0.15	−0.35	−0.39
		β_2	−7.39	**−38.77**	**−37.67**	0.22	**−21.94**	**−21.66**	1.07	−6.26	−6.39
		β_3	−3.01	−6.27	−6.42	−0.97	−7.70	−7.66	−0.14	−1.45	−1.12
Sim 3	0.1	β_1	−0.91	−2.82	−2.64	−0.06	−1.06	−0.93	0.01	0.53	0.59
		β_2	2.09	**−10.49**	**−10.26**	0.53	0.12	−0.37	0.26	7.57	7.56
		β_3	1.60	−9.38	−8.89	0.71	−5.22	−5.33	0.10	2.49	2.39
	0.5	β_1	−0.43	−4.71	−4.74	−0.15	−2.42	−2.49	0.23	−0.03	−0.04
		β_2	1.76	**−11.29**	**−11.36**	2.03	−4.05	−3.79	−0.10	4.67	4.57
		β_3	1.96	**−34.66**	**−36.51**	2.85	**−15.97**	**−17.67**	2.26	−7.33	−7.61
Sim 4	0.1	β_1	−0.27	−2.00	−2.23	0.10	−0.58	−0.52	0.27	0.39	0.41
		β_2	−0.93	**−14.00**	**−14.27**	−1.31	−9.07	−8.94	0.54	0.94	0.80
		β_3	0.82	−6.61	−7.33	0.18	−3.19	−3.16	0.29	−1.41	−1.46
	0.5	β_1	−0.13	−3.35	−3.31	−0.56	−1.96	−1.96	0.17	0.45	0.39
		β_2	−0.62	−6.62	−6.60	1.82	−1.37	−1.62	−0.75	0.46	0.68
		β_3	0.34	−6.24	−6.49	−0.88	−1.73	−1.72	−0.57	1.24	1.41
Sim 5	0.1	β_1	−0.60	7.08	6.88	−0.27	**12.21**	**12.12**	−0.11	**16.20**	**16.17**
		β_2	2.26	−5.39	−4.82	0.25	−3.83	−3.67	0.37	−2.30	−2.51
		β_3	−0.96	**15.93**	**16.57**	2.17	**31.36**	**31.30**	0.42	**31.69**	**31.95**
	0.5	β_1	0.15	2.16	1.96	−0.07	4.78	4.54	−0.05	8.17	8.10
		β_2	−1.61	**−14.79**	**−14.64**	−0.35	−9.53	−9.40	−0.62	−6.77	−6.71
		β_3	0.65	**11.10**	**11.43**	−0.21	**17.52**	**17.89**	−0.08	**21.93**	**21.82**
Sim 5 Reversed	0.1	β_1	−0.73	−0.89	−0.67	0.23	−0.75	−0.54	0.05	−1.14	−1.02
		β_2	0.88	−1.34	−1.17	1.54	2.81	2.68	0.44	4.66	4.58
		β_3	−0.30	**−13.88**	**−13.83**	0.29	**−12.41**	**−12.19**	−0.29	**−12.43**	**−12.31**
	0.5	β_1	−0.02	3.50	3.59	0.26	3.07	3.17	0.04	2.81	2.83
		β_2	0.10	−4.68	−4.64	0.48	−1.92	−2.03	0.17	−0.55	−0.51
		β_3	−1.64	−8.53	−8.34	−0.32	−9.90	−9.88	0.43	−9.70	−9.62

Note: The values are percent bias (Equation 18) for the regression coefficients as a function of level-2 sample size while holding level-1 sample size constant for complete data estimates (CMP), bayesian estimation (BAY), and factored regression imputation (IMP). Percent bias values with a magnitude greater than 10% are bolded.

estimates were consistent with the first simulation, with nothing new to highlight. Finally, coverage values also largely mimicked the findings of Simulation 1, so we point interested readers to the online supplement.

Simulation 3: non-normal latent means at level-2

For Simulation 3, the marginal distribution of the level-2 latent group means was a single degree of freedom chi-square that was rescaled to the desired mean and variance (0 and σ^2_{bX}, respectively). The main purpose of this simulation was to investigate a misspecification where the normal distribution for μ_{Xj} in Equation 6 is applied to skewed and kurtotic latent group means. This misspecification is interesting to consider because there is currently no obvious solution to the problem. For example, the Yeo–Johnson transformation (Yeo & Johnson, 2000) that we consider in the next two simulations is appropriate for manifest variables, but no analog currently exists for non-normal latent variables. Note that the level-2 moderator was normally distributed, as it was in the first simulation.

Tables 2 to 5 give the percentage bias values for each estimator in the Simulation 3 rows. The main new finding of this study is that non-normal latent means primarily impacted the interaction coefficient. Moreover, bias depended on the level-2 sample size, particularly in the ICC = 0.50 conditions. Although

Table 4. Percent bias by level-1 sample sizes for variance parameters with 30% missing data rate.

			L1 = 5; L2 = 50			L1 = 15; L2 = 50			L1 = 25; L2 = 50		
	ICC		CMP	BAY	IMP	CMP	BAY	IMP	CMP	BAY	IMP
Sim 1	0.1	$\sigma^2_{u_0}$	7.82	−1.36	1.80	0.42	−2.25	1.66	−0.04	−2.13	2.06
		$\sigma^2_{u_1}$	9.20	**22.74**	**31.81**	−0.98	−0.25	3.67	0.66	0.06	4.12
		σ_{u_0,u_1}	−3.10	**−40.13**	**−79.73**	−1.55	**−13.89**	**−46.06**	−0.84	−6.21	**−30.46**
	0.5	$\sigma^2_{u_0}$	−0.45	−2.18	2.70	−0.31	−1.31	3.34	0.19	−0.59	3.89
		$\sigma^2_{u_1}$	5.55	**20.07**	**28.64**	−0.83	−1.97	0.75	0.40	−0.92	2.06
		σ_{u_0,u_1}	−0.44	**−30.06**	**−60.09**	0.46	**−12.00**	**−34.67**	−2.94	**−10.08**	**−26.67**
Sim 2	0.1	$\sigma^2_{u_0}$	6.21	−3.96	−1.77	−0.38	−2.97	0.56	−0.01	−1.49	2.67
50/50		$\sigma^2_{u_1}$	8.59	**21.46**	**28.70**	−1.66	−2.22	0.98	−0.53	−0.67	3.13
		σ_{u_0,u_1}	−8.22	**−54.42**	**−85.39**	−0.23	**−29.30**	**−56.05**	−0.77	**−21.74**	**−41.83**
	0.5	$\sigma^2_{u_0}$	−1.31	−3.17	1.64	0.34	−0.26	4.40	0.35	−0.22	4.30
		$\sigma^2_{u_1}$	4.44	**15.08**	**21.05**	0.44	−0.29	1.13	1.29	1.42	3.88
		σ_{u_0,u_1}	−0.34	**−46.70**	**−70.22**	2.58	**−23.53**	**−42.47**	1.96	**−16.88**	**−32.61**
Sim 2	0.1	$\sigma^2_{u_0}$	6.40	−3.12	−1.46	−0.42	−3.12	0.46	0.16	−2.02	1.89
80/20		$\sigma^2_{u_1}$	**10.41**	**19.64**	**27.95**	−1.42	−3.91	−0.46	−0.68	−2.75	0.88
		σ_{u_0,u_1}	−5.45	**−49.13**	**−82.92**	0.09	**−17.83**	**−49.76**	−1.19	**−14.19**	**−36.92**
	0.5	$\sigma^2_{u_0}$	−1.07	−3.99	0.75	0.25	−0.61	3.89	0.35	−1.07	3.43
		$\sigma^2_{u_1}$	5.12	**13.60**	**19.27**	0.76	−2.07	−0.71	1.19	−1.61	1.39
		σ_{u_0,u_1}	1.00	**−41.83**	**−68.65**	2.56	**−17.53**	**−39.25**	1.89	**−13.51**	**−30.09**
Sim 3	0.1	$\sigma^2_{u_0}$	9.05	3.11	6.99	0.55	−1.49	2.08	−0.60	−1.82	2.41
		$\sigma^2_{u_1}$	9.25	**22.17**	**30.68**	0.16	0.04	4.35	0.85	−0.59	3.50
		σ_{u_0,u_1}	−6.94	**−33.69**	**−78.30**	−1.39	**−11.78**	**−46.18**	0.67	−4.36	**−29.92**
	0.5	$\sigma^2_{u_0}$	1.01	1.05	5.50	0.27	1.51	6.00	−0.11	0.40	4.73
		$\sigma^2_{u_1}$	9.23	**26.57**	**35.47**	0.68	−0.90	1.95	−1.01	−4.40	−1.81
		σ_{u_0,u_1}	4.29	**−28.20**	**−57.91**	1.96	**−11.85**	**−34.70**	−0.91	−9.98	**−28.22**
Sim 4	0.1	$\sigma^2_{u_0}$	9.22	2.94	5.79	0.86	−1.62	2.53	−0.57	−1.27	3.12
		$\sigma^2_{u_1}$	**11.00**	**28.78**	**35.85**	−0.15	3.29	7.66	0.41	6.52	**11.10**
		σ_{u_0,u_1}	−8.88	**−40.13**	**−78.28**	−0.81	−5.47	**−38.74**	1.08	2.89	**−22.79**
	0.5	$\sigma^2_{u_0}$	−0.92	−1.28	3.49	−0.05	0.17	5.01	0.83	1.44	6.03
		$\sigma^2_{u_1}$	7.88	**28.86**	**37.96**	−0.73	2.73	5.54	−0.96	5.06	8.82
		σ_{u_0,u_1}	0.28	**−25.15**	**−56.27**	−3.47	−7.51	**−30.34**	1.41	3.23	**−15.24**
Sim 5	0.1	$\sigma^2_{u_0}$	3.17	−7.39	−1.27	1.73	−3.98	−0.07	−0.17	−2.33	2.48
		$\sigma^2_{u_1}$	7.71	**223.33**	**245.80**	2.43	**148.17**	**161.26**	−0.68	**121.42**	**133.39**
		σ_{u_0,u_1}	**−10.92**	**68.50**	**−22.70**	1.78	**75.41**	**18.74**	0.04	**76.54**	**36.27**
	0.5	$\sigma^2_{u_0}$	−0.50	−5.46	−1.01	0.62	−2.73	1.64	−0.43	−3.20	1.11
		$\sigma^2_{u_1}$	9.63	**110.63**	**127.61**	1.02	**77.82**	**84.33**	0.63	**69.30**	**75.51**
		σ_{u_0,u_1}	−1.01	**−16.68**	**−48.27**	2.14	6.58	**−17.64**	−2.15	**12.14**	−8.45
Sim 5	0.1	$\sigma^2_{u_0}$	0.60	1.42	4.21	−0.30	−0.08	3.90	−0.89	−0.32	3.90
Reversed		$\sigma^2_{u_1}$	7.04	0.55	**14.58**	0.22	**−19.40**	**−14.30**	−0.48	**−22.25**	**−17.79**
		σ_{u_0,u_1}	−7.21	**24.20**	**−50.38**	0.09	**14.05**	**−24.98**	−0.79	**11.89**	**−14.61**
	0.5	$\sigma^2_{u_0}$	−0.78	−0.09	4.20	0.97	3.06	7.68	−0.63	1.13	5.64
		$\sigma^2_{u_1}$	7.91	1.77	**14.36**	0.42	**−16.06**	**−11.70**	−0.10	**−17.73**	**−13.78**
		σ_{u_0,u_1}	0.91	**12.31**	**−30.77**	−0.91	9.48	**−14.55**	0.23	**10.34**	−8.14

Note: The values are percent bias (Equation 18) for the random effect variance parameter as a function of level-1 sample size while holding level-2 sample size constant for complete data estimates (CMP), bayesian estimation (BAY), and factored regression imputation (IMP). Percent bias values with a magnitude greater than 10% are bolded.

we know of no studies that have considered consistency properties in this context, the simulation results suggest that estimates might be consistent under this misspecification, as β_3's bias decreased as the level-2 sample size increased; reading from left to right in Table 3, the estimates exhibited large biases when $J = 25$ that dropped below 10% when $J = 200$. Examining a wider range of level-2 sample sizes could identify the more precise point at which asymptotic properties take hold (if they do). Random effect estimates were consistent with the first two simulations, with nothing new to highlight. Finally, coverage values largely mimic the findings of Simulation 1, so we point interested readers to the online supplement.

Simulation 4: non-normal level-2 predictor

For Simulation 4, the marginal distribution of the level-2 moderator was a single degree of freedom chi-square that was rescaled to the desired mean and variance (0 and σ^2_{bZ}, respectively). The main purpose of this simulation was to investigate a misspecification where the normal distribution for Z in Equation 6 is applied to a skewed and kurtotic moderator. The Simulation 3 section of Tables 2 to 5 give the percentage bias values for the three estimators.

The main new finding of this study is that non-normality primarily impacted the moderator's lower-order slope, and the interaction coefficient generally exhibited

Table 5. Percent bias by level-2 sample sizes for variance parameters with 30% missing data rate.

			L1 = 15; L2 = 25			L1 = 15; L2 = 50			L1 = 15; L2 = 200		
	ICC		CMP	BAY	IMP	CMP	BAY	IMP	CMP	BAY	IMP
Sim 1	0.1	$\sigma^2_{u_0}$	1.36	−3.68	4.25	0.42	−2.25	1.66	0.18	−0.46	0.49
		$\sigma^2_{u_1}$	1.22	2.32	**11.26**	−0.98	−0.25	3.67	−0.65	−0.71	0.29
		σ_{u_0,u_1}	−4.98	**−21.74**	**−63.38**	−1.55	**−13.89**	**−46.06**	0.59	−2.56	**−14.57**
	0.5	$\sigma^2_{u_0}$	−0.46	−2.20	7.63	−0.31	−1.31	3.34	0.06	−0.26	0.85
		$\sigma^2_{u_1}$	2.44	−0.69	6.93	−0.83	−1.97	0.75	0.55	−0.28	0.10
		σ_{u_0,u_1}	1.24	**−22.35**	**−53.71**	0.46	**−12.00**	**−34.67**	1.08	−3.99	**−10.93**
Sim 2 50/50	0.1	$\sigma^2_{u_0}$	0.76	−3.69	3.68	−0.38	−2.97	0.56	−0.26	−0.63	0.25
		$\sigma^2_{u_1}$	2.29	2.57	**11.07**	−1.66	−2.22	0.98	−0.26	−1.00	−0.18
		σ_{u_0,u_1}	−3.90	**−41.23**	**−73.50**	−0.23	**−29.30**	**−56.05**	0.59	**−10.21**	**−20.99**
	0.5	$\sigma^2_{u_0}$	−0.92	−2.72	7.18	0.34	−0.26	4.40	0.38	0.37	1.48
		$\sigma^2_{u_1}$	−0.76	−0.84	4.91	0.44	−0.29	1.13	0.57	−0.37	−0.25
		σ_{u_0,u_1}	−2.55	**−39.59**	**−65.18**	2.58	**−23.53**	**−42.47**	0.08	−7.52	**−14.34**
Sim 2 80/20	0.1	$\sigma^2_{u_0}$	1.85	−2.03	5.44	−0.42	−3.12	0.46	−0.23	−0.73	0.12
		$\sigma^2_{u_1}$	2.81	0.30	8.86	−1.42	−3.91	−0.46	−0.24	−1.13	−0.35
		σ_{u_0,u_1}	−0.88	**−27.59**	**−67.43**	0.09	**−17.83**	**−49.76**	0.53	−6.04	**−17.26**
	0.5	$\sigma^2_{u_0}$	−0.80	−2.98	6.72	0.25	−0.61	3.89	0.39	−0.03	1.08
		$\sigma^2_{u_1}$	−0.80	−2.86	2.11	0.76	−2.07	−0.71	0.58	−1.04	−0.74
		σ_{u_0,u_1}	−1.29	**−34.59**	**−61.76**	2.56	**−17.53**	**−39.25**	0.08	−6.40	**−13.60**
Sim 3	0.1	$\sigma^2_{u_0}$	2.81	−2.94	4.63	0.55	−1.49	2.08	−0.10	1.47	2.40
		$\sigma^2_{u_1}$	2.04	1.88	**10.41**	0.16	0.04	4.35	0.29	1.75	2.87
		σ_{u_0,u_1}	−1.48	**−17.94**	**−63.52**	−1.39	**−11.78**	**−46.18**	0.86	2.46	**−10.50**
	0.5	$\sigma^2_{u_0}$	0.66	0.48	**10.14**	0.27	1.51	6.00	0.12	1.78	2.84
		$\sigma^2_{u_1}$	1.31	−1.58	4.18	0.68	−0.90	1.95	0.69	1.97	2.44
		σ_{u_0,u_1}	0.72	**−25.34**	**−57.16**	1.96	**−11.85**	**−34.70**	0.16	2.38	−5.76
Sim 4	0.1	$\sigma^2_{u_0}$	0.75	−4.32	3.31	0.86	−1.62	2.53	0.68	0.31	1.33
		$\sigma^2_{u_1}$	1.50	7.77	**17.32**	−0.15	3.29	7.66	−0.49	6.31	7.41
		σ_{u_0,u_1}	−0.15	**−16.93**	**−59.89**	−0.81	−5.47	**−38.74**	0.01	4.28	−7.93
	0.5	$\sigma^2_{u_0}$	−0.05	−0.94	9.03	−0.05	0.17	5.01	0.15	1.15	2.27
		$\sigma^2_{u_1}$	2.10	6.51	**14.63**	−0.73	2.73	5.54	0.39	6.90	7.58
		σ_{u_0,u_1}	3.00	**−14.12**	**−47.28**	−3.47	−7.51	**−30.34**	0.35	6.58	−1.35
Sim 5	0.1	$\sigma^2_{u_0}$	−0.71	−8.08	−0.63	1.73	−3.98	−0.07	0.63	−0.92	0.51
		$\sigma^2_{u_1}$	3.94	**150.32**	**174.09**	2.43	**148.17**	**161.26**	0.56	**138.54**	**142.54**
		σ_{u_0,u_1}	−0.91	**58.82**	**−16.59**	1.78	**75.41**	**18.74**	0.77	**99.54**	**79.53**
	0.5	$\sigma^2_{u_0}$	−1.68	−5.46	3.81	0.62	−2.73	1.64	0.28	−2.77	−1.77
		$\sigma^2_{u_1}$	3.36	**90.37**	**106.83**	1.02	**77.82**	**84.33**	0.30	**74.43**	**75.84**
		σ_{u_0,u_1}	0.64	−6.76	**−42.94**	2.14	6.58	**−17.64**	1.36	**24.10**	**15.68**
Sim 5 Reversed	0.1	$\sigma^2_{u_0}$	0.97	−1.67	6.01	−0.30	−0.08	3.90	0.36	4.58	5.86
		$\sigma^2_{u_1}$	0.90	**−17.45**	−7.62	0.22	**−19.40**	**−14.30**	0.51	**−20.39**	**−18.51**
		σ_{u_0,u_1}	1.79	9.16	**−44.36**	0.09	**14.05**	**−24.98**	2.18	**26.66**	**12.73**
	0.5	$\sigma^2_{u_0}$	−0.17	2.62	**12.24**	0.97	3.06	7.68	−0.06	1.62	2.68
		$\sigma^2_{u_1}$	0.38	**−16.02**	−7.58	0.42	**−16.06**	**−11.70**	0.59	**−16.36**	**−15.24**
		σ_{u_0,u_1}	−6.38	2.95	**−33.91**	−0.91	9.48	**−14.55**	0.96	**14.29**	6.94

Note: The values are percent bias (Equation 18) for the random effect variance parameter as a function of level-2 sample size while holding level-1 sample size constant for complete data estimates (CMP), bayesian estimation (BAY), and factored regression imputation (IMP). Percent bias values with a magnitude greater than 10% are bolded.

little or no bias. The ICC again played a role, although this time bias values were largest in the ICC = 0.10 condition. The simulation results again suggest that estimates might be consistent under this misspecification, as β_2 's bias decreased as the level-2 sample size increased; reading from left to right in Table 3, bias values dropped below 10% when $J = 50$ and were effectively zero when $J = 200$. Random effect estimates were consistent with the first two simulations, with nothing new to highlight. Finally, coverage values largely mimic the findings of Simulation 1, so we point interested readers to the online supplement.

As discussed earlier, one advantage of using a sequential predictor model is that it is straightforward to accommodate non-normal continuous conditional distributions. In the single-level regression context, Lüdtke et al. (2020b) proposed the Yeo–Johnson (Yeo & Johnson, 2000) power transformation for non-normal data. We conducted a follow-up simulation that applied the Yeo–Johnson transformation to the non-normal level-2 variable. Importantly, these analyses are still misspecified because the Yeo–Johnson transformation is tasked with approximating the true data-generating distribution, which is a chi-square. Our goal for this simulation is to mimic applied practice where a researcher, without knowledge of the true population distribution, would inappropriately apply a Yeo–Johnson transformation. Because the single

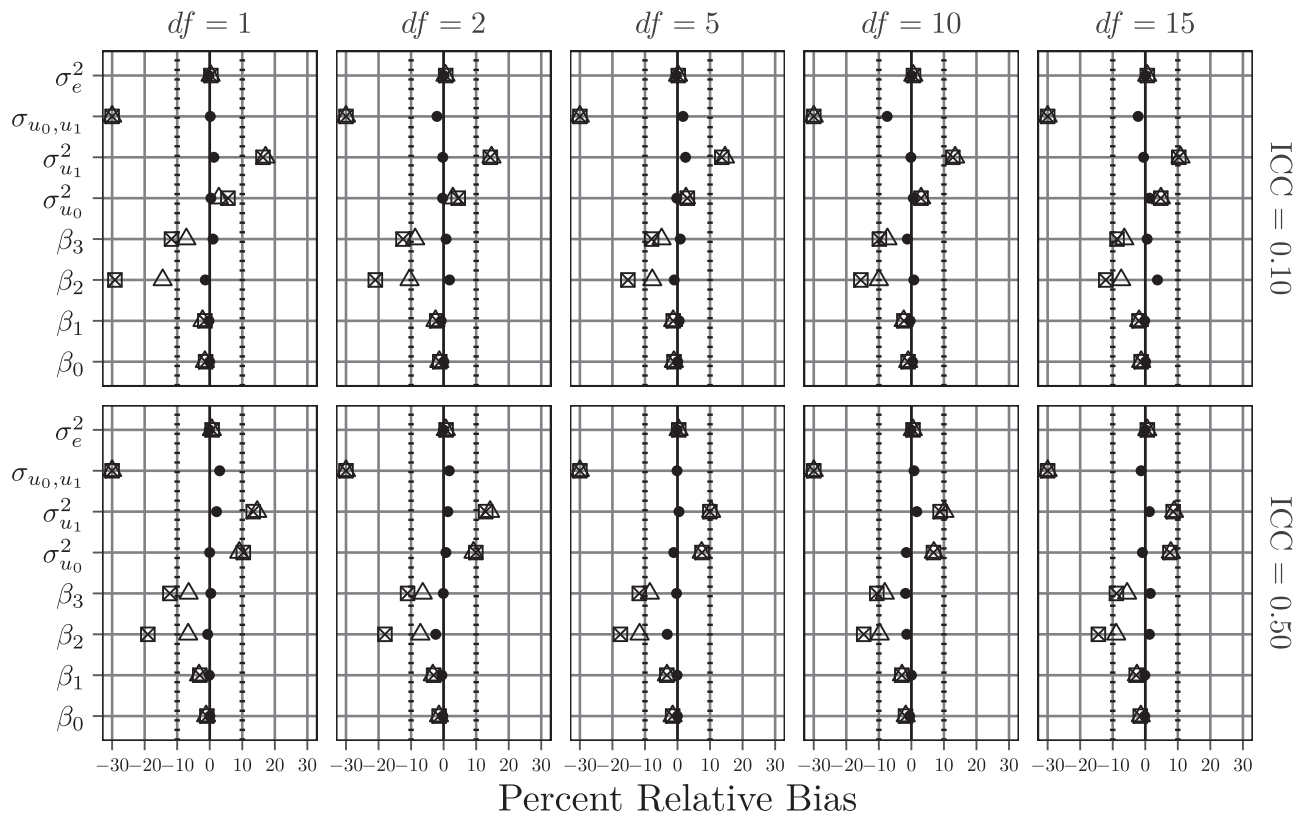

Figure 2. Bias for Bayesian estimation with varying degrees of non-normal level-2 moderator.
Note: Trellis plots displaying percentage bias values from Bayesian estimation with a Yeo-Johnson transformation (⊠) applied to a level-2 predictor with varying degrees of non-normality. Complete-data estimates (●) from restricted maximum likelihood and Bayesian estimation without the transformation (△) are shown as a comparison. The within-cluster sample size is 15 and the number of level-2 clusters is 25. Values are truncated to ±30 bias. The sample sizes were chosen based on the conditions with observed bias.

degree of freedom chi-square distribution is arguably a worse-case scenario misspecification (skewness and excess kurtosis values were approximately 2.8 and 12), the follow-up simulations varied the shape of the moderator's distribution using chi-square variables with degrees of freedom values equal to 2 (skewness ≈ 2, kurtosis ≈ 6), 5 (skewness ≈ 1.26, kurtosis ≈ 2.40), 10 (skewness ≈ .90, kurtosis ≈ 1.19), and 15 (skewness ≈ .73, kurtosis ≈ .79). For this simulation, we fixed the level-1 and level-2 sample sizes at values that exhibited the most bias, $n_j = 15$ and $J = 25$.

Figure 2 shows trellis plots of Bayesian estimation with and without the Yeo–Johnson transformation (it is also possible to use multiple imputation with the transformation, but we exclude these results in the interest of space) and the tabular results are offered in the online supplemental (Section C). Although the bias values generally aren't very extreme to begin with, the figure shows that the Yeo–Johnson transformation did not reduce bias. In fact, percentage bias values for the transformation were somewhat larger than the untransformed results. We attribute this seemingly counterintuitive finding to the fact that the level-2 sample size is very small at $J = 25$[3], which makes it difficult to estimate the shape parameter (e.g., even a single outlying observation could exert substantial influence on the transformation); therefore, a more informative prior on the shape parameter may be required.

Simulation 5: non-normal within-cluster variance

For Simulation 5, recall that the distribution of the within-cluster residuals was a single degree of freedom chi-square that was rescaled to the desired mean and variance (0 and σ^2_{wX}, respectively). The main purpose of this simulation was to investigate a misspecification where the within-cluster normal distribution in the top expression of Equation 6 is applied to a level-1 predictor with skewed and kurtotic within-cluster residuals. Tables 2 to 5 give the percentage bias values for each estimator in the Simulation 5 rows.

[3]We also investigated the $J = 50$ condition and included this figure in Section I of the online supplemental material; however, the findings were essentially the same, and the regression slopes for the normal distribution model were unbiased in all conditions.

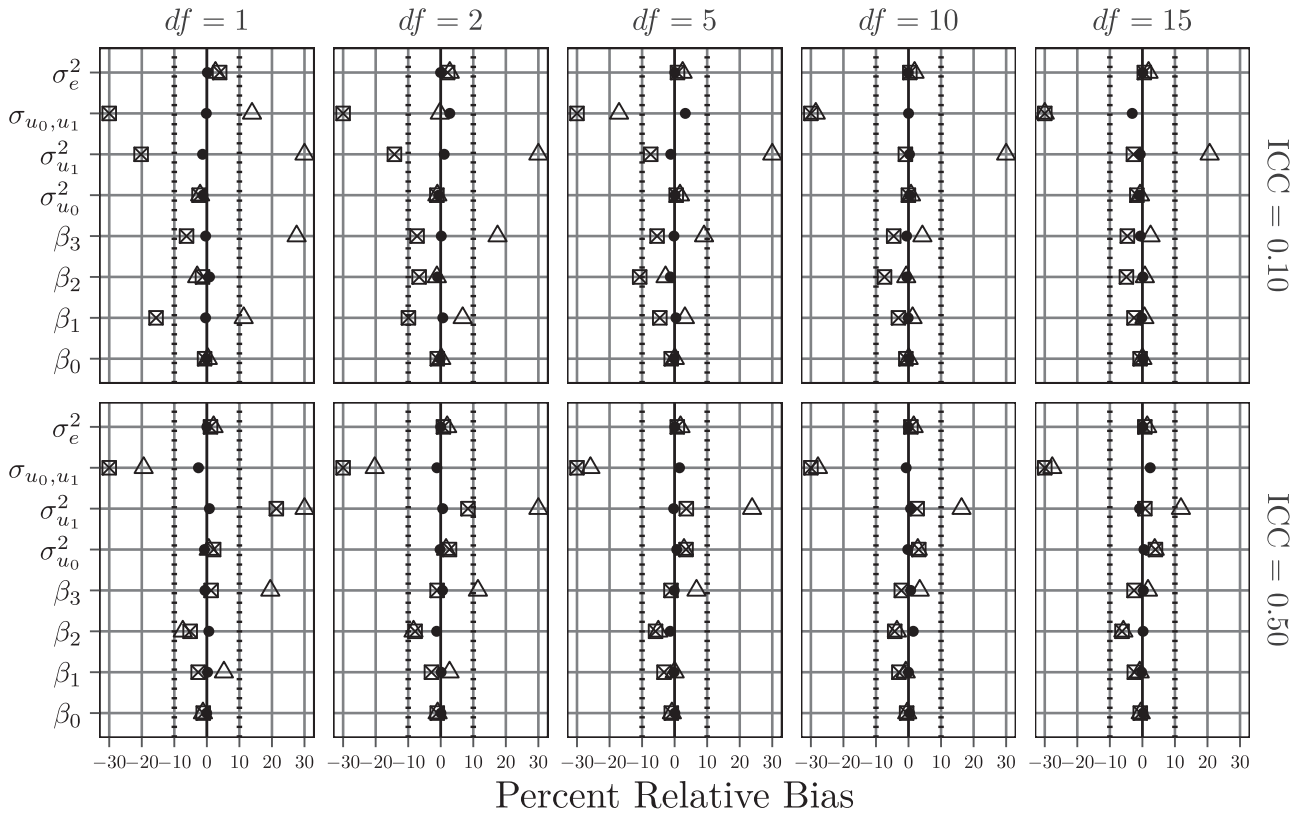

Figure 3. Bias for Bayesian estimation with varying degrees of non-normal level-1 residual.
Note: Trellis plots displaying percentage bias values from Bayesian estimation with a Yeo-Johnson transformation (☒) applied to a level-1 predictor with varying degrees of within-cluster non-normality. Complete-data estimates (●) from restricted maximum likelihood and Bayesian estimation without the transformation (△) are shown as a comparison. The within-cluster sample size is 15 and the number of level-2 clusters is 50. The separation prior was used for both methods. Values are truncated to ±30 bias. The sample sizes were chosen based on the conditions with observed bias.

The main new finding of this study is that within-cluster non-normality affected both the interaction coefficient and the random slope variance estimates; percentage bias values of 20% (in absolute value) were typical for the former, and the random slope bias values ranged between 80% to over 200%. Unlike Simulations 3 and 4, bias did not decrease as the sample size increased. To better understand the large random slope bias, we examined within-cluster regression of *Y* on *X* for several level-2 units. Because the missing data selection mechanism was defined such that higher Y values were more likely to have missing *X* observations, the bulk of missing values were in the sparsely populated upper tail of *X's* distribution. The scatterplots revealed that the deleted points had high leverage, flattening the within-cluster slopes. Replacing these values with less extreme imputations in the middle of the distribution (a consequence of the normality assumption) dramatically increased the within-cluster slopes and their variability. The online supplement (Section J) shows exemplar scatterplots that illustrate this point.

To investigate the impact of the missing data selection mechanism, we ran a follow-up simulation that reversed the sign of the relationship between *Y* and *X*. This operation created a missing at random selection mechanism that retained values in the upper tail of *X's* distribution and instead removed values from the large peak at the low end of the distribution. The "Sim 5 Reversed" rows of Tables 2 to 5 give the percentage bias from the follow-up simulation. Comparing the original data generation to the reversed relationship, one can see that the magnitude of the bias in both the cross-level interaction coefficient and random slope variance decreased drastically. In addition, the direction of the bias also changed from positive to negative. These results highlight that it is important to consider the functional form of the missing data mechanism when considering the impact of model misspecifications with asymmetrical distributions.

Simulation 5 is another example where the Yeo–Johnson transformation could potentially mitigate the negative impact of non-normal data. Following the procedure from Simulation 4, we again

performed follow-up simulations varied the shape of the within-cluster residual distribution using chi-square variables with degrees of freedom values equal to 2 (skewness ≈ 2, kurtosis ≈ 6), 5 (skewness ≈ 1.26, kurtosis ≈ 2.40), 10 (skewness ≈ .90, kurtosis ≈ 1.19), and 15 (skewness ≈ .73, kurtosis ≈ .79). For this simulation, we again fixed the level-1 and level-2 sample sizes at values that might be typical in applied practice, $n_j = 15$ and $J = 50$. To reiterate, these analyses are misspecified because the Yeo–Johnson density only approximates the true data-generating distribution. Again, our goal is to mimic a situation where a researcher would inappropriately apply this transformation without knowledge of the true population distribution.

Figure 3 shows trellis plots of Bayesian estimation with and without the Yeo–Johnson transformation (it is also possible to use multiple imputation with the transformation, but we exclude these results in the interest of space). The figure shows that the Yeo–-Johnson transformation effectively reduced bias, especially for the fixed effects, where percentage values were generally below 10%. The transformation also improved estimates of the random slope variance, which exhibited bias values of 10% or lower when the residual distribution was a chi-square with 2 degrees of freedom (skewness ≈ 2, kurtosis ≈ 6). Unsurprisingly, bias associated with the untransformed results decreased as the chi-square distribution's degrees of freedom (and thus non-normality) decreased. Considered as a whole, these results are encouraging given that Yeo–Johnson transformation was not the true data-generating function. We attribute the good performance of the Yeo–Johnson transformation in this simulation to the fact that the number of level-1 units is large, which improves the accuracy of the shape parameter.

Real data example

The data for the example are from an organizational study with $n_j = 6$ employees nested within $J = 105$ workgroups or teams. The analysis features a construct known as leader–member exchange (employee–supervisor relationship quality) as a within-team predictor of employee empowerment, the effect of which is moderated by team-level leadership climate (Aguinis et al., 2013; Chen et al., 2007). This interpretation is achieved by centering the predictor at iteratively-sampled latent group means. Although we did not investigate such a model in the simulations, we also include the leader–member exchange latent group means and their between-cluster interaction with the team-level moderator, and we use a self-reported biological sex dummy code (0 = female, 1 = male) and team-level cohesion rating as level-1 and level-2 covariates, respectively.

$$\begin{aligned} empower_{ij} = {} & \left(\beta_0 + u_{0j}\right) + \left(\beta_1 + u_{1j}\right)\left(lmx_{ij} - \mu_{LMXj}\right) \\ & + \beta_2\left(\mu_{LMXj} - \mu_{LMXj}\right) \\ & + \beta_3\left(climate_j - \mu_{CLIMATE}\right) \\ & + \beta_4\left(\mu_{LMXj} - \mu_{LMX}\right)\left(climate_j - \mu_{CLIMATE}\right) \\ & + \beta_5\left(lmx_{ij} - \mu_{LMXj}\right)\left(climate_j - \mu_{CLIMATE}\right) \\ & + \beta_6\left(male_{ij}\right) + \beta_7\left(cohesion_j\right) + \varepsilon_{ij} \end{aligned} \tag{20}$$

About 16.2% and 4.1% of the individual-level empowerment and leader–member exchange scores are missing, respectively, and 9.5% and 5.7% of the team-level climate and cohesion scores are missing.

A joint specification for the predictors is ideally suited for a Bayesian analysis because the latent group means (Hamaker & Muthén, 2020; Lüdtke et al., 2008) and grand means are explicit model parameters in the supporting predictor models. The factorization for this approach is as follows.

$$\begin{aligned} & f(EMPOWER, LMX, CLIMATE, MALE, COHESION) = \\ & f(EMPOWER \mid LMX, CLIMATE, MALE, COHESION) \\ & f(LMX, CLIMATE, MALE^*, COHESION) \end{aligned} \tag{21}$$

The asterisk superscript indicates that the sex dummy code appears as a latent response variable in the multivariate predictor distribution. Although treating a binary nominal predictor like biological sex as an ordinal latent response variable may seem odd, the multinomial probit model for multicategorical data is identical to the ordinal probit model in this case (Goldstein et al., 2009). An even simpler parameterization treats the incomplete predictors as multivariate normal, conditional on the complete biological sex code.

The strength of a sequential specification is that it accommodates some configurations of non-normal variables via the Yeo–Johnson transformation (Yeo & Johnson, 2000). We do not use it here because the latent group and grand means needed for centering are not byproducts of estimation. However, one could use the procedure to create non-normal multiple imputations and center at the manifest means of the imputed data. The most natural way to order variables in a sequential specification is by level, so the level-2 variables are factored last.

$$\begin{aligned} &f(EMPOWER, LMX, CLIMATE, MALE, COHESION) = \\ &f(EMPOWER \mid LMX, CLIMATE, MALE, COHESION) f(MALE^{*} \mid LMX, CLIMATE, COHESION) \times \\ &f(LMX \mid CLIMATE, COHESION) \times f(COHESION \mid CLIMATE) \times f(CLIMATE) \end{aligned} \tag{22}$$

In our experience, ordering predictors such that any transformed variables appear in one of the rightmost factorizations usually facilitates convergence, as does centering the non-normal variable at its median.

We fit the above model using the same three prior distributions from the simulations. For multiple imputation, we created $M = 100$ filled-in data sets by saving the imputations every 200 iterations across two separate chains. We set the burn-in periods conservatively to 20,000 after inspecting the potential scale reduction factor (Gelman & Rubin, 1992) diagnostic from preliminary runs; the inverse Wishart priors required a 5,000 iteration burn-in period, whereas the separation prior required closer to 10,000 initial cycles. We used Blimp 3 (Keller & Enders, 2021) to implement Bayesian estimation and model-based multiple imputation. After creating the multiple imputations, we centered the leader–member exchange scores at the imputed latent group means and centered all other variables at the grand means of the filled-in data sets. Finally, we used the R package *mitml* (Grund et al., 2021) to fit the random coefficient model to each data set and pool the resulting estimates and standard errors. The raw data and analysis scripts for the joint and sequential specifications are available on GitHub (*https://github.com/blimp-stats/FACTORED-REGRESSION-WITH-CROSS-LEVEL-INTERACTIONS*). In addition, Section D of the online supplemental shows a Blimp script that implements the separation prior. The script also generates multiple imputations for a frequentist analysis along with estimates of the latent group means needed for centering. Interested readers can consult Keller and Enders (2022) for additional details about specific commands.

Table 6. Bayesian and multiple imputation results from the real data analysis example.

		Bayesian Prior Specification							
		Wishart 1		Wishart 2		Separation		Imputation	
Parameter		Med.	*SD*	Med.	*SD*	Med.	*SD*	Est.	*SE*
Intercept	β_0	26.78	0.84	26.73	0.87	26.77	0.83	26.74	0.86
LMX_w	β_1	0.66	0.08	0.66	0.08	0.66	0.08	0.66	0.08
LMX_b	β_2	0.11	0.34	0.15	0.31	0.12	0.32	0.08	0.34
Climate	β_3	0.20	0.06	0.20	0.06	0.20	0.06	0.20	0.05
$LMX_w \times$ Climate	β_4	0.04	0.02	0.04	0.02	0.04	0.02	0.04	0.02
$LMX_b \times$ Climate	β_5	–0.05	0.08	–0.03	0.07	–0.05	0.08	–0.06	0.08
Male code	β_6	1.70	0.33	1.70	0.33	1.72	0.33	1.69	0.33
Cohesion	β_7	0.21	0.16	0.20	0.17	0.21	0.16	0.22	0.17
Intercept Var.	$\sigma^2_{u_0}$	0.39	0.30	0.81	0.54	0.28	0.24	0.42	—
Covariance	$\sigma_{u_0 u_1}$	–0.04	0.10	–0.07	0.15	0.002	0.06	–0.06	—
Slope Var.	$\sigma^2_{u_1}$	0.20	0.08	0.19	0.09	0.22	0.07	0.19	—
Residual Var.	σ^2_e	11.96	0.86	12.01	0.90	12.07	0.88	12.05	—
	Rights and Sterba (2019) R^2 Effect Sizes								
Fixed effects	$\boldsymbol{\beta}$	.310	.038	.304	.039	.301	.039	.305	—
Intercept Var.	$\sigma^2_{u_0}$	.019	.014	.038	.025	.013	.011	.021	—
Slope Var.	$\sigma^2_{u_1}$	.081	.028	.073	.033	.097	.031	.076	—

Note: Bayes estimates were based on the posterior median (Med.) and posterior standard deviation (*SD*). Imputation estimates (Est.) and standard errors (*SE*) were based on 100 imputations using the separation specification.

Table 6 gives the posterior summaries for the three Bayesian analyses and factored regression multiple imputation results from the separation prior. The first two columns are inverse Wishart priors: the prior in the first column should be more informative because it adds to the data's degrees of freedom, whereas the prior in the second column should be less informative because it subtracts from it. Consistent with the simulation results, the choice of prior distribution had a noticeable impact on intercept and slope variances; differences in the intercept variance were between approximately 0.35 and 1.35 posterior standard deviation units, and differences in the slope variance were between 0.26 and 0.78 standard deviation units. It is important to highlight that this example most aligns with the ICC = .10 and $n_j = 5$ conditions, a situation where none of the priors and estimators were optimal.

With a larger number of level-1 units, our simulation results would predict that the separation prior should produce superior estimates of the intercept and slope variance, but our results say much less about which method is "best" with this particular

configuration. From a practical perspective, a researcher might approach this uncertainty by presenting a sensitivity analysis that considers the stability of the model's conclusions across different prior distributions. To facilitate such a comparison, Table 6 also reports Rights and Sterba (2019) effect sizes for the fixed effects, random intercepts, and random slopes; Blimp automatically reports these variance explained statistics with a Wishart prior, and the code in Section D of the online supplemental illustrates how to obtain these quantities with the separation prior. As one might expect, the variance explained by the fixed effects was virtually unaffected by the choice of prior, but the R^2 statistics for the random effects were more variable; as a proportion of the total model-implied outcome variance, the random intercept variance estimates ranged from $R^2 = .013$ to .038, and the random slope effect sizes ranged from $R^2 = .031$ to .081. Gauging whether the effect size differences are meaningfully is somewhat subjective, but we believe these quantities convey the impact of the prior distributions on a more practical metric than the raw variance components. Online supplemental documents are an ideal vehicle for reporting alternate sets of results like those in Table 6.

Discussion

A large body of research supports the use of multiple imputation as a method to handle missing data. Despite being a mature methodology, methodologists have only investigated and fine-tuned multiple imputation's ability to handle interactions and other nonlinear terms in the last decade. This is usually achieved with some type of factored regression specification that factorizes the joint distribution of the analysis variables into a primary part representing the analysis model of interest and a secondary part representing the distributions of the incomplete predictors. A growing body of recent research supports these methods (Bartlett et al., 2015; Enders et al., 2020; Erler et al., 2019; Grund et al., 2021; Lüdtke et al., 2020a; Lüdtke et al., 2020b; Zhang & Wang, 2017), and our study adds to this growing knowledge base.

The first two simulations examined ideal circumstances where the model used to generate the data matched the true data-generating model. A major contribution of these studies is the focus on Bayesian inference with a novel separation strategy that imposes unique priors on the random effect variances and their correlation. This approach parameterizes the random intercepts and random slopes as correlated phantom latent variables following a specification described by Merkle and Rosseel (2018, p. 8), but it is conceptually equivalent to other separation priors suggested in the literature (Barnard et al., 2000; Liu et al., 2016). Overall, our simulations generally supported the separation prior in the multilevel context, as the strategy reduced the bias in the variance components relative to inverse Wishart priors. An apparent downside of the separation prior is that it systematically underestimates the correlation between the random intercepts and slopes, although the mean squared error of this parameter was substantially smaller than that of multiple imputation (i.e., Bayesian estimates were closer to the true value, on average, despite their bias).

Another major contribution of the paper was the application of the Yeo–Johnson transformation to non-normal predictors. Both Simulations 3 and 4 investigated forms of between-cluster non-normality. Simulation 3 considered the case where the level-1 predictor's latent group means were non-normally distributed, and Simulation 4 examined a non-normal level-2 predictor. In Simulation 3, non-normal latent means affected the recovery of the interaction slope coefficient, whereas the non-normal moderator in Simulation 4 primarily impacted the variable's lower-order slope. We also conducted a follow-up to Simulation 4 that considered the Yeo–Johnson transformation, but the procedure was ineffective at decreasing bias. We attribute this finding to the fact that the level-2 sample size was very small ($J = 25$ clusters), making it difficult to estimate the shape parameter accurately. It is also important to note that the transformation was misspecified, as it did not match the true data-generating function. Future studies should consider continuous non-normal variables generated via a Yeo–Johnson transformation.

Finally, Simulation 5 investigated a misspecification where the level-1 predictor's within-cluster residuals were non-normal. This type of non-normality had a rather catastrophic impact on the slope variance estimates, which were substantially overestimated. However, follow-up simulations revealed that this bias was largely due to the missing data process, which culled values from the sparse upper tail of the within-cluster distribution, as removing missing values from the distribution's dense lower tail dramatically reduced bias. This simulation also investigated the Yeo–Johnson transformation and found that it was generally effective, even though it only approximated the true data-generating function. This finding is important because it suggests that the procedure may

hold promise for applied practice where the population distribution functions are virtually never known.

Considered as a whole, our simulations suggest that there is no reason to prefer Bayesian estimation to multiple imputation (or vice versa) because the two procedures are effectively equivalent (setting aside their important philosophical differences). That said, there may be practical reasons to choose one over the other. For example, Bayesian inference is arguably simpler because estimation happens in one stage rather than two, as required by multiple imputation. This is an important practical advantage that positions Bayesian estimation as a potential replacement for maximum likelihood estimation, another single-stage method that is not yet well-suited for these types of analysis problems. Although Bayesian estimation readily accommodates auxiliary variables as additional endogenous variables in the factored regression specification (Keller & Enders, 2022, Chapter 4), one might instead prefer multiple imputation because auxiliary variables can simply be added as additional covariates in the imputation model and ignored in the subsequent analysis phase (Lüdtke et al., 2020b). Frequentist analyses may also be more normative in some disciplines.

The simulation studies had features that potentially limit their generalizability. First, we only investigated an analysis model with one predictor at level-1, one predictor at level-2, and a cross-level interaction between the predictors. The basic premise of our study could be expanded to other multilevel interaction effects (e.g., a within-cluster interaction involving a pair of level-1 variables; a between-cluster interaction involving the group means and the level-2 predictor; Preacher et al., 2010), nonlinear polynomials (e.g., quadratic models), or even combinations of both. Second, we focused on an analysis model where the dependent variable was complete and normally distributed. Future simulations could investigate categorical outcomes, as the factored regression procedure readily accommodates binary, ordinal, multicategorical, and even count outcomes. Third, future studies could consider different forms of misspecifications (e.g., nonlinear relationships among predictors) and different types of missing data mechanisms. Simulation 5 underscores the importance of this future direction, as the deletion process we used produced a worse-case scenario that is virtually impossible to model correctly. Finally, future studies are needed that thoroughly investigate the application of the Yeo–Johnson transformation to multilevel models. Our study provided an initial foray into this topic, but more research is needed.

In conclusion, the simulation results illustrate the flexibility and limitations of the factored regression specification for multilevel data with cross-level interaction effects. The five simulations illustrated instances where the method was largely robust to model misspecifications and revealed situations where it was vulnerable. At its worst, factored regression was still far superior to ad hoc methods such as just-another-variable imputation (see online supplemental material). Consistent with the growing body of literature, this study offers broad support for factored regression models, but additional research is needed to understand its behavior, especially with non-normal continuous variables and the promising Yeo–Johnson transformation.

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 R305D190002 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 authors would like to thank the anonymous reviewers and the associate editor Sarah Depaoli for their comments on prior versions of this manuscript. 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.) Wiley.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528. https://doi.org/10.1177/0149206313478188

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.2307/2290350

Allison, P. D. (2002). *Missing data*. Sage.

Asparouhov, T., & Muthén, B. (2021a). Bayesian estimation of single and multilevel models with latent variable interactions. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 314–328. https://doi.org/10.1080/10705511.2020.1761808

Asparouhov, T., & Muthén, B. (2021b). Expanding the Bayesian structural equation, multilevel and mixture models to logit, negative-binomial and nominal variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 622–637. https://doi.org/10.1080/10705511.2021.1878896

Asparouhov, T., Muthén, B. (2010a). Bayesian analysis using Mplus: Technical implementation. https://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., Muthén, B. (2010b). Multiple imputation with Mplus. https://www.statmodel.com/download/Imputations7.pdf

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*, 1281–1311.

Bartlett, J., Keogh, R., Bonneville, C. T. (2021). Package 'smcfcs'. https://cran.r-project.org/web/packages/smcfcs/index.html

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R, Alzheimer's Disease Neuroimaging Initiative* (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N. (2021). Package 'lme4'. https://cran.r-project.org/web/packages/lme4/index.html

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *26*(2), 211–243.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Browne, W. J. (1998). [*Applying MCMC methods to multilevel models*]. [Doctoral dissertation]. University of Bath).

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. https://doi.org/10.18637/jss.v045.i05

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics (Oxford, England)*, *2*(4), 485–500. https://doi.org/10.1093/biostatistics/2.4.485

Chen, G., Kirkman, B. L., Kanfer, R., Allen, D., & Rosen, B. (2007). A multilevel study of leadership, empowerment, and performance in teams. *The Journal of Applied Psychology*, *92*(2), 331–346.

Darnieder, W. F. (2011). [*Bayesian methods for data-dependent priors*]. [Doctoral dissertation]. The Ohio State University).

Draper, D. (2008). Bayesian multilevel analysis and MCMC. In *Handbook of multilevel analysis*. (pp. 77–139) Springer.

Du, H., Enders, C., Keller, B. T., Bradbury, T. N., & Karney, B. R. (2022). A Bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research*, *57*(2-3), 478–512.

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*(1), 39–55. https://doi.org/10.1037/a0035314

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and other nonlinear terms. *Psychological Methods*, *25*(1), 88–112. https://doi.org/10.1037/met0000228

Enders, C. K., Hayes, T., & Du, H. (2018). A comparison of multilevel imputation schemes for random coefficient models: Fully conditional specification and joint model imputation with random covariance matrices. *Multivariate Behavioral Research*, *53*(5), 695–713. https://doi.org/10.1080/00273171.2018.1477040

Enders, C. K., & Keller, B. T. (2019). Blimp technical appendix: Centering covariates in a bayesian multilevel analysis.

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, *23*(2), 298–317. https://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. https://doi.org/10.1037/met0000063

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, *28*(2), 555–568. https://doi.org/10.1177/0962280217730851

Erler, N. S., Rizopoulos, D., Rosmalen, J., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (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. https://doi.org/10.1002/sim.6944

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-Ba117a

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. (3rd ed.) CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*(3), 173–197. https://doi.org/10.1177/1471082x0800900301

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. https://doi.org/10.1111/rssa.12022

Gomes, M., Radice, R., Camarena Brenes, J., & Marra, G. (2019). Copula selection models for nonGaussian outcomes that are missing not at random. *Statistics in Medicine*, *38*(3), 480–496. https://doi.org/10.1002/sim.7988

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: The Official Journal of the Society for Prevention Research*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: A review. *Quality & Quantity*, *49*(3), 967–976. https://doi.org/10.1007/s11135-014-0060-5

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Press.

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. https://doi.org/10.1177/1094428117703686

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, *53*(6), 2631–2649.

Grund, S., Robitzsch, A., Lüdke, O. (2021). Package 'mitml'. https://cran.r-project.org/web/packages/mitml/index.html

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: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*(9), 876–883. https://academic.oup.com/aje/article-abstract/149/9/876/113681

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.2307/2334940

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. https://doi.org/10.3102/0162373707299706

Hollenbach, F. M., Bojinov, I., Minhas, S., Metternich, N. W., Ward, M. D., & Volfovsky, A. (2021). Multiple imputation using Gaussian copulas. *Sociological Methods & Research*, *50*(3), 1259–1283. https://doi.org/10.1177/0049124118799381

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765–769. https://doi.org/10.1080/01621459.1990.10474938

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. https://doi.org/10.2307/3315865

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, *51*(10), 5142–5154. https://doi.org/10.1016/j.csda.2006.05.021

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. Springer.

Käärik, E. (2006). Imputation algorithm using copulas. *Advances in Methodology and Statistics*, *3*(1), 109–120–109–120. https://doi.org/10.51936/qqac6077

Käärik, E., & Käärik, M. (2009). Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference*, *139*(11), 3830–3835. https://doi.org/10.1016/j.jspi.2009.05.020

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. https://doi.org/10.1207/s15327906mbr2301_4

Keller, B. T., & Du, H. (2019). A Fully Conditional Specification Approach to Multilevel Multiple Imputation with Latent Cluster Means. *Multivariate Behavioral Research*, *54*(1), 149–150. https://doi.org/10.1080/00273171.2018.1556085

Keller, B. T., Enders, C. K. (2022). Blimp user's guide (Version 3). Retrieved from www.appliedmissingdata.com/blimp

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Lipsitz, S. R., & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, *83*(4), 916–922. https://doi.org/10.1093/biomet/83.4.916

Liu, H. Y., Zhang, Z. Y., & Grimm, K. J. (2016). 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*, *23*(3), 354–367. https://doi.org/10.1080/10705511.2015.1057285

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel

latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Lüdtke, O., Robitzsch, A., & West, S. G. (2020a). Analysis of interactions and nonlinear effects with missing data: A factored regression modeling approach using maximum likelihood estimation. *Multivariate Behavioral Research*, *55*(3), 361–381. https://doi.org/10.1080/00273171.2019.1640104

Lüdtke, O., Robitzsch, A., & West, S. G. (2020b). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, *25*(2), 157–181. https://doi.org/10.1037/met0000233

Lunn, D., Jackson, C., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book*. CRC Press.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer.

Maas, C. J., & Hox, J. J. (2004a). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Maas, C. J., & Hox, J. J. (2004b). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, *26*(3), 388–402. https://doi.org/10.1214/11-STS361

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250x.1975.9989847

McNeish, D. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

McNeish, D. (2016b). 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*(1), 27–56. https://doi.org/10.3102/1076998615621299

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Murray, D. M., & Blistein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, *27*(1), 79–103. https://doi.org/10.1177/0193841X02239019

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

Quartagno, M., Carpenter, J. (2020). Package 'jomo'. https://cran.r-project.org/web/packages/jomo/index.html

R Core Team (2020). *A language and environment for statistical computing*. In R Foundation for Statistical Computing.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

Robbins, M. W., Ghosh, S. K., & Habiger, J. D. (2013). Imputation in high-dimensional economic data as applied to the Agricultural Resource Management Survey. *Journal of the American Statistical Association*, *108*(501), 81–95. https://doi.org/10.1080/01621459.2012.734158

Robitzsch, A., Lüdke, O. (2021). Package 'mdmb'. https://cran.r-project.org/web/packages/mdmb/index.html

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

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

Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, *58*(4), 298–302. https://doi.org/10.1198/000313004X6355

Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change*. (pp. 355377) American Psychological Association.

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. https://doi.org/10.1198/106186002760180608

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixedeffects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, *12*, 46–13. https://doi.org/10.1186/1471-2288-12-46

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.) Sage.

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) Routledge.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217–221. https://doi.org/10.1080/01621459.1996.10476679

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *366*(1874), 2389–2403. https://doi.org/10.1098/rsta.2008.0038

Yucel, R. M. (2011). Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, *11*(4), 351–370. https://doi.org/10.1177/1471082X1001100404

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, *22*(4), 649–666. https://doi.org/10.1037/met0000104