Psychological Methods

© 2017 American Psychological Association
1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000148

A Fully Conditional Specification Approach to Multilevel Imputation of Categorical and Continuous Variables

Craig K. Enders and Brian T. Keller
University of California, Los Angeles

Roy Levy
Arizona State University

Abstract

Specialized imputation routines for multilevel data are widely available in software packages, but these methods are generally not equipped to handle a wide range of complexities that are typical of behavioral science data. In particular, existing imputation schemes differ in their ability to handle random slopes, categorical variables, differential relations at Level-1 and Level-2, and incomplete Level-2 variables. Given the limitations of existing imputation tools, the purpose of this manuscript is to describe a flexible imputation approach that can accommodate a diverse set of 2-level analysis problems that includes any of the aforementioned features. The procedure employs a fully conditional specification (also known as chained equations) approach with a latent variable formulation for handling incomplete categorical variables. Computer simulations suggest that the proposed procedure works quite well, with trivial biases in most cases. We provide a software program that implements the imputation strategy, and we use an artificial data set to illustrate its use.

Translational Abstract

Multiple imputation is a missing data handling technique that creates several copies of the incomplete data, each with different estimates of the missing values. The researcher analyzes each data set, and the resulting estimates and standard errors are averaged into a single set of results. The primary goal of this article was to outline a novel multiple imputation approach to multilevel data sets and examine its performance. Multilevel data are exceedingly common throughout psychology and the behavioral sciences. Examples of such nested data structures include children within classrooms, individuals within families, employees within workgroups, and repeated measurements within individuals, to name a few. Current approaches to handling multilevel missing data have limitations, and our approach addresses practical problems that are common in applied research (e.g., incomplete categorical variables, complex model structures). The study used computer simulation to create many artificial data sets with missing values, after which it imputed each data set and examined the accuracy of the resulting estimates. The computer simulation results indicated that the proposed procedure works quite well, with trivial biases in most cases. We provide a software program for MacOS and Windows that implements the imputation strategy, and the paper illustrates its use.

Keywords: missing data, multiple imputation, multilevel models, imputation software

Supplemental materials: http://dx.doi.org/10.1037/met0000148.supp

A rather large body of methodological literature supports the use of missing data handling methods that assume a missing at random (MAR) mechanism, whereby the probability of missing data on a particular variable is fully determined by the observed values of other variables (Little & Rubin, 2002; Rubin, 1976). The multiple imputation procedure proposed by Rubin (1987) is an MAR-based approach that enjoys widespread use in a variety of disciplines, including the behavioral sciences. To implement multiple imputation, a researcher first creates several copies of the incomplete data set, filling in each with a different set of plausible replacement values. The complete data sets are then analyzed, and the resulting parameter estimates and standard errors are pooled into a single set of results. Multiple imputation is preferable to older approaches such as deletion because it can reduce nonresponse bias and improve power. Detailed descriptions of multiple imputation are readily available in the methods literature (Enders, 2010; Graham, 2012; Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001; van Buuren, 2012).

Joint modeling and fully conditional specification (FCS; also known as sequential regression and chained equations imputation) are the principal imputation frameworks for single-level data.

Craig K. Enders and Brian T. Keller, Department of Psychology, University of California, Los Angeles; Roy Levy, School of Social and Family Dynamics, Arizona State University.

This work has not been presented in other outlets. This work was supported by Institute of Educational Sciences award R305D150056.

Correspondence concerning this article should be addressed to Craig K. Enders, Department of Psychology, University of California, Los Angeles, Box 951963, Los Angeles, CA 90095-1563. E-mail: cenders@psych.ucla.edu

Schafer's (1997) classic text popularized the joint modeling strategy that assumes a common distribution for the incomplete variables. In the context of normally distributed data, Schafer's approach repeatedly samples plausible population parameters (typically a covariance matrix and a mean vector) from a probability distribution and uses those parameters to define a multivariate normal distribution, from which it draws replacement data values. FCS uses a similar two-step algorithmic approach (sample parameter values, use the parameters to define a distribution of replacement values), but it draws imputations from a series of univariate conditional distributions (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; van Buuren, 2007, 2012; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Under this scheme, variables are imputed one at a time, with the filled-in variable from one step serving as a predictor in all subsequent imputation steps.

Methodologists have extended the joint model and FCS to multilevel data (Asparouhov & Muthén, 2010; Carpenter, Goldstein, & Kenward, 2011; Goldstein, Bonnet, & Rocher, 2007; Goldstein, Carpenter, Kenward, & Levin, 2009; Schafer, 2001; Schafer & Yucel, 2002; van Buuren, 2011, 2012; Yucel, 2008), and specialized imputation routines are widely available in software packages. For example, the joint model framework is implemented in M*plus* (Muthén & Muthén, 1998–2012), the PAN and MLMMM packages in R (Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008), MLwiN and Stata (Carpenter et al., 2011), and SAS (Mistler, 2013), and FCS is available in the R package MICE (van Buuren et al., 2014). The joint model is equivalent to FCS with single-level data and multivariate normal variables (Hughes et al., 2014), but multilevel imputation routines apply different underlying models, and software packages offer different functionality (Enders, Mistler, & Keller, 2016). Simulation and analytic work suggest that the joint model and FCS can readily accommodate basic random intercept analyses with normally distributed variables, but they differ beyond that (Carpenter & Kenward, 2013; Enders et al., 2016; Mistler & Enders, 2016).

Enders, Mistler, and Keller (2016) conclude that existing multilevel imputation routines are good for very specific tasks, but these methods are generally not equipped to handle a wide range of complexities that are typical of behavioral science data. In particular, the joint model and FCS differ in their ability to handle random slopes, categorical variables, differential relations at Level-1 and Level-2 (e.g., contextual effect models, multilevel structural equation models), and incomplete Level-2 variables. Given the limitations of existing imputation tools, our primary goal for this paper is to outline a flexible FCS imputation strategy that can accommodate a diverse set of two-level analysis problems that includes any (or all) of the aforementioned features. We provide an application named Blimp for Mac and Windows that implements our FCS approach, and we use computer simulations to evaluate its performance.

The organization of the article is as follows. We begin with a brief description of complete-data Bayesian estimation for a two-level model, as this provides the mathematical machinery for FCS imputation. Second, we review how FCS is currently applied to multilevel data. Third, we describe complete-data Bayesian estimation for a two-level probit model, as this provides the basis for imputing nominal and ordinal variables. Fourth, we outline an extension to FCS that accommodates incomplete nominal and ordinal variables at Level-1 and Level-2. Fifth, we outline a modification to FCS that partitions relations among Level-1 variables into within- and between-cluster components. Sixth, we propose an extension to FCS that can accommodate missing data on Level-2 variables. Finally, we use computer simulations to evaluate the modifications to FCS, and we conclude with a data analysis example that demonstrates our custom-built FCS software application.

Bayesian Estimation for a Two-Level Regression Model

Like other multilevel imputation schemes (Asparouhov & Muthén, 2010; Goldstein et al., 2007, 2009; Schafer & Yucel, 2002), FCS borrows from established complete-data Bayesian estimation methods for multilevel regression models. Multilevel imputation via FCS can be viewed as a complete-data Bayesian analysis with an additional step that fills in the data, conditional on the model parameters and Level-2 residuals from a particular iteration. To provide some necessary background, this section gives a brief overview of the Gibbs sampling algorithm for a Bayesian analysis. We focus on a traditional multilevel model for univariate normal data because it provides the mathematical machinery for FCS imputation. However, it is important to emphasize that FCS is not limited to traditional multilevel regression models, as the resulting imputations are applicable to other multilevel frameworks with nested factors (e.g., multilevel structural equation models).

To motivate the ensuing discussion, consider a multilevel model with two Level-1 predictors and a random slope.[1] Using notation from Scott, Shrout, and Weinberg (2013), the model is

$$Y_{1ij} = \beta_0 + \beta_1 Y_{2ij} + \beta_2 Y_{3ij} + u_{0j} + u_{1j} Y_{2ij} + \varepsilon_{ij} \quad (1)$$

where Y_{1ij} is the outcome score for observation i in cluster j, Y_{2ij} and Y_{3ij} are Level-1 predictors, β_0 is the intercept, and β_1 and β_2 are slope coefficients for Y_2 and Y_3, respectively. Turning to the random effects, u_{0j} is a residual that captures between-cluster residual variation (i.e., mean differences) in the outcome, and u_{1j} is a random slope residual that allows the influence of Y_2 to vary across clusters. Finally, ε_{ij} is a within-cluster residual that captures unexplained Level-1 variation. In line with a traditional multilevel analysis, we assume that Level-2 residuals are multivariate normal with zero means and an unstructured covariance matrix Σ_u, and we assume that Level-1 residuals are normally distributed with a constant variance σ^2_ε. This latter assumption can be relaxed, as the Bayesian framework readily accommodates heteroscedastic within-cluster residual variation (Kasim & Raudenbush, 1998; van Buuren, 2011).

A Bayesian analysis views regression coefficients, Level-2 residuals, and variance parameters as random variables, and a joint distribution describes the relative probability of different combinations of parameter values and Level-2 residual terms, given the

[1] Multilevel models are also known in the literature as mixed effects and random effects models. We use the phrase multilevel model to emphasize that our imputation routine is designed for data structures with nested factors. Not all mixed models that incorporate random effects feature this multilevel nesting structure, and these examples fall outside of the scope of our work.

data. Bayesian estimation expresses this joint distribution as a set of full conditional distributions, and a Gibbs sampler algorithm iteratively samples values from each distribution. The joint distribution for the model in Equation (1) requires four such conditional distributions, one each for the regression coefficients, the Level-2 residuals, the within-cluster residual variance, and the Level-2 covariance matrix. Accordingly, the Gibbs algorithm samples these quantities in a series of four steps, with each step conditioning on (i.e., treating as known) the values from previous steps: (a) sample regression coefficients from a distribution that conditions on the data, the current variance estimates, and the current Level-2 residuals, (b) sample Level-2 residuals from a distribution that conditions on the data, the coefficients from the previous step, and the current variance estimates, (c) sample a Level-1 residual variance from a distribution that conditions on the data, the current Level-2 covariance matrix, and the coefficients and residuals from previous two steps, and (d) sample a Level-2 covariance matrix from a distribution that conditions on the data and the values from the first three steps.

More formally, the sampling steps for a single iteration t of the Gibbs algorithm are

$$\begin{aligned} \boldsymbol{\beta}^{(t)} &\sim \mathrm{MVN}\left(\boldsymbol{\beta} \mid \mathbf{Y},\ \mathbf{u}^{(t-1)},\ \sigma_{\varepsilon}^{2(t-1)},\ \boldsymbol{\Sigma}_{u}^{(t-1)}\right) \\ \mathbf{u}^{(t)} &\sim \mathrm{MVN}\left(\mathbf{u} \mid \mathbf{Y},\ \boldsymbol{\beta}^{(t)},\ \sigma_{\varepsilon}^{2(t-1)},\ \boldsymbol{\Sigma}_{u}^{(t-1)}\right) \\ \sigma_{\varepsilon}^{2(t)} &\sim \mathrm{IG}\left(\boldsymbol{\sigma}_{\varepsilon}^{2} \mid \mathbf{Y},\ \boldsymbol{\beta}^{(t)},\ \mathbf{u}^{(t)},\ \boldsymbol{\Sigma}_{u}^{(t-1)}\right) \\ \boldsymbol{\Sigma}_{u}^{(t)} &\sim \mathrm{IW}\left(\Sigma_{u} \mid \mathbf{Y},\ \boldsymbol{\beta}^{(t)},\ \mathbf{u}^{(t)},\ \sigma_{\varepsilon}^{2(t)}\right) \end{aligned} \tag{2}$$

where ~MVN denotes a multivariate normal distribution, ~IG is the inverse Gamma distribution, and ~IW indicates the inverse Wishart distribution. Each of the above distributions has a location and scale parameter that defines its expected value and variance, and these quantities also depend on hyperparameters (i.e., the expected value and variance) of a corresponding prior distribution. The supplemental online material includes a technical document that gives specific details for each distribution, as do a number of published resources (Browne & Draper, 2000; Goldstein et al., 2007, 2009; Kasim & Raudenbush, 1998; Schafer & Yucel, 2002; van Buuren, 2012; Yucel, 2008).

Iterating the sampling steps from Equation (2) many (e.g., several thousand) times gives an empirical estimate of each parameter's marginal posterior distribution, the mean and standard deviation of which are analogous to a frequentist point estimate and standard error, respectively. In the context of FCS imputation, the previous sampling steps are unchanged, but each iteration features an additional fifth step that generates imputations based on the current model parameters and Level-2 residual terms. Thus, each Gibbs cycle uses the current imputations to execute a complete-data Bayesian analysis, after which it uses the resulting parameter values to generate a new set of imputations. The next section details this procedure.

FCS Imputation for Two-Level Data

This section describes the current implementation of multilevel FCS. To be consistent with existing literature and software (van Buuren, 2011, 2012; van Buuren et al., 2014), we restrict our attention to normally distributed Level-1 variables, but subsequent sections outline modifications to FCS that extend its current capabilities. For brevity, we focus on imputation here and refer interested readers to other resources that describe analysis and pooling procedures for multiply imputed data (Enders, 2010; Rubin, 1987; Schafer, 1997; Schafer & Olsen, 1998; Sinharay et al., 2001; van Buuren, 2012).

Multilevel FCS imputes variables one at a time, drawing replacement values from a series of univariate distributions that condition on a set of multilevel model parameters, Level-2 residual terms, and complete and previously imputed variables. To illustrate, consider a set of Q incomplete Level-1 variables, indexed $q = 1, \ldots, Q$. The imputation steps from a single iteration t of FCS can be summarized symbolically as follows

$$\begin{aligned} &Y_{1(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(Y_{1(\mathrm{mis})} \mid Y_{2}^{(t-1)}, \ldots, Y_{Q}^{(t-1)},\ \mathbf{X},\ \boldsymbol{\theta}_{(1)}^{(t)},\ \mathbf{u}_{(1)}^{(t)}\right) \\ &Y_{2(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(Y_{2(\mathrm{mis})} \mid Y_{1}^{(t)},\ Y_{3}^{(t-1)}, \ldots, Y_{Q}^{(t-1)},\ \mathbf{X},\ \boldsymbol{\theta}_{(2)}^{(t)},\ \mathbf{u}_{(2)}^{(t)}\right) \\ &Y_{q(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(Y_{q(\mathrm{mis})} \mid Y_{1}^{(t)}, \ldots, Y_{q-1}^{(t)},\ Y_{q+1}^{(t-1)}, \ldots, Y_{Q}^{(t-1)},\ \mathbf{X},\ \boldsymbol{\theta}_{(q)}^{(t)},\ \mathbf{u}_{(q)}^{(t)}\right) \\ &Y_{Q(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(Y_{Q(\mathrm{mis})} \mid Y_{1}^{(t)}, \ldots, Y_{Q-1}^{(t)},\ \mathbf{X},\ \boldsymbol{\theta}_{(Q)}^{(t)},\ \mathbf{u}_{(Q)}^{(t)}\right) \end{aligned} \tag{3}$$

where $Y_{q(\mathrm{mis})}^{(t)}$ is the variable to be imputed at step q of iteration t, $Y_{q}^{(\cdot)}$ is a completed version of this variable that contains the current imputations, ~N denotes a univariate normal distribution, $\mathbf{X}$ is a set of complete variables (Level-1 or Level-2), $\boldsymbol{\theta}_{(q)}^{(t)}$ represents the current set of multilevel model parameters for incomplete variable q (i.e., $\boldsymbol{\theta}_{(q)}^{(t)} = \{\boldsymbol{\beta}_{(q)}^{(t)},\ \Sigma_{u(q)}^{(t)},\ \sigma_{\varepsilon(q)}^{2(t)}\}$), and $\mathbf{u}_{(q)}^{(t)}$ denotes the current values of the Level-2 residuals for that variable. In words, the equation says to draw missing values from a normal distribution, the mean and variance of which depend on previously imputed and complete variables, multilevel model parameters, and Level-2 residual terms. As noted previously, each step of Equation (3) can be viewed as a sequence that performs a complete-data Bayesian analysis with $Y_{q}^{(t-1)}$ as the outcome followed by an imputation step that uses $\boldsymbol{\theta}_{(q)}^{(t)}$ and $\mathbf{u}_{(q)}^{(t)}$ to generate updated imputations for $Y_{q}^{(t)}$.

To illustrate multilevel FCS, reconsider the random slope analysis from Equation (1), and assume that all variables are incomplete. This analysis is useful because it highlights that FCS can tailor the composition of each imputation step to accommodate the specific features of a particular analysis model (e.g., some variables require a random slope, others require only random intercepts). However, it is important to reiterate that FCS is not limited to univariate regression models, as the resulting imputations are applicable to other multilevel frameworks with nested factors (e.g., multilevel structural equation models).

To begin, FCS applies the Bayesian estimation steps from Equation (2) to the filled-in data from the previous iteration, treating Y_1 as an outcome and Y_2 and Y_3 as predictors. The resulting parameter values and residual terms define a normal distribution that generates Y_1 imputations

$$\begin{aligned} &Y_{1ij(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(\hat{Y}_{1ij}^{(t)},\ \ \sigma_{\varepsilon(Y_1)}^{2}\right) \\ &\hat{Y}_{1ij}^{(t)} = \beta_{0(Y_1)} + \beta_{1(Y_1)} Y_{2ij}^{(t-1)} + \beta_{2(Y_1)}\, Y_{3ij}^{(t-1)} + u_{0(Y_1)} +\ u_{1(Y_1)} Y_{2ij}^{(t-1)} \end{aligned} \tag{4}$$

where $\hat{Y}_{1ij}^{(t)}$ is a predicted value from the multilevel model, and $\sigma_{\varepsilon(Y_1)}^{2}$ is the within-cluster residual variance. To simplify the notation, we omit the iteration superscript on the parameters and residual terms because these quantities are drawn at iteration t prior to imputation. Further, we include the incomplete variable's name in the subscripts on the right side of the equations to emphasize that each imputation model requires unique values of $\boldsymbol{\theta}^{(t)}_{(q)}$ and $\mathbf{u}^{(t)}_{(q)}$.

Having updated Y_1, FCS performs a second set of Bayesian estimation steps, this time treating Y_2 as the outcome and Y_1 and Y_3 as predictors. As before, the resulting parameter values and residual terms define a normal distribution, from which the algorithm draws new Y_2 imputations.

$$\begin{aligned} &Y^{(t)}_{2ij(\text{mis})} \sim \mathrm{N}\left(\hat{Y}^{(t)}_{2ij},\ \ \sigma^2_{\varepsilon(Y_2)}\right) \\ &\hat{Y}^{(t)}_{2ij} = \beta_{0(Y_2)} + \beta_{1(Y_2)} Y^{(t)}_{1ij} + \beta_{2(Y_2)}\, Y^{(t-1)}_{3ij} + u_{0(Y_2)} + \ \ u_{1(Y_2)} Y^{(t)}_{1ij} \end{aligned} \tag{5}$$

Notice that the distribution attempts to preserve the random influence of Y_1 on Y_2 in the analysis model by incorporating a symmetric random effect for the regression of Y_2 on Y_1. Although relatively little work has investigated imputation for random slopes, this so-called "reversed random coefficient" specification reflects the current implementation of multilevel FCS (Grund, Lud̈tke, & Robitzsch, 2016; van Buuren, 2011, 2012; van Buuren et al., 2014).

Finally, FCS performs a third Bayesian analysis that treats Y_3 as the outcome, after which it draws new Y_3 imputations from the following distribution.

$$\begin{aligned} Y^{(t)}_{3ij(\text{mis})} &\sim \mathrm{N}\left(\hat{Y}^{(t)}_{3ij},\ \ \sigma^2_{\varepsilon(Y_3)}\right) \\ \hat{Y}^{(t)}_{3ij} &= \beta_{0(Y_3)} + \beta_{1(Y_3)} Y^{(t)}_{1ij} + \beta_{2(Y_3)}\, Y^{(t)}_{2ij} + u_{0(Y_3)} \end{aligned} \tag{6}$$

Because the analysis model does not posit a random slope for Y_3, the distribution's mean incorporates a Level-2 residual term for only the intercept.

Thus far, we have considered the current incarnation of multilevel FCS, as described by (van Buuren, 2011, 2012) and implemented in the MICE package for R (van Buuren et al., 2014). The FCS framework is very flexible and can accommodate a number of useful extensions that are not readily available to researchers. The remainder of the article outlines three such modifications that address important practical problems that arise in behavioral research: (a) incomplete nominal and ordinal variables; (b) analyses that partition relations into within- and between-cluster relations (e.g., contextual effects analyses; multilevel structural equation models); and (c) incomplete Level-2 variables. This functionality is implemented in the Blimp application for Mac and Windows.

Bayesian Estimation for Ordinal and Nominal Outcomes

The current application of FCS to multilevel data is limited to normally distributed variables, and published studies have yet to extend FCS to incomplete categorical variables. The categorical imputation routine that we outline in this manuscript borrows from established Bayesian estimation procedures for probit regression models (Agresti, 2012; Albert & Chib, 1993; Finney & DiStefano, 2013; Johnson & Albert, 1999), variants of which are implemented in the joint model imputation framework (Asparouhov & Muthén, 2010; Carpenter, Goldstein, & Kenward, 2011; Carpenter & Kenward, 2013; Goldstein, Carpenter, Kenward, & Levin, 2009). To provide some necessary background, this section gives a brief overview of the complete-data estimation steps for a multilevel probit model. Consistent with FCS for normally distributed variables, categorical imputation can be viewed as a complete-data Bayesian analysis with an additional step that fills in the data, conditional on multilevel model parameters and Level-2 residuals. For now, we focus on an analysis with Level-1 variables, but the procedure readily generalizes to higher-level variables.

To motivate the ensuing discussion, consider a simple random intercept model with a single Level-1 predictor and binary outcome variable with discrete values of zero and one. Probit regression views discrete responses as arising from a normally distributed latent variable, often denoted Y^* in the literature. For example, if Y is a clinical depression indicator (e.g., 0 = not depressed, 1 = clinically depressed), the model defines a corresponding Y^* latent variable representing a normally distributed propensity for clinical depression. The resulting model for the underlying latent variable is as follows.

$$\begin{aligned} &Y^*_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + \varepsilon_{ij} \\ &u_{0j} \sim \mathrm{N}(0,\ \ \sigma^2_u) \quad \varepsilon_{ij} \sim \mathrm{N}(0,\ \ 1) \end{aligned} \tag{7}$$

Conceptually, Equation (7) is standard linear multilevel regression model with a latent outcome variable. However, because the latent variable is not observed, the model constrains the within-cluster residual variance to unity to define a scale (i.e., Y^* is a within-cluster *z*-score).

The cumulative probit model for ordinal data uses a threshold parameter (or parameters) to link the latent variable distribution to the discrete responses. In the case of a binary outcome, a single threshold parameter τ divides the latent variable distribution into two regions, such that discrete values of one and zero correspond to latent scores above and below the threshold, respectively. The threshold is typically fixed at zero because it is redundant with the regression intercept, but an equivalent parameterization fixes the intercept to zero and estimates the threshold. More generally, ordered categorical variables with $K > 2$ response options ($k = 1, \ldots, K$) require $K - 1$ threshold parameters, and the following function relates the discrete and latent scores.

$$Y = f(Y^*) = \begin{cases} 1 & \text{if } -\infty < Y^* < \tau_1 \\ 2 & \text{if } \tau_1 < Y^* < \tau_2 \\ \vdots & \\ K & \text{if } \tau_{K-1} < Y^* < \infty \end{cases} \tag{8}$$

With $K > 2$ response categories, the first threshold τ_1 is often fixed at zero, but the decision to estimate the intercept instead of this threshold (or vice versa) is arbitrary. The top panel of Figure 1 depicts the within-cluster latent variable distributions for a binary outcome at three values of X, and the bottom panel shows a five-category ordinal variable with four threshold parameters.

The latent variable formulation for categorical variables offers computational advantages because it integrates with established Bayesian estimation procedures for normally distributed outcomes. Specifically, the Gibbs sampler begins by updating the threshold parameters (if $K > 2$) and sampling latent scores for the entire sample, after which it uses identical steps from Equation (2) to update parameters and Level-2 residual terms from the model in Equation (7). More formally, the sampling steps for a single iteration t of the algorithm are

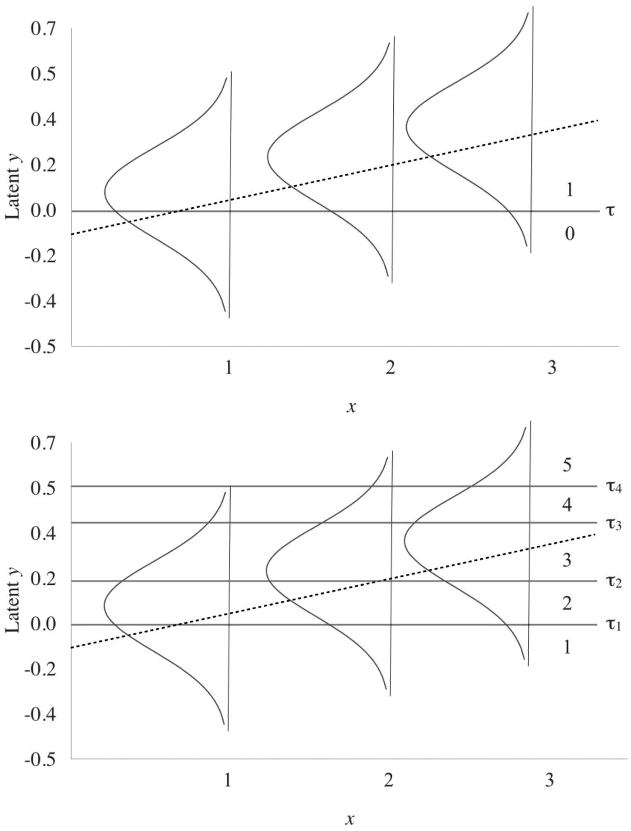

Figure 1. Latent Y^* distributions for a categorical Y variable at three values of X. The dashed line represents the within-cluster regression (i.e., $\beta_{0j} = \beta_0 + u_{0j}$ and $\beta_{1j} = \beta_1 + u_{1j}$), and the horizontal line(s) denotes the threshold(s). The top panel depicts a binary Y variable where a discrete score of $Y = 1$ occurs when the measurement process yields a Y^* value above the threshold τ, and a discrete score of $Y = 0$ occurs when Y^* falls below the threshold. The bottom panel depicts a five-category ordinal variable with four threshold parameters.

$$\begin{aligned}
\boldsymbol{\tau}^{(t)} &\sim \text{N}\left(\boldsymbol{\tau} \mid \mathbf{Y},\ \mathbf{Y}^{*(t-1)},\ \boldsymbol{\beta}^{(t-1)},\ \mathbf{u}^{(t-1)},\ \boldsymbol{\Sigma}_u^{(t-1)}\right) \\
\mathbf{Y}^{*(t)} &\sim \text{TN}\left(\mathbf{Y}^* \mid \mathbf{Y},\ \boldsymbol{\tau}^{(t)},\ \boldsymbol{\beta}^{(t-1)},\ \mathbf{u}^{(t-1)},\ \boldsymbol{\Sigma}_u^{(t-1)}\right) \\
\boldsymbol{\beta}^{(t)} &\sim \text{MVN}\left(\boldsymbol{\beta} \mid \mathbf{Y},\ \boldsymbol{\tau}^{(t)},\ \mathbf{Y}^{*(t)},\ \mathbf{u}^{(t-1)},\ \boldsymbol{\Sigma}_u^{(t-1)}\right) \qquad (9) \\
\mathbf{u}^{(t)} &\sim \text{MVN}\left(\mathbf{u} \mid \mathbf{Y},\ \boldsymbol{\tau}^{(t)},\ \mathbf{Y}^{*(t)},\ \boldsymbol{\beta}^{(t)},\ \boldsymbol{\Sigma}_u^{(t-1)}\right) \\
\boldsymbol{\Sigma}_u^{(t)} &\sim \text{IW}\left(\boldsymbol{\Sigma}_u \mid \mathbf{Y},\ \boldsymbol{\tau}^{(t)},\ \mathbf{Y}^{*(t)},\ \boldsymbol{\beta}^{(t)},\ \mathbf{u}^{(t)}\right)
\end{aligned}$$

where TN denotes the truncated normal distribution (explained below), and the remaining terms are the same as before. Note that the sampling step for the within-cluster residual variance is absent because this parameter is a constant. Published resources (Albert & Chib, 1993; Cowles, 1996; Goldstein et al., 2007) and the technical document in the supplemental online material give a description of the updating step for thresholds, and the remaining distributions are the same as those for a linear multilevel model.

A brief description of the process that generates latent variable scores provides insight into categorical imputation, which we describe in the next section. To begin, reconsider the regression model from Equation (7), first assuming that Y is a binary outcome (e.g., a clinical depression indicator). The model from Equation (7) implies the following latent variable distribution for each case.

$$Y_{ij}^{*(t)} \sim \text{N}(\beta_0 + \beta_1 X_{ij} + u_{0j}, 1) \qquad (10)$$

For clarity, we omit iteration superscripts on the parameters and residual terms, noting that these quantities carry forward from the previous Gibbs cycle. Recall that a threshold parameter divides the latent variable distribution into two regions, such that a discrete score of zero requires a Y^* value below the threshold, and a score of one requires a Y^* value above the threshold. The sampling procedure honors this linkage, drawing latent variable scores that are restricted to the appropriate region of the normal distribution. More formally, the sampling procedure draws latent variable scores from a truncated normal distribution (Robert, 1995), denoted by ~TN in Equation (9). The procedure for ordinal variables with $K > 2$ response options is identical but requires additional threshold parameters. For example, consider a five-category variable with response options $k = 1, \ldots, 5$. Cases with $Y = 1$ must have latent scores between negative infinity and τ_1, cases with $Y = 2$ must have latent values between τ_1 and τ_2, and so on.

The multinomial probit model can accommodate nominal variables with $K > 2$ categories (Aitchison & Bennett, 1970; Albert & Chib, 1993; Goldstein et al., 2009). For example, suppose that the outcome variable is a three-category depression diagnosis (e.g., 1 = clinical depression, 2 = subclinical depression, 3 = no depression). The multinomial model defines an underlying normal variable U^* for each of the K discrete response options that can be viewed as the latent propensity of endorsing a particular category (e.g., a normally distributed propensity for each diagnosis). The latent variables can be expressed more succinctly as a set of $K - 1$ latent difference scores, each of which contrasts the U^* value for a particular category against that of an arbitrary reference group (e.g., the response with the highest numeric code). For example, the latent difference scores for a nominal variable with $K = 3$ response options ($k = 1, 2, 3$) are

$$\begin{aligned}
Y_1^* &= U_1^* - U_3^* \\
Y_2^* &= U_2^* - U_3^*
\end{aligned} \qquad (11)$$

where the highest code (e.g., 3 = no depression) is the reference group. In the context of the depression example, Y_1^* and Y_2^* represent the propensity for clinical and subclinical depression, respectively, relative to the no-depression comparison group.

The resulting model for the underlying normal latent variables is as follows.

$$\begin{aligned}
Y_{1ij}^* &= \beta_{0(Y_1^*)} + \beta_{1(Y_1^*)}X_{ij} + u_{0j(Y_1^*)} + \varepsilon_{ij(Y_1^*)} \\
Y_2^* &= \beta_{0(Y_2^*)} + \beta_{1(Y_2^*)}X_{ij} + u_{0j(Y_2^*)} + \varepsilon_{ij(Y_2^*)} \qquad (12) \\
u_{0j} &\sim \text{MVN}(\mathbf{0},\ \boldsymbol{\Sigma}_u) \quad \varepsilon_{ij} \sim \text{MVN}(\mathbf{0}, \mathbf{1})
\end{aligned}$$

The variable name subscripts on the coefficients and residual terms indicate that these quantities can differ across latent variables, such that X may be a stronger predictor of Y_1^* than Y_2^* (or vice versa). Specifying the within-cluster covariance matrix as an identity matrix defines the scales of the latent variables, as it did in the ordinal model.

The multinomial probit model does not require threshold parameters. Rather, category membership implies a particular rank order and magnitude for the latent variable difference scores. Specifically, for a nominal variable with K response options ($k = 1, \ldots, K$), the following function relates the discrete and latent scores.

$$Y = g(Y^*)$$
$$= \begin{cases} 1 & \text{if } Y_1^* = \max[Y_1^*, Y_2^*, \ldots, Y_{K-1}^*] \text{ and } Y_1^* > 0 \\ 2 & \text{if } Y_2^* = \max[Y_1^*, Y_2^*, \ldots, Y_{K-1}^*] \text{ and } Y_2^* > 0 \\ & \vdots \\ K-1 & \text{if } Y_{K-1}^* = \max[Y_1^*, Y_2^*, \ldots, Y_{K-1}^*] \text{ and } Y_{K-1}^* > 0 \\ K & \text{if } \max\,[Y_1^*, Y_2^*, \ldots, Y_{K-1}^*] < 0 \end{cases} \tag{13}$$

Returning to the three-category depression example, membership in the first category implies that $U_1^* > U_3^*$ or equivalently $Y_1^* > 0$ and $Y_1^* > Y_2^*$ (i.e., the latent propensity for the first category, clinical depression, is greater than that of both the second and third categories, subclinical and no depression, respectively). Similarly, membership in the second category implies that $U_2^* > U_3^*$ or equivalently $Y_2^* > 0$ and $Y_2^* > Y_1^*$. Finally, membership in the third category (the reference group) implies that both Y_1^* and Y_2^* are less than zero (i.e., the latent propensities of belonging to the first and second category are less that of the third category).

Because the multinomial model does not employ thresholds, the initial sampling step that generates the latent variable scores uses an accept-reject algorithm (Goldstein et al., 2009) to repeatedly draw a set of Y^* values for each case until it obtains values that satisfy the rules from Equation (13). After sampling latent scores for the entire sample, the algorithm uses the final three steps of Equation (9) to update the parameters and Level-2 residual terms from the model in Equation (12). These updating steps are identical to the corresponding steps for normally distributed variables in Equation (2). The technical document in the supplemental online material gives additional details.

Categorical Variable Imputation

Consistent with the procedure for normal variables, FCS imputation for categorical variables is essentially a complete-data Bayesian analysis with an additional step that fills in the missing data. Missing values necessitate three changes to the Gibbs sampler from Equation (9). We summarize these in text, and refer readers to the online supplemental material for additional details. First, the second estimation step applies only to the complete cases because the procedure for drawing latent scores from a truncated normal distribution must condition on an observed discrete response. Second, each estimation cycle concludes with an additional step that generates imputations based on the current model parameters and Level-2 residual terms. As illustrated in Equations (7) and (12), the Bayesian estimation steps are modeling the underlying normal variable, and so imputation is also performed on the latent variable metric. However, the procedure for drawing Y^* imputations is somewhat different than that for the complete cases because it is no longer possible to condition on a discrete response. Rather, the imputation step accounts for missing data uncertainty by drawing latent variable imputes from a normal distribution with no truncation or restrictions on the latent variable's range. Finally, after completing imputation, the algorithm converts the latent variable imputes to discrete scores by applying Equation (8) to ordinal variables or Equation (13) to nominal variables.

To illustrate categorical imputation, consider the following analysis model

$$Y_{1ij}^{(\text{o})} = \beta_0 + \beta_1 Y_{2ij}^{(\text{n})} + \beta_2 Y_{3ij}^{(\text{n})} + u_{0j} + \varepsilon_{ij} \tag{14}$$

where Y_1 is an incomplete ordinal variable, and Y_2 and Y_3 are dummy codes representing an incomplete nominal variable with three categories. We use (o) and (n) in the superscripts to remind readers that the variables are ordinal and nominal, respectively. To begin, FCS applies the Bayesian estimation steps from Equation (9), treating Y_1 as the outcome and Y_2 and Y_3 as predictors. The resulting parameter values and residual terms define a normal distribution that generates Y_1 imputations on the latent variable metric, as follows.

$$Y_{1ij}^{*(t)} \sim \text{N}\left(\beta_{0(Y_1^*)} + \beta_{1(Y_1^*)} Y_{2ij}^{(t-1)} + \beta_{2(Y_1^*)} Y_{3ij}^{(t-1)} + u_{0j(Y_1^*)},\ 1\right) \tag{15}$$

After applying the rules from Equation (8) to create discrete imputes, a second sequence of Bayesian estimation steps provides parameter values and residual terms for the nominal imputation models.

$$\begin{aligned} Y_{2ij}^{*(t)} &\sim \text{N}\left(\beta_{0(Y_2^*)} + \beta_{1(Y_2^*)} Y_{1ij}^{(t)} + u_{0j(Y_2^*)},\ 1\right) \\ Y_{3ij}^{*(t)} &\sim \text{N}\left(\beta_{0(Y_3^*)} + \beta_{1(Y_3^*)} Y_{1ij}^{(t)} + u_{0j(Y_3^*)},\ 1\right) \end{aligned} \tag{16}$$

The algorithm next applies the categorization rules from Equation (13) to the latent variable imputations, after which it begins the next round of Y_1 imputation.

It is important emphasize that categorical imputation is very different from rounding schemes that have appeared in the literature (Allison, 2002, 2005; Bernaards, Belin, & Schafer, 2007; Yucel, He, & Zaslavsky, 2008), most of which are capable of introducing substantial biases (Horton, Lipsitz, & Parzen, 2003). For example, the so-called naïve rounding approach imputes discrete variables as though they are normally distributed and subsequently rounds the fractional imputes to the nearest integer. Unlike these ad hoc methods, the procedure we outline is grounded in statistical theory (Agresti, 2012; Aitchison & Bennett, 1970; Albert & Chib, 1993; Carpenter & Kenward, 2013; Johnson & Albert, 1999) and applies established Bayesian estimation steps from the literature (Albert & Chib, 1993; Cowles, 1996; Goldstein et al., 2007). Further, the latent variable approach is an established method for joint model imputation (Asparouhov & Muthén, 2010; Carpenter & Kenward, 2013; Muthén & Muthén, 1998–2012) that appears to work well with single-level data (Wu, Jia, & Enders, 2015) and two-level random intercept models (Enders et al., 2016).

Partitioning Within- and Between-Cluster Variation With FCS

Many multilevel analyses apply models that partition relations among Level-1 variables into within- and between-cluster components. One common example is the classical contextual effects analysis from the multilevel regression literature (Longford, 1989; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Lüdtke et al., 2008; Raudenbush & Bryk, 2002; Shin & Raudenbush, 2010). An example of this model is

$$Y_{1ij} = \beta_0 + \beta_1 Y_{2ij} + \beta_2 \overline{Y}_{2j} + \beta_3 X_{ij} + u_{0j} + \varepsilon_{ij} \tag{17}$$

where β_1 is the pooled within-cluster regression of Y_1 on Y_2, and β_2 is the difference between the within-cluster regression and the between-cluster regression of $\overline{Y}_{1j}$ on $\overline{Y}_{2j}$ (i.e., the contextual effect), and X is a covariate. Raudenbush and Bryk (2002) gave an example of a contextual effects analysis where student-level socioeconomic status and school-average socioeconomic status (e.g., Y_{2ij} and $\overline{Y}_{2j}$, respectively, in the above equation) predict academic achievement, and published applications of this model are common in the literature (e.g., Harker & Tymms, 2004; Kenny & La Voie, 1985; Lütdke, Koller, Marsh, & Trautwein, 2005; Miller & Murdock, 2007; Simons, Wills, & Neal, 2014). Multilevel structural equation modeling is a second analysis framework that models unique within- and between-cluster covariance structures. As an example, Martin, Malmberg, and Liem (2010) reported the results from a multilevel factor analysis where the internal structure of individual and school-average academic motivation and engagement differed. Other similar applications are common in the substantive literature (Dunn, Masyn, Jones, Subramanian, & Koenen, 2015; Huang & Cornell, 2015; Muthén, 1991; Reise, Ventura, Nuechterlein, & Kim, 2005; Toland & De Ayala, 2005).

Although not immediately obvious, the standard formulation of FCS described in the previous sections is incapable of partitioning relations among Level-1 variables into within- and between-cluster components because it places implicit equality constraints on functions of the within- and between-cluster covariance matrices (Mistler & Enders, 2016). Analytic work and computer simulation results show that applying FCS to models such as that in Equation (17) can introduce considerable bias, even under a missing completely at random (MCAR) mechanism (Carpenter & Kenward, 2013; Enders et al., 2016; Mistler & Enders, 2016). Carpenter and Kenward (2013, p. 220) outlined a modification to FCS (attributed to a personal communication from Ian White) that addresses this problem by introducing the cluster means of Level-1 variables into the imputation model.

Implementing Carpenter and Kenward's modification is straightforward. Following each imputation step, the FCS algorithm computes the cluster means from the filled-in data, and both the Level-1 variable and its cluster means function as predictors in subsequent imputation steps. To illustrate this modification, reconsider the contextual effects analysis model from Equation (17), and assume that X is complete and Y_1 and Y_2 are incomplete. We further assume that all variables are normally distributed, but the procedure works the same with categorical variables. Omitting the supporting sampling steps that provide the parameter values and Level-2 residual terms, FCS draws imputations from the following distributions

$$\begin{aligned} Y^{(t)}_{1ij(\text{mis})} &\sim \text{N}(\beta_{0(Y_1)} + \beta_{1(Y_1)} Y^{(t-1)}_{2ij} + \beta_{2(Y_1)} X_{1ij} \\ &\quad + \beta_{3(Y_1)} \overline{Y}^{(t-1)}_{2j} + \beta_{4(Y_1)} \overline{X}_{1j} + u_{0(Y_1)},\ \sigma^2_{\varepsilon(Y_1)}) \\ Y^{(t)}_{2ij(\text{mis})} &\sim \text{N}(\beta_{0(Y_2)} + \beta_{1(Y_2)} Y^{(t)}_{1ij} + \beta_{2(Y_2)} X_{1ij} \\ &\quad + \beta_{3(Y_2)} \overline{Y}^{(t)}_{1j} + \beta_{4(Y_2)} \overline{X}_{1j} + u_{0(Y_2)},\ \sigma^2_{\varepsilon(Y_2)}) \end{aligned} \tag{18}$$

where $\overline{Y}^{(t-1)}_{2j}$, $\overline{Y}^{(t)}_{1j}$ and $\overline{X}_{1j}$ are cluster means.

The FCS imputation models from Equation (18) are more general than the analysis model because they partition all relations into within- and between-cluster components, whereas the analysis model does so only for Y_1 and Y_2. This generality is not detrimental and would be important for certain analytic contexts. For example, suppose that the three variables were indicators of a latent factor in a multilevel confirmatory factor analysis. The imputation models in Equation (18) do not impose constraints on the within- and between-cluster covariance matrices and thus could accommodate a model that posits a different factor structure at Level-1 and Level-2, different loading magnitudes, factor variances, and so on (e.g., Martin et al., 2010). Analytic work and computer simulations suggest that Carpenter and Kenward's (2013) modification to FCS adequately preserves the within- and between-cluster covariance matrices (Carpenter & Kenward, 2013; Enders et al., 2016; Mistler & Enders, 2016). The Blimp application incorporates cluster means by default, but users can disable this option.

FCS Imputation for Incomplete Level-2 Variables

Imputation for incomplete Level-2 variables is straightforward with some, but not all, incarnations of joint model imputation (Asparouhov & Muthén, 2010; Carpenter et al., 2011; Goldstein et al., 2009). Briefly, the joint model is a multivariate approach that uses saturated within- and between-cluster covariance matrices to generate imputations (e.g., for an overview, see Enders et al., 2016). This framework defines all variables as having two levels, and it constrains to zero all elements of the within-cluster covariance matrix that correspond to the Level-2 variables. These constraints produce Level-2 imputations that are effectively the sum of a grand mean and a between-cluster residual term. Methodologists have described an analogous model specification for maximum likelihood estimation of two-level models (Liang & Bentler, 2004).

Despite the ease with which the joint model generates Level-2 imputations, current applications of this approach have little or no capacity for preserving random slope variation because they assume a common within-cluster covariance matrix for all clusters (Enders et al., 2016).[2] In our view, this limitation provides a compelling rationale for building out the FCS framework, which can readily accommodate random slopes (e.g., see the earlier description and illustration of FCS). However, the FCS literature has thus far addressed only incomplete Level-1 variables (van Buuren, 2011, 2012), and methods for imputing incomplete Level-2 variables are not automatically available in software. Methodologists have suggested that Level-2 missingness may be addressed by aggregating the data and applying single-level imputation to a cluster-level data set with J records (Gelman & Hill, 2007; Yucel, 2008), and analytic work from Carpenter and Kenward (2013, pp. 220–221) provides a formal mathematical rationale for this strategy.

This section outlines a Level-2 imputation strategy that applies the following steps: (a) use the procedure from the previous sections to impute all Level-1 variables, conditioning on the current Level-2 imputations; (b) aggregate the data, creating a J-record data set where each row contains the cluster means and Level-2 scores for cluster j; (c) apply single-level FCS to the incomplete Level-2 variables; and (d) carry the Level-2 imputes forward to the next round of Level-1 imputation. Consistent with Level-1 imputation, the Level-2 procedure can be viewed as a complete-data Bayesian analysis with an additional step that fills in the data, conditional on the model parameters from a particular iteration. The key difference is that a series of single-level regres-

[2] Yucel (2011) outlined a Gibbs sampler for cluster-specific covariance matrices, but this approach has not been evaluated in the literature.

sion models define the distributions of missing data. Complete-data Bayesian estimation for linear regression requires sampling steps for the coefficients and residual variance, as follows.

$$\begin{aligned} \boldsymbol{\beta}^{(t)} &\sim \mathrm{MVN}(\boldsymbol{\beta} \,|\, \mathbf{Y},\ \sigma_{\mu}^{2(t-1)}) \\ \sigma_{\mu}^{2(t)} &\sim \mathrm{IG}(\sigma_{\mu}^{2} \,|\, \mathbf{Y},\ \boldsymbol{\beta}^{(t)}) \end{aligned} \tag{19}$$

Note that we use a u and σ_{μ}^{2} to denote the residual and residual variance, respectively, to emphasize that variation is at the between-cluster level. The specific details of each distribution are found in the online supplemental material and in published resources (Gelman et al., 2014; Lynch, 2007; Sinharay et al., 2001; van Buuren, 2012; van Buuren et al., 2006).

To illustrate Level-2 imputation more concretely, consider the following analysis model

$$Y_{1ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Y_{2j} + \beta_3 Y_{3j} + u_{0j} + \varepsilon_{ij} \tag{20}$$

where Y_1 and X are incomplete and complete Level-1 variables, respectively, and Y_2 and Y_3 are incomplete Level-2 predictors. Further, temporarily assume that all variables are normally distributed. Following Y_1 imputation, FCS aggregates the Level-1 variables and creates a cluster-level data set where each row contains the Level-2 scores and the cluster means of Y_1 and X. FCS first applies the Bayesian estimation steps from Equation (19) to the filled-in data, treating Y_2 as the outcome, and the remaining variables as predictors. The resulting parameter values define a normal distribution that generates updated Y_2 imputations. A second sequence of Bayesian estimation steps provides the parameters for Y_3 imputation. The Level-2 imputation models are as follows

$$\begin{aligned} Y_{2j(\mathrm{mis})}^{(t)} &\sim \mathrm{N}(\beta_{0(Y_2)} + \beta_{1(Y_2)} Y_{3j}^{(t-1)} + \beta_{2(Y_2)} \overline{Y}_{1j}^{(t)} + \beta_{3(Y_2)} \overline{X}_{1j}, \sigma_{\mu(Y_2)}^{2}) \\ Y_{3j(\mathrm{mis})}^{(t)} &\sim \mathrm{N}(\beta_{0(Y_3)} + \beta_{1(Y_3)} Y_{2j}^{(t)} + \beta_{2(Y_3)} \overline{Y}_{1j}^{(t)} + \beta_{3(Y_3)} \overline{X}_{1j}, \sigma_{\mu(Y_3)}^{2}) \end{aligned} \tag{21}$$

where a predicted value and residual variance again define the center and spread of the distributions, respectively. We reiterate that Equation (21) is a single-level imputation model, with the σ_{μ}^{2} parameters capturing between-cluster residual variation. As noted previously, analytic work from Carpenter and Kenward (2013, pp. 220–221) shows that including aggregated Level-1 variables (e.g., $\overline{Y}_{1j}^{(t)}$ and $\overline{X}_{1j}$) in the Level-2 imputation models is important for preserving the between-cluster covariance structure.

The categorical imputation procedure described earlier in the article readily extends to Level-2 variables. In this situation, the estimation steps from Equation (9) simplify because the residuals and their covariance matrix (i.e., $\mathbf{u}$ and Σ_{u}) are no longer needed for a single-level probit model. Rather, the estimation steps update threshold parameters (ordinal variables with $K > 2$ categories), latent scores for the complete cases, and regression coefficients, after which latent variable imputations are drawn from an unrestricted normal distribution. For example, suppose that Y_2 and Y_3 from the previous analysis model are incomplete ordinal variables. The imputation steps for these variables are as follows.

$$\begin{aligned} Y_{2j(\mathrm{mis})}^{*(t)} &\sim \mathrm{N}(\beta_{0(Y_2)} + \beta_{1(Y_2)} Y_{3j}^{(t-1)} + \beta_{2(Y_2)} \overline{Y}_{1j}^{(t)} + \beta_{3(Y_2)} \overline{X}_{1j}, 1) \\ Y_{3j(\mathrm{mis})}^{*(t)} &\sim \mathrm{N}(\beta_{0(Y_3)} + \beta_{1(Y_3)} Y_{2j}^{(t)} + \beta_{2(Y_3)} \overline{Y}_{1j}^{(t)} + \beta_{3(Y_3)} \overline{X}_{1j}, 1) \end{aligned} \tag{22}$$

As before, the variance of the latent variable distributions is fixed at unity for identification, and the latent imputes are categorized at the end of each step using the current threshold values and the rules from Equation (8).

Simulation Study

To investigate the performance of our FCS approach, we designed a Monte Carlo simulation study with four between-subjects factors: number of clusters (J = 25, 50, and 200), within-cluster sample size (n_j = 5, 15, 25, and 50), intraclass correlation (ICC = .20 and .50), and the MAR missing data rate (0%, 5%, 15%, and 25%). We generated 2,000 replications within each of the 96 design cells, resulting in 192,000 replications. In choosing the levels of each factor we considered guidelines from the literature, conditions implemented in published Monte Carlo studies, and generalizability to typical behavioral science data sets. For example, ICC values of .20 and .50 are representative of cross-sectional (e.g., students nested in schools) and repeated measures (e.g., observations nested in subjects) designs, respectively (Spybrook et al., 2011), and these values are typical of ICCs from published research (Gulliford, Ukoumunne, & Chinn, 1999; Hedges & Hedberg, 2007; Murray & Blistein, 2003). Similarly, the Level-2 sample sizes we implement represent values that researchers might choose after consulting the methodological literature (e.g., Kreft & de Leeuw (1998) recommend at least 30 clusters, and Maas and Hox (2005) suggest that 50 clusters is a common value in educational and organizational settings). For within-cluster sample sizes, Maas and Hox (2005) suggest that n_j = 30 is typical of educational research settings, and we chose n_j = 5 as a lower limit for the within-cluster sample size because smaller values are known to produce imprecise random effect estimates in some situations (Clark & Wheaton, 2007; Raudenbush, 2008). Finally, it is difficult to determine appropriate missing data rates because authors rarely report this information. Nevertheless, the rates that we examine here are common in the missing data simulations and are sufficiently large to expose practical problems with imputation (e.g., a 25% missing data rate on every variable in the analysis model is probably uncommon in most applied scenarios).

Population Model and Data Generation

We used a two-level regression model with a random slope as the population data-generating model. For the ICC = .20 condition, the population regression model was

$$\begin{aligned} Y_{ij}^{(\mathrm{c})} = \beta_0 + \ & \beta_1 X_{1ij}^{(\mathrm{o})} + \beta_2 X_{21ij}^{(\mathrm{n})} + \beta_3 X_{22ij}^{(\mathrm{n})} + \beta_4 X_{3j}^{(\mathrm{o})} + \beta_5 X_{4j}^{(\mathrm{o})} + u_{0j} \\ & + u_{1j} X_{1ij}^{(\mathrm{o})} + \varepsilon_{ij} \end{aligned} \tag{23}$$

where Y is a continuous Level-1 outcome, X_1 is a six-category ordinal variable, and X_{21} and X_{22} are binary dummy codes representing a three-category nominal variable, and X_3 and X_4 are binary Level-2 covariates. We use alphanumeric superscripts on the variable names to remind readers of the metrics (i.e., c = continuous, o = ordinal, and n = nominal). Throughout the article, we have used X and Y to denote complete and incomplete variables, respectively, but we break from that convention here and use X to denote a predictor in the analysis model. We chose this model because it is sufficiently complex to represent published applications of MLMs and because it incorporates a combination of features that are difficult or impossible to handle with existing imputation frameworks. We acknowledge that some researchers may prefer to code the 6-category X_1 variable as a set of dummy variables, but we treat this variable as ordinal in order to evaluate the imputation routine; doing so does not inherently violate model assumptions because the multilevel analysis does not impose distributional assumptions on predictor variables. Although not depicted in the analysis model, the simulation also includes a normally distributed auxiliary variable at each level, A_1 and A_2. As described below, these variables determine missingness probabilities.

The data generation process first created random normal variables and subsequently used threshold parameters to form discrete values for the categorical variables. To facilitate the determination of model parameters, we began by specifying within- and between-cluster covariance matrices for the underlying normal variables, shown in Table 1. These matrices had the following properties: (a) predictors measured at the same level (e.g., X_3 and X_4) had correlations of .30; (b) auxiliary variables had .40 correlations with other variables measured at the same level (e.g., A_1 and Y) but were uncorrelated with variables at the opposite level (e.g., A_1 and X_3); and (c) all predictors had a .30 correlation with the outcome variable. We chose correlations of .30 to align with Cohen's (1988) definition of a medium effect size, and we specified somewhat stronger correlations for the auxiliary variables to ensure that omitting these variables from imputation would introduce bias (Collins, Schafer, & Kam, 2001).

The following steps produced the underlying normally distributed versions of the variables. First, we created the Level-2 variables (i.e., A_2, X_3, X_4) and the between-cluster components of the Level-1 predictors (i.e., $\overline{X}_{1j}$ and $\overline{X}_{2j}$). Second, we generated the within-cluster components of the Level-1 variables (i.e., A_1, X_1 and X_2). These steps first generated standard normal variables and then used Cholesky decomposition to transform the z-scores to the desired covariance structure from Table 1. Third, we generated Level-2 residuals as z-scores and again used Cholesky decomposition to transform them to the desired covariance structure. We determined the residual intercept variance by solving for the regression of $\overline{Y}_j$ on the between-cluster variables. Based on some preliminary power simulations, we set the slope variance equal to 30% of the total Level-2 variance, and we arbitrarily specified a .30 correlation between the intercept and slope residuals. Fourth, we computed a vector of predicted scores that conditioned on the within- and between-cluster predictors and the Level-2 residual terms, and defined Y as the sum of a predicted score and a within-cluster residual. We again used the appropriate elements of the covariance matrices in Table 1 to obtain the coefficients and residual variance for this step.

Table 1
Within- and Between-Cluster Covariance Matrices for Data Generation

	A_1	A_2	Y	X_1	X_2	X_3	X_4
			ICC = .20				
A_1	**0**	**0**	**0**	**0**	**0**	**0**	**0**
A_2	0	**1.00**	**0**	**0**	**0**	**.40**	**.40**
Y	.40	0	**.25**	**.08**	**.08**	**.15**	**.15**
X_1	.40	0	.30	**.25**	**.08**	**0**	**0**
X_2	.40	0	.30	.30	**.25**	**0**	**0**
X_3	0	0	0	0	0	**1.00**	**.30**
X_4	0	0	0	0	0	0	**1.00**
			ICC = .50				
A_1	**0**	**0**	**0**	**0**	**0**	**0**	**0**
A_2	0	**1.00**	**0**	**0**	**0**	**.40**	**.40**
Y	.40	0	**1.00**	**.30**	**.30**	**.30**	**.30**
X_1	.40	0	.30	**1.00**	**.30**	**0**	**0**
X_2	.40	0	.30	.30	**1.00**	**0**	**0**
X_3	0	0	0	0	0	**1.00**	**.30**
X_4	0	0	0	0	0	0	**1.00**

Note. The diagonal displays the between-cluster variances. The within-cluster variances of all Level-1 variables equal 1, and these quantities are zero for Level-2 variables. The lower-diagonal gives the average within-cluster covariances, and the upper diagonal elements in bold typeface give the between-cluster variances covariances. Because the ICC = .50 condition has within- and between-cluster variances set to 1.00, all off-diagonal elements can be viewed as correlations.

After generating underlying normal variables, we used the cumulative distribution function of the normal distribution to determine threshold parameters for categorizing the predictor variables. Specifically, we chose thresholds that approximately recoded (a) X_1 into a six-category ordinal variable with proportions equal to .10, .25, .30, .15, .10, and .10; (b) X_2 into three discrete groups with proportions of .20, .20, and .60; and (c) X_4 and X_5 into a binary variables with category proportions of .40 and .60 and .60 and .40, respectively. The choice of category proportions is somewhat arbitrary, but we chose the above values to mimic background variables that would not follow a uniform or symmetric distribution (e.g., education level, ethnicity, etc.).

The final step of data generation imposed MAR missing values on every variable in the analysis model. Recall that the data generation process included a pair of normally distributed auxiliary variables, A_1 and A_2. These variables determined missingness, such that higher scores on A_1 (or A_2) produced higher rates of missing data at Level-1 (or Level-2). We used logistic regression to relate the auxiliary variables to the missingness probabilities as follows. First, we used the latent variable formulation for logistic regression (Agresti, 2012; Johnson & Albert, 1999) to define a latent propensity of missingness at Level-1 and Level-2. To ensure a relatively strong selection mechanism, we set the correlation between this latent variable and the auxiliary variables at .40, from which we derived a logistic regression intercept and slope. Substituting the values of A_1 into the equation produced an N-row vector of Level-1 missingness probabilities, and doing the same with A_2 gave a J-row vector of Level-2 probabilities. For each Level-1 variable, we created an N-row vector of missing data indicators (0 = observed, 1 = missing) by sampling from a binomial distribution, such that the success rate for each observation was equal to its corresponding missingness probability. We applied the same procedure to the Level-2 variables, coding each variable as missing if its corresponding indicator equaled one. We used R version 3.2.3 to execute the data generation steps, and the syntax is available upon request.

Imputation and Estimation

We used the Blimp application to generate 50 imputations for each artificial data set. After examining the potential scale reduction diagnostic (Gelman & Rubin, 1992) from several data sets, we specified a burn-in period of 1,000 iterations and a thinning interval of 1,000 iterations (i.e., starting at the 1,000th Gibbs cycle, we saved a data set every 1000th iteration thereafter). We then used full maximum likelihood estimation in M*plus* 7 to fit the analysis model to each imputed data set, and we wrote a custom R program to pool the resulting estimates and standard errors. It is difficult to identify a useful comparison against which to evaluate our FCS approach because existing methods are unable to produce adequate imputations for the analysis model in Equation (23). For example, joint model approaches that use a latent variable formulation for categorical variables (e.g., the MLwiN and M*plus* programs) cannot preserve random slope variation and thus would yield biased random effects. Although it can accommodate random slopes, the current implementation of FCS (e.g., in the R package MICE) does not accommodate categorical variables. Interested readers can consult Enders et al. (2016) for a demonstration of these problems. Some structural equation modeling programs (e.g., M*plus*) could apply full information maximum likelihood estimation to the analysis model, but this approach is not a useful benchmark because it necessarily treats the categorical predictors as normally distributed random variables. Finally, listwise deletion is problematic in this simulation because it requires an MCAR mechanism. Thus, we restrict our attention to FCS.

We examined two outcomes, relative bias and confidence interval coverage. As noted previously, we used standard matrix expressions to derive regression parameters for generating the underlying normal variables. However, these parameters are no longer applicable after categorizing the predictors. Because it is difficult or impossible to analytically derive true values for a model with discrete explanatory variables, we instead used the estimates from a complete data set with a million cases (10,000 clusters with 100 cases each) to define the true values. The population parameters for the ICC = .20 and .50 conditions are given below.

$$\begin{aligned}
Y_{ij}^{(c)} &= 5.006 + .191\left(X_{1ij}^{(o)}\right) + .219\left(X_{21ij}^{(n)}\right) + .496\left(X_{22ij}^{(n)}\right)\\
&\quad + .198\left(X_{3j}^{(o)}\right) + .207\left(X_{4j}^{(o)}\right) + u_{0j} + u_{1j}X_{1ij}^{(o)} + \varepsilon_{ij}\\
\Sigma_u &= \begin{bmatrix} .198 & .030 \\ .030 & .034 \end{bmatrix} \qquad \sigma_\varepsilon^2 = .891\\
Y_{ij}^{(c)} &= 5.026 + .239\left(X_{1ij}^{(o)}\right) + .250\left(X_{21ij}^{(n)}\right) + .553\left(X_{22ij}^{(n)}\right)\\
&\quad + .402\left(X_{3j}^{(o)}\right) + .404\left(X_{4j}^{(o)}\right) + u_{0j} + u_{1j}X_{1ij}^{(o)} + \varepsilon_{ij}\\
\Sigma_u &= \begin{bmatrix} .800 & .119 \\ .119 & .128 \end{bmatrix} \qquad \sigma_\varepsilon^2 = .914
\end{aligned} \tag{24}$$

We defined relative bias as the difference between an average estimate and the true value divided by the true value (i.e., bias as a proportion of the true value). Authors regularly suggest that relative bias values less than .10 in absolute value are acceptable (Finch, West, & MacKinnon, 1997; Kaplan, 1988). Finally, we used the pooled standard errors to construct 95% confidence intervals for each estimate and computed confidence interval coverage as the proportion of replications where the normal-theory interval (i.e., the estimate plus or minus 1.96 standard error units) contained the true (complete-data) parameter value. With an alpha level of .05, an accurate imputation routine should produce coverage rates of .95. Values below the nominal rate indicate Type I error inflation (e.g., a coverage value of 90% suggests a twofold increase in Type I errors), whereas values exceeding .95 reflect conservative inference. When estimates are unbiased, confidence interval coverage unambiguously reflects the quality of the estimated standard errors, but biased estimates will distort coverage, even when standard errors are accurate.

Results

Figures 2 through 4 give trellis plots of relative bias by the number of Level-2 units, with dashed lines denoting the ±0.10 bias thresholds that authors routinely apply in simulation studies (Finch et al., 1997; Kaplan, 1988). For readers who want to inspect the numeric estimates and bias values in more detail, Tables 1 through 6 in the supplemental online material give the average parameter estimates and relative bias values for all combinations of conditions. With almost no exceptions, relative bias values for the fixed effects estimates fell below ±0.10, and the design factors had very little impact on parameter recovery. There was a slight tendency for accuracy to improve when the within-cluster sample size was $n_j = 15$ or larger, as relative bias values were generally near zero in these situations.

Turning to variance estimates, the sample sizes and missing data rate influenced the between-cluster covariance matrix estimates. Figures 2 through 4 highlight a number of trends. First, the intercept-slope covariance exhibited the largest bias values, followed by the slope and intercept variance, respectively. The covariance bias is likely an artifact of dividing by a population value that is relatively close to zero, so we are hesitant to emphasize this finding. The intercept variance estimates generally exhibited tolerable biases, and this parameter was largely unaffected by the missing data rate. Slope variance estimates were typically too low, with bias values reach or exceeding 10% at missing data rates of 15% or higher. Second, bias decreased as the within-cluster sample size increased, presumably because the reliability of the Level-2 residuals improved. Third, increasing the number of clusters from 25 to 50 influenced the estimates, but further increasing the number of clusters to 200 had virtually no impact. Comparing Figures 2 and 3, we see that random effect biases were generally smaller with 25 clusters than with 50 clusters. Although this trend may seem counterintuitive, the difference is attributable to the prior distribution, the influence of which depends on the number of clusters. Specifically, when the number of clusters was small, the inverse Wishart prior distribution counteracted negative bias by shifting the mass of the marginal posterior distributions to a higher positive value. Judging from the similarity of Figures 3 and 4, the influence of the prior effectively vanished with 50 or more clusters. As noted in the online supplemental material, our choice of prior distributions was informed by the literature and extensive simulation work. Nevertheless, we caution against overgeneralizing these results, as the influence of the prior may depend on features of the data or the analysis model. A variety of resources discuss the influence of prior distributions with small samples (e.g., Depaoli, 2014; McNeish, 2016; McNeish & Stapleton, 2016a).

Figures 5 through 7 give trellis plots of confidence interval coverage for the fixed effects slopes by the number of Level-2 units, with dashed lines at .925 and .975 denoting the so-called liberal criterion from Bradley (1978). We do not consider coverage

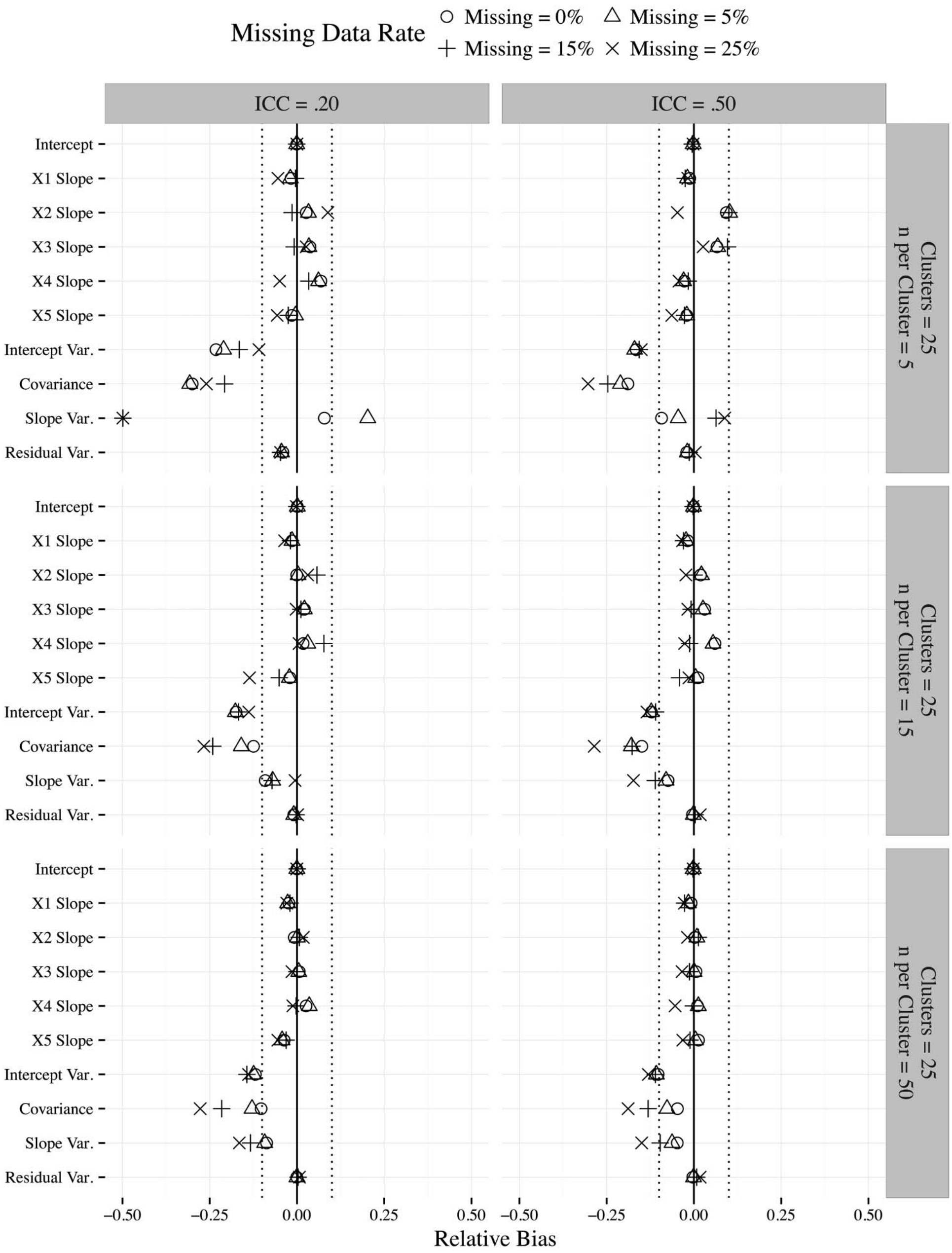

Figure 2. Average relative bias values for design cells with $j = 25$ clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ±0.10.

for variance estimates because the literature suggests that symmetric confidence intervals for these parameters are inappropriate (e.g., Maas & Hox, 2005; Snijders & Bosker, 2012), and we also omit the intercept because this parameter is typically not central to substantive hypotheses.[3] As seen in Figures 6 and 7, when the number of clusters was 50 or higher, coverage values for all slope coefficients generally fell within Bradley's liberal criterion. However, with 25 clusters, the Level-1 predictors had adequate coverage, but the values for Level-2 predictors were generally too low, with most values ranging between .88 and .925. Complete-data estimates—including those from our study—exhibit the same pattern (McNeish & Stapleton, 2016b; Stegmueller, 2013), so it does not appear that imputation exacerbates coverage problems.

[3] The intercept coefficient generally suffered from low coverage, with most values ranging between .90 and .925; this finding was independent of the missing data rate, with the complete data estimates exhibiting the same pattern.

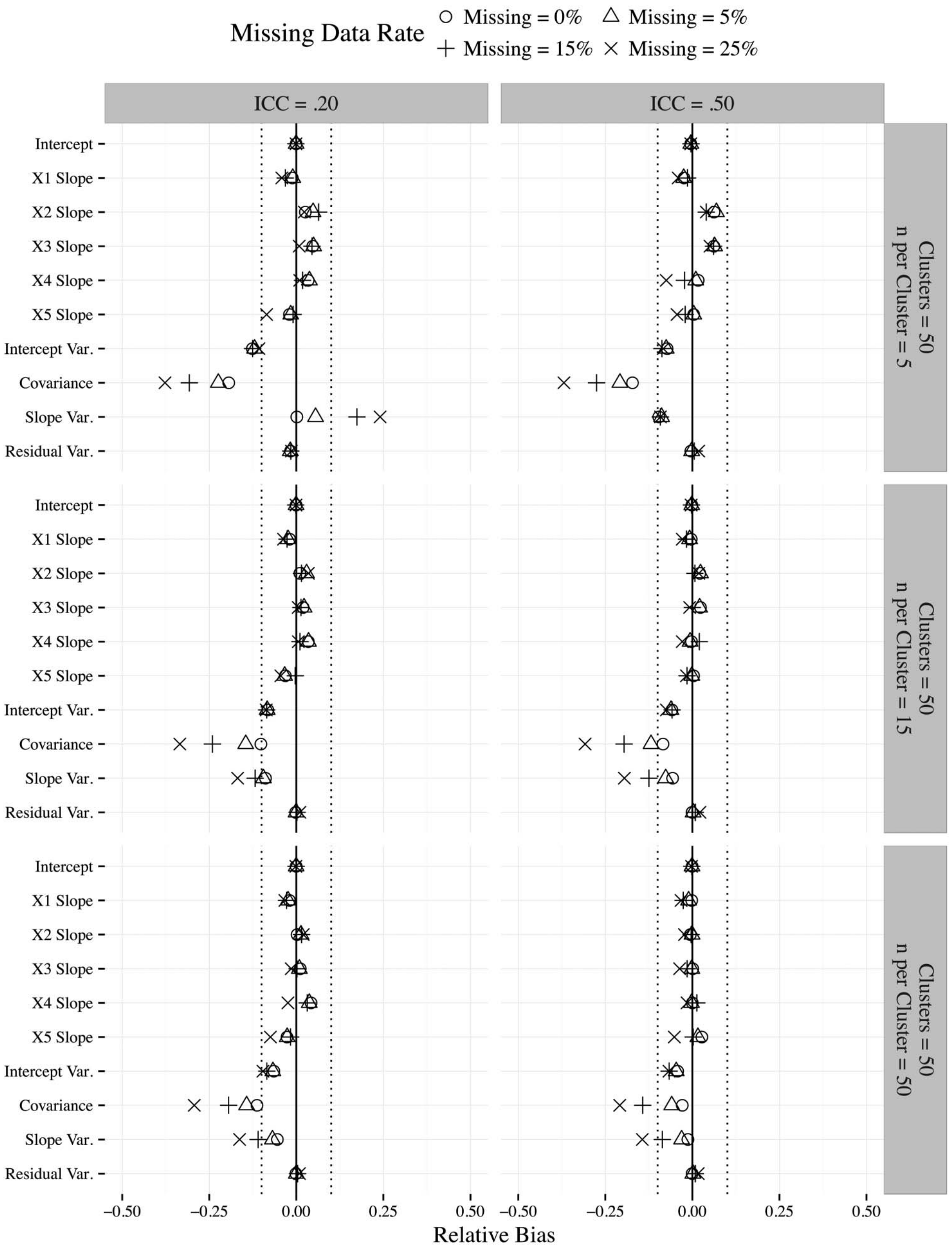

Figure 3. Average relative bias values for design cells with $j = 50$ clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ±0.10.

Software Implementation

The FCS imputation routine that we propose in this article is available in Blimp, an application for the Mac and Windows operating systems. Researchers can work from a simple command language or from a graphical interface. To illustrate the program, we consider the Blimp syntax for the analysis model in Equation (23). The syntax and the corresponding data file are available at www.appliedmissingdata.com/multilevel-imputation.html, as are a number of supporting documents and tutorials.

The Blimp syntax consists of a relatively small number of commands (shown in caps, although the program is not case sensitive), each of which ends in a colon. Commands are followed by one or more options or specifications, with a semicolon terminating each list. Briefly, the DATA command gives the file path to the raw ASCII data file, the VARNAMES command lists the order of the variables in the data file, and MISSING specifies a common missing value code for all incomplete variables. The MODEL command specifies a Level-2 identifier variable (the variable to the left of the tilde), the variables in the imputation model (the list to the right of the tilde), and any random associations between pairs of Level-1 variables (two or more variables joined by a colon). Variables

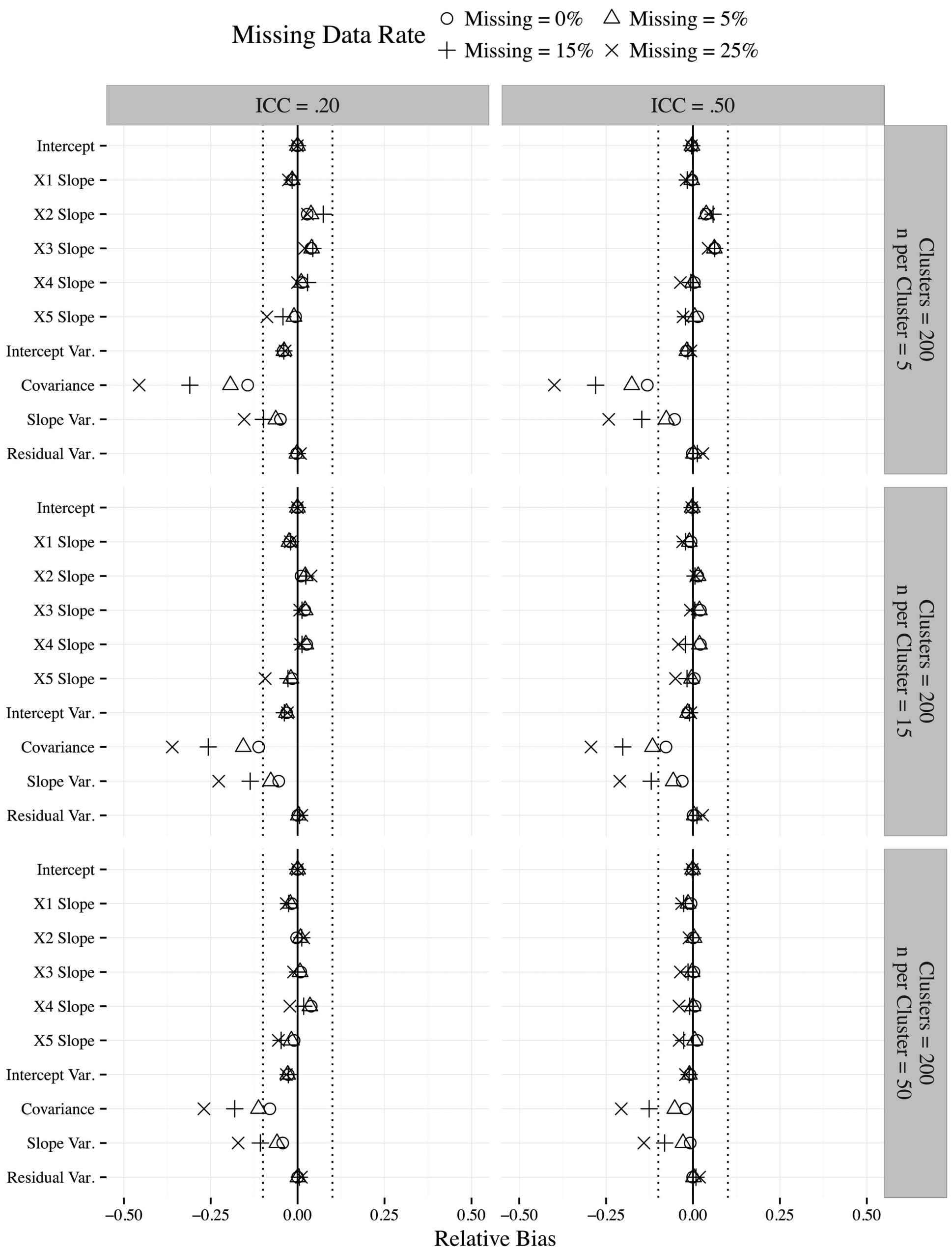

Figure 4. Average relative bias values for design cells with j = 200 clusters. Relative bias is defined as the difference between an average estimate and the true value expressed as a proportion of the true value. The dashed lines represent bias values of ±0.10.

listed on the MODEL command are automatically defined as continuous (normal) unless the user lists the variables on the ORDINAL or NOMINAL lines. The MODEL command automatically introduces random intercepts for all Level-1 variables, and random slopes are specified by joining two or more Level-1 variables with a colon (e.g., "y:x1" specifies a random slope). NIMPS gives the desired number of imputations, BURN and THIN are algorithmic options that determine the burn-in and thinning (i.e., between-imputation) intervals, respectively, and SEED provides a random number seed. In our example, NIMPS = 20 requests 20 data imputed sets, and the BURN = 1,000 and THIN = 1,000 options instruct the program to save the first data set after the 1,000th computational cycle and subsequent data sets every 1,000th cycle thereafter. Finally, the OUTFILE command gives the file path for the text file(s) containing the imputations, and the OPTIONS command specifies a number of miscellaneous computational and output preferences. In our example, the "prior1" keyword specifies the prior distribution for variance and covariance parameters (see the technical document from the online supplemental material), "hov"

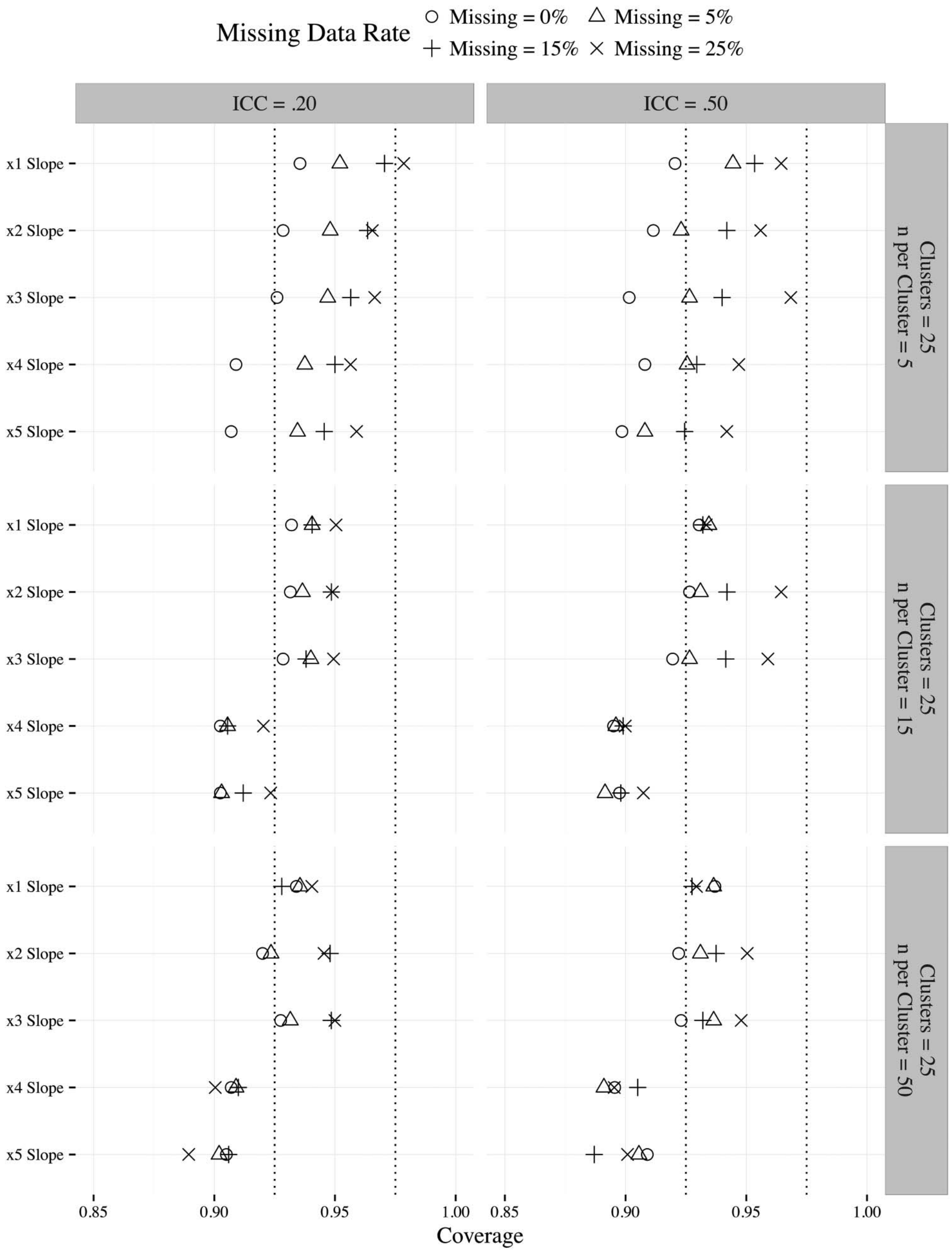

Figure 5. 95% confidence interval coverage for design cells with J = 25 clusters. The solid line at .95 represents the nominal value, and the dashed lines at .925 and .975 represent Bradley's (1978) liberal criterion.

invokes homogeneous within-cluster residual variances, "separate" saves imputed data sets to separate text files (e.g., for analysis with M*plus*), "psr" requests a table of potential scale reduction factors (Gelman & Rubin, 1992), and "clmean" introduces cluster means in the imputation model, as in Equation (18). As noted previously, all facets of imputation can also be specified using a graphical interface that bypasses the need for syntax.

Blimp is written in C++ and is provided as an optimized compiled executable for Mac and Windows operating systems. This architecture makes the program substantially faster than an R package, for example. To provide some rough benchmarks, we generated 50 imputations for two data sets from the simulation study. With 25% missing data on every variable, a 2014 iMac took approximately 22 s to complete imputation with N = 125 observations (25 clusters and five observations per cluster), and it took roughly 7 min to complete imputation with N = 10,000 observations (200 clusters with 50 cases per cluster). These runtimes put Blimp on par with commercial packages such as M*plus*.

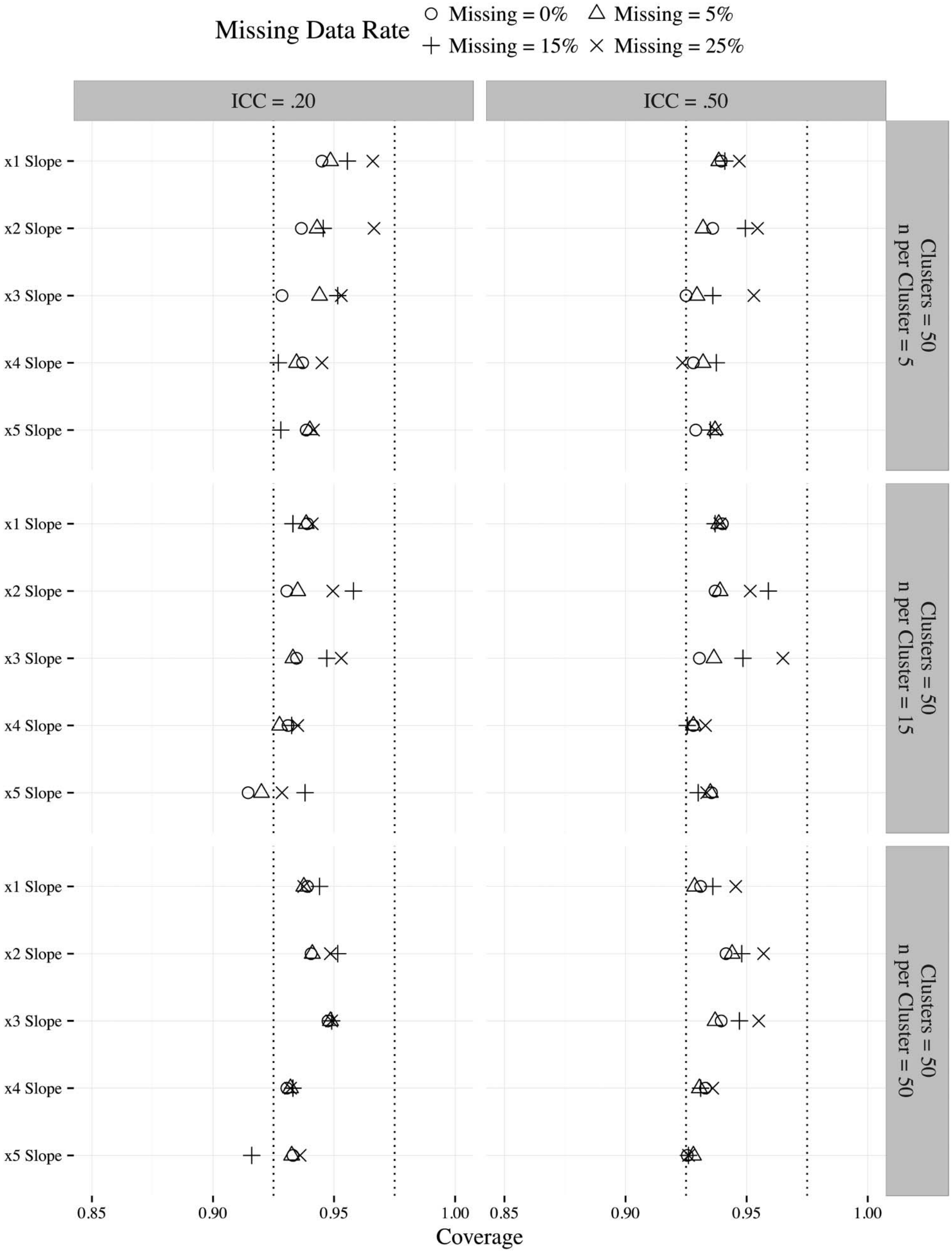

Figure 6. 95% confidence interval coverage for design cells with J = 50 clusters. The solid line at .95 represents the nominal value, and the dashed lines at .925 and .975 represent Bradley's (1978) liberal criterion.

Discussion

Multiple imputation is an MAR-based approach that has enjoyed widespread use in a variety of disciplines. The joint model and FCS are the predominant imputation frameworks for single-level data, and both have multilevel extensions. The multilevel imputation literature is still relatively nascent, and existing imputation routines are diverse and offer different functionality; all approaches can readily accommodate basic random intercept analyses with normally distributed variables, but they differ in their ability to handle random slopes, categorical variables, different within- and between-cluster covariance matrices, and incomplete Level-2 variables (Enders et al., 2016). This article outlined an FCS imputation approach that can accommodate these common analysis features. Our simulation results suggest that FCS gives good performance across a variety of conditions that are typical of behavioral science data. In virtually conditions that we examined, regression coefficients were relatively free of bias, even in small samples with a large proportion of missing data. Random effect estimates were somewhat mixed, however. Intercept variance esti-

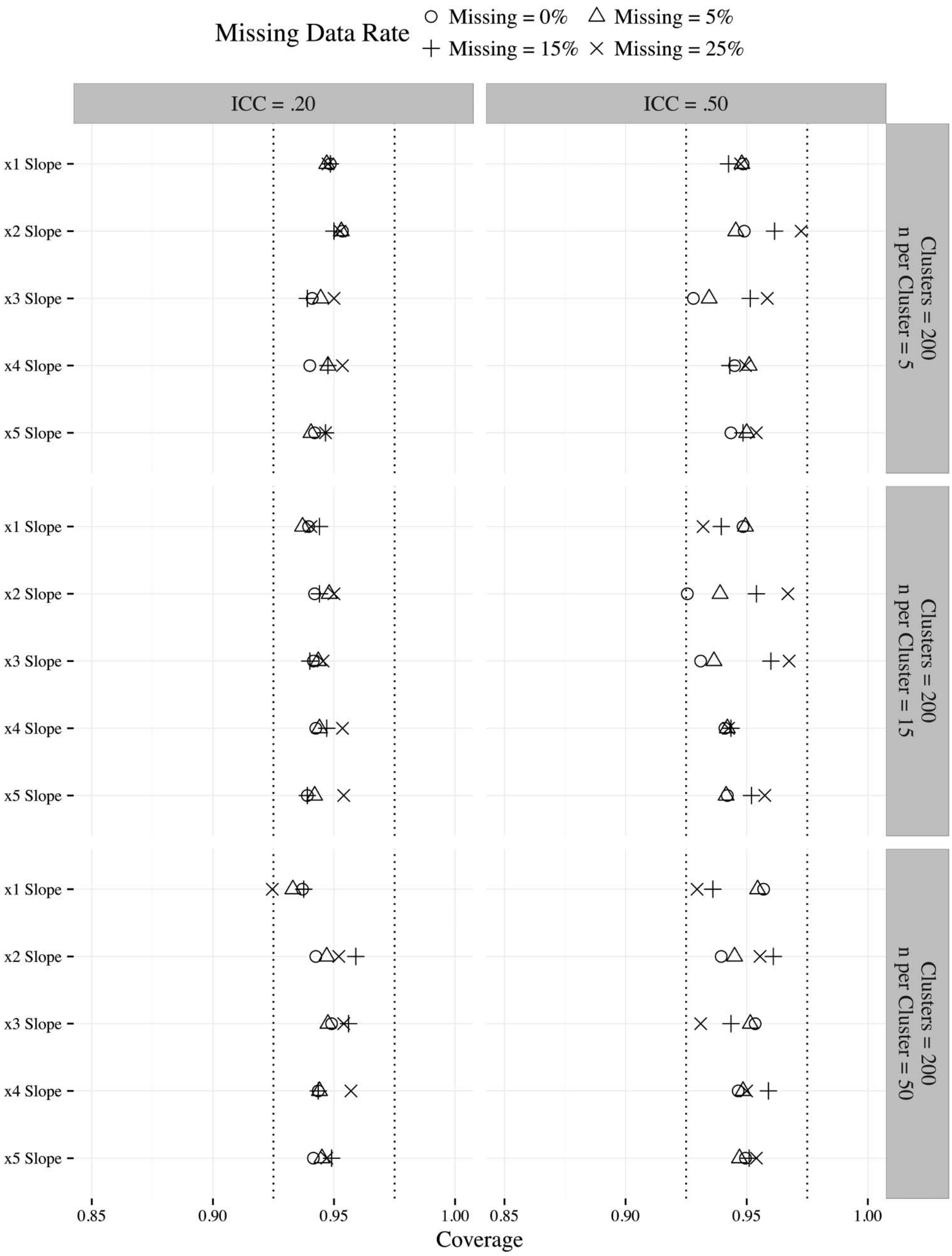

Figure 7. 95% confidence interval coverage for design cells with J = 200 clusters. The solid line at .95 represents the nominal value, and the dashed lines at .925 and .975 represent Bradley's (1978) liberal criterion.

mates were generally accurate and were unaffected by the missing data rate. Slope variance estimates, on the other hand, were often too low. A 15% missing data rate appeared to be a tipping point where slope variance estimates began to exhibit biases exceeding 10%, particularly when the within-cluster sample size was small.

Our work developing and testing FCS imputation allows us to offer a number of practical recommendations for researchers. In the context of single-level imputation, the literature often suggests that a single set of well-conceived imputations can serve as input data for a wide variety of statistical analyses. Given the complexities of multilevel data, we recommend that researchers limit their focus to a single analysis or a small family of analyses when generating imputations. Employing parsimonious imputation models mitigates computational problems that can arise with large numbers of random effects (Schafer, 2001), and it avoids an excessive number of Level-2 variables (Level-2 imputation requires fewer variables than clusters). Related to model complexity, we recommend that researchers perform preliminary analyses to determine which variables in a particular analysis family require random slopes. Although our procedure can accommodate more than one random slope predictor, we expect the Gibbs sampler algorithm to experience computational problems if the number of random associations is too large (Schafer, 2001). To simplify imputation, researchers could first estimate models with listwise deletion, retaining random slopes that reach some liberal significance criterion (e.g., $p < .20$); because these tests are exploratory, approximate probability values from standard Wald z tests can be used for this purpose, or researchers can use more appropriate mixture-based chi-square tests (Molenberghs & Verbeke, 2004; Savalei & Kolenikov, 2008). Finally, we strongly encourage researchers to examine convergence diagnostics prior to creating a set of imputations for analysis. In our experience, even relatively simple models can require very long burn-in periods (e.g., several hundred, perhaps 1,000 or more iterations) in order for the MCMC algorithm to achieve stationarity. Currently, our software implements Asparouhov and Muthén's (2010) modification of the Gelman and Rubin (1992) potential scale reduction factor, and we recommend that researchers examine PSR values from two or more long chains (e.g., 2,000 or more iterations) prior to generating imputed data sets.

Although our preliminary simulation results are promising, a great deal of methodological work remains. First, interaction effects are often of interest in multilevel research, and our program currently requires users to treat product terms like any other incomplete variable (von Hippel, 2009). A growing body of methodological research has demonstrated that interactive effects are problematic for MAR-based missing data handling methods (Carpenter & Kenward, 2013; Enders, Baraldi, & Cham, 2014; Seaman, Bartlett, & White, 2012; Yuan & Savalei, 2014), and imputing product terms generally requires an MCAR mechanism (Carpenter & Kenward, 2013). Methodologists have recently developed FCS-based imputation routines that work well with interactive effects (Bartlett, Seaman, White, & Carpenter, 2015), and we hope to extend these procedures to the multilevel context in the future. Second, we limited our simulations to a normally distributed outcome and normally distributed random effects. Violating either of these assumptions is potentially problematic for multiple imputation (Yuan, Yang-Wallentin, & Bentler, 2012; Yucel & Demirtas, 2010). Nonnormal data are probably the norm in many behavioral research settings, so it is important for future studies to evaluate FCS with nonnormal continuous variables. The impact of nonnormality could be most pronounced on Level-2 imputation where the sample size is very small (Yuan et al., 2012). Third, all simulation studies necessarily lack generalizability, and ours is no different, as we chose to investigate a rather limited set of conditions and parameter values. For example, we limited our focus to medium effect sizes in the context of a traditional multilevel regression model, and we restricted our attention to categorical predictors because the literature has largely focused on continuous variables. In developing our imputation routine, we performed numerous simulation studies with different models (e.g., random intercepts, random slopes) and different configurations of variables (e.g., all continuous, mixtures of categorical and continuous). The results from these test simulations were largely consistent with those reported here, and summaries are available upon request. One difference from the simulations here is that regression coefficients for continuous Level-2 predictors tend to exhibit mild biases when the number of clusters is small and the percentage of missing data is large (e.g., $J = 25$ with 25% missing data can produce bias values of 10%–15%). This bias is consistent with published studies on single-level imputation with small sample sizes (Yuan et al., 2012). Nevertheless, future studies should examine different analysis models (e.g., multilevel structural equation models), different data structures (e.g., dyadic data structures), different configurations of random effects (e.g., more than one random slope, smaller or larger intraclass correlations), and different effect sizes, to name a few.

In sum, multiple imputation has a long history in the methodological literature, but its extension to multilevel data is more recent. Given the limitations associated with existing imputation routines, our goal was to develop and test an imputation procedure that can accommodate a wide range of complexities that are typical of behavioral science data. Our computer simulations suggest that the FCS approach has good performance across many scenarios, but a great deal of methodological work is needed to advance this important topic.

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Aitchison, J., & Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika, 57,* 253–262. http://dx.doi.org/10.1093/biomet/57.2.253

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88,* 669–679. http://dx.doi.org/10.1080/01621459.1993.10476321

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage. http://dx.doi.org/10.4135/9781412985079

Allison, P. D. (2005, April). *Imputation of categorical variables with PROC MI*. Paper presented at the SAS Users Group International, Philadelphia, PA.

Asparouhov, T., & Muthén, B. (2010). *Multiple imputation with Mplus*. Retrieved from http://www.statmodel.com/download/Imputations7.pdf

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

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine, 26,* 1368–1382. http://dx.doi.org/10.1002/sim.2619

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

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15,* 391–420. http://dx.doi.org/10.1007/s001800000041

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software, 45,* 1–14. http://dx.doi.org/10.18637/jss.v045.i05

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. West Sussex, UK: Wiley. http://dx.doi.org/10.1002/9781119942283

Clark, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research. *Sociological Methods and Research, 35,* 311–351.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cowles, K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing, 6,* 101–111. http://dx.doi.org/10.1007/BF00162520

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling, 21,* 239–252. http://dx.doi.org/10.1080/10705511.2014.882686

Dunn, E. C., Masyn, K. E., Jones, S. M., Subramanian, S. V., & Koenen, K. C. (2015). Measuring psychosocial environments using individual responses: An application of multilevel factor analysis to examining students in schools. *Prevention Science, 16,* 718–733. http://dx.doi.org/10.1007/s11121-014-0523-x

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

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

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21,* 222–240. http://dx.doi.org/10.1037/met0000063

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4,* 87–107. http://dx.doi.org/10.1080/10705519709540063

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (2nd ed., pp. 439–492). Charlotte, NC: Information Age.

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

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

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

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics, 32,* 252–286. http://dx.doi.org/10.3102/1076998606298042

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling, 9,* 173–197. http://dx.doi.org/10.1177/1471082X0800900301

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-4018-5

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

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,* 876–883. http://dx.doi.org/10.1093/oxfordjournals.aje.a009904

Harker, H., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15,* 177–199. http://dx.doi.org/10.1076/sesi.15.2.177.30432

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29,* 60–87. http://dx.doi.org/10.3102/0162373707299706

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57,* 229–232. http://dx.doi.org/10.1198/0003130032314

Huang, F. L., & Cornell, D. G. (2015). Using multilevel factor analysis with clustered data: Investigating the factor structure of the positive values scale. *Journal of Psychoeducational Assessment, 34,* 3–14. http://dx.doi.org/10.1177/0734282915570278

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology, 14,* 28. http://dx.doi.org/10.1186/1471-2288-14-28

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

Kaplan, D. (1988). The impact of specification error on the estimation, testing and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 69–86. http://dx.doi.org/10.1207/s15327906mbr2301_4

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 23,* 93–116. http://dx.doi.org/10.3102/10769986023002093

Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology, 48,* 339–348. http://dx.doi.org/10.1037/0022-3514.48.2.339

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage. http://dx.doi.org/10.4135/9781849209366

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika, 69,* 101–122. http://dx.doi.org/10.1007/BF02295842

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119013563

Longford, N. (1989). Contextual effects and group means. *Multilevel Modelling Newsletter, 1,* 5.

Lüdtke, O., Koller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish–little-pond effect. *Contemporary Educational Psychology, 30,* 263–285. http://dx.doi.org/10.1016/j.cedpsych.2004.10.002

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16,* 444–467. http://dx.doi.org/10.1037/a0024376

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,* 203–229. http://dx.doi.org/10.1037/a0012869

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-71265-9

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 86–92. http://dx.doi.org/10.1027/1614-2241.1.3.86

Martin, A. J., Malmberg, L.-E., & Liem, G. A. D. (2010). Multilevel motivation and engagement: Assessing construct validity across students and schools. *Educational and Psychological Measurement, 70,* 973–989. http://dx.doi.org/10.1177/0013164410378089

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23,* 750–773. http://dx.doi.org/10.1080/10705511.2016.1186549

McNeish, D., & Stapleton, L. M. (2016a). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51,* 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008

McNeish, D. M., & Stapleton, L. M. (2016b). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28,* 295–314. http://dx.doi.org/10.1007/s10648-014-9287-x

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32,* 83–104. http://dx.doi.org/10.1016/j.cedpsych.2006.10.006

Mistler, S. A. (2013, August). *A SAS macro for applying multiple imputation to multilevel data*. Paper presented at the SAS Global Forum, San Francisco, CA.

Mistler, S. A., & Enders, C. K. (2016). *A comparison of joint model and fully conditional specification imputation for multilevel missing data*. Manuscript submitted for publication.

Molenberghs, G., & Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Statistica Sinica, 14,* 989–1020.

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354. http://dx.doi.org/10.1111/j.1745-3984.1991.tb00363.x

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

Raudenbush, S. W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 207–236). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-73186-5_5

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

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84,* 126–136. http://dx.doi.org/10.1207/s15327752jpa8402_02

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing, 5,* 121–125. http://dx.doi.org/10.1007/BF00143942

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9780470316696

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13,* 150–170. http://dx.doi.org/10.1037/1082-989X.13.2.150

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall. http://dx.doi.org/10.1201/9781439821862

Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355–377). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-012

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571. http://dx.doi.org/10.1207/s15327906mbr3304_5

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

Scott, M. A., Shrout, P. E., & Weinberg, S. L. (2013). Multilevel model notation—Establishing the commonalities. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The Sage handbook of multilevel modeling* (pp. 21–38). Newbury Park, CA: Sage. http://dx.doi.org/10.4135/9781446247600.n2

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. http://dx.doi.org/10.1186/1471-2288-12-46

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35,* 26–53. http://dx.doi.org/10.3102/1076998609345252

Simons, J. S., Wills, T. A., & Neal, D. J. (2014). The many faces of affect: A multilevel model of drinking frequency/quantity and alcohol dependence symptoms among young adults. *Journal of Abnormal Psychology, 123,* 676–694. http://dx.doi.org/10.1037/a0036926

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329. http://dx.doi.org/10.1037/1082-989X.6.4.317

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

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" Software Version 3.0*. Retrieved from http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of Frequentist and Bayesian approaches. *American Journal of Political Science, 57,* 748–761. http://dx.doi.org/10.1111/ajps.12001

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65,* 272–296. http://dx.doi.org/10.1177/0013164404268667

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16,* 219–242. http://dx.doi.org/10.1177/0962280206074463

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

van Buuren, S. (2012). *Flexible imputation of missing data*. New York, NY: Chapman & Hall. http://dx.doi.org/10.1201/b11826

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76,* 1049–1064. http://dx.doi.org/10.1080/10629360600810434

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2014). *Package "mice."* Retrieved from cran.r-project.org/web/packages/mice/mice.pdf

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

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies to ordinal missing data for Likert scale variables. *Multivariate Behavioral Research, 50,* 484–503. http://dx.doi.org/10.1080/00273171.2015.1022644

Yuan, K.-H., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution based MLE: The role of auxiliary variables. *Journal of Multivariate Analysis, 124,* 353–370. http://dx.doi.org/10.1016/j.jmva.2013.11.006

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distributional conditions. *Sociological Methods & Research, 41,* 598–629. http://dx.doi.org/10.1177/0049124112460373

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transaction A, 366,* 2389–2403.

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

Yucel, R. M., & Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis, 54,* 790–801. http://dx.doi.org/10.1016/j.csda.2009.01.016

Yucel, R. M., He, Y., & Zaslavsky, A. M. (2008). Using calibration to improve rounding in imputation. *The American Statistician, 62,* 1–5. http://dx.doi.org/10.1198/000313008X300912

Received November 16, 2015
Revision received February 21, 2017
Accepted March 11, 2017 ■