Psychological Methods
2016, Vol. 21, No. 2, 222–240

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000063

Multilevel Multiple Imputation: A Review and Evaluation of Joint Modeling and Chained Equations Imputation

Craig K. Enders, Stephen A. Mistler, and Brian T. Keller
Arizona State University

Although missing data methods have advanced in recent years, methodologists have devoted less attention to multilevel data structures where observations at level-1 are nested within higher-order organizational units at level-2 (e.g., individuals within neighborhoods; repeated measures nested within individuals; students nested within classrooms). Joint modeling and chained equations imputation are the principal imputation frameworks for single-level data, and both have multilevel counterparts. These approaches differ algorithmically and in their functionality; both are appropriate for simple random intercept analyses with normally distributed data, but they differ beyond that. The purpose of this paper is to describe multilevel imputation strategies and evaluate their performance in a variety of common analysis models. Using multiple imputation theory and computer simulations, we derive 4 major conclusions: (a) joint modeling and chained equations imputation are appropriate for random intercept analyses; (b) the joint model is superior for analyses that posit different within- and between-cluster associations (e.g., a multilevel regression model that includes a level-1 predictor and its cluster means, a multilevel structural equation model with different path values at level-1 and level-2); (c) chained equations imputation provides a dramatic improvement over joint modeling in random slope analyses; and (d) a latent variable formulation for categorical variables is quite effective. We use a real data analysis to demonstrate multilevel imputation, and we suggest a number of avenues for future research.

Keywords: missing data, multiple imputation, multilevel modeling

During the last three decades, the methodological literature has advocated for missing data handling approaches that assume a missing at random (MAR) missing data mechanism, whereby the propensity for missing data on a variable y is related to other variables in the analysis (or imputation) model but not to the would-be values of y itself (Enders, 2010; Little & Rubin, 2002; Rubin, 1976; Schafer & Graham, 2002). Multiple imputation and maximum likelihood estimation are the principal MAR-based procedures, and both are widely available in software packages. Multiple imputation creates several copies of the data (e.g., $m = 20$), each of which contains a different set of plausible replacement values, whereas maximum likelihood estimates model parameters directly from the observed data. Relative to traditional missing data handling techniques (e.g., excluding incomplete cases from an analysis), these approaches can reduce nonresponse bias and increase power (Enders, 2010; Schafer & Graham, 2002).

Multiple imputation has been widely available for many years, and numerous accessible discussions of this method appear in the literature (Enders, 2010; Graham, 2012; Schafer, 1997; Schafer & Graham, 2002; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001; van Buuren, 2012). However, familiar imputation routines such as Schafer's (1997) data augmentation algorithm are inappropriate for multilevel data because they ignore clustering and generate imputations from a model that assumes zero intraclass correlations. Perhaps not surprisingly, applying single-level imputation methods to multilevel data can introduce substantial bias, even when MAR is satisfied. Specialized imputation routines for multilevel data are now widely available in a variety of software packages, including Mplus (Muthén & Muthén, 1998–2012), the PAN and MLMMM packages in R (Schafer, 2001; Schafer & Yucel, 2002), the MICE package in R (van Buuren & Groothuis-Oudshoorn, 2011), MLwiN and Stata (Carpenter, Goldstein, & Kenward, 2011), and SAS (Mistler, 2013), among others. However, our experience suggests that many behavioral science researchers are unfamiliar with these techniques because much of the literature to date appears in specialized technical journals.

Joint modeling and chained equations imputation (also known as fully conditional specification and sequential regression imputation) are the principal imputation frameworks for single-level data, and both have multilevel counterparts (Schafer, 2001; Schafer & Yucel, 2002; van Buuren, 2011, 2012; Yucel, 2008). These approaches differ algorithmically and in their functionality; both are appropriate for a two-level random intercept analysis with normally distributed data, but they differ beyond that. First, the joint model can accommodate analyses that posit distinct within- and between-cluster relations (e.g., a contextual effects model that

This article was published Online First December 21, 2015.

Craig K. Enders, Stephen A. Mistler, and Brian T. Keller, Department of Psychology, Arizona State University.

Craig K. Enders and Brian T. Keller are now at the University of California, Los Angeles. Stephen A. Mistler is now at SAS, Cary North Carolina.

Correspondence concerning this article should be addressed to Craig K. Enders, University of California, Los Angeles, Department of Psychology, Box 951563, Los Angeles, CA, 90095-1563. E-mail: cenders@psych.ucla.edu

includes a level-1 predictor and its cluster means, a multilevel structural equation model with different path values at level-1 and level-2), but chained equations imputation currently assumes common regressions at both levels. Second, the joint model is usually restricted to random intercept analyses, whereas chained equations imputation is well suited for random slopes. Third, the joint model can accommodate incomplete categorical variables, whereas chained equations imputation is currently restricted to normally distributed variables. Finally, the methods differ in how they deal with incomplete variables at level-2.

Because neither imputation framework addresses the full range of complexities that are typical of behavioral science data, it is important to understand the strengths and limitations of each approach. Table 1 provides an advanced organizer for the paper, summarizing the major features of the joint model and chained equations imputation. The primary purpose of this paper is to describe these strategies and evaluate their performance. Specifically, we consider four missing data handling techniques: single-level imputation (Schafer, 1997); single-level imputation with dummy coded level-2 units (i.e., fixed effect imputation; Andridge, 2011; Reiter, Raghunathan, & Kinney, 2006); multilevel joint imputation (Asparouhov & Muthén, 2010; Schafer, 2001; Schafer & Yucel, 2002); and multilevel chained equations (van Buuren, 2011, 2012). We use analytic work and computer simulations to study these imputation approaches, deriving results that support the major conclusions from Table 1.

The organization of the paper is as follows. We begin with a brief overview of the multilevel model and Bayesian estimation, as this provides the groundwork for understanding multilevel imputation schemes. Second, we briefly discuss an ad hoc method that uses single-level imputation with dummy codes to preserve cluster-level variation. Third, we outline the primary multilevel imputation frameworks, joint modeling and chained equations imputation. We describe these methods in the context of normally distributed data, and we use simulation studies to examine their differences. Fourth, we describe strategies for dealing with incomplete categorical variables, with an emphasis on latent variable imputation. Finally, we use a real data analysis to illustrate joint imputation and chained equations.

Our paper makes a number of contributions to the literature. As noted previously, our experience suggests that researchers are generally unfamiliar with multilevel imputation. This is not surprising, as the multilevel imputation literature is relatively technical, focusing on algorithmic development rather than application. Comprehensive comparisons of multilevel imputation procedures are virtually nonexistent, and contemporary missing data texts generally devote only a few paragraphs to the topic. This paper provides researchers with an accessible review of existing multilevel imputation routines, so that they can choose an imputation procedure that best suits their analytic goals. Further, whereas joint modeling and chained equations are often equivalent with single-level data (Hughes et al., 2014), these methods can be quite different in the multilevel context, even with normally distributed variables; in fact, important differences exist among software packages that implement joint imputation. We know of no work that systematically describes and illustrates the subtle differences among these competing methods. Finally, the nascent literature on multilevel imputation is surprisingly devoid of empirical studies comparing joint modeling to chained equations imputation, and the simulation studies that we describe later in the paper are the first to systematically compare these approaches.

Multilevel Models and Bayesian Estimation

To establish the necessary background information, this section provides a brief overview of multilevel models and Bayesian estimation. In particular, we focus on two-level random intercept and random slope analysis models with a pair of level-1 predictors, x and z. We restrict our attention to level-1 variables because this situation has received the most attention in the multilevel imputation literature. Although these analysis models are relatively simple, they allow us to demonstrate key differences between the joint model and chained equations imputation, and the conclusions that we derive readily generalize to more complex analyses and modeling frameworks (e.g., multilevel structural equation models).

Table 1
Summary of Multilevel Imputation Model Features

Feature	Imputation model		
	Joint model (JM-SY)	Joint model (JM-AM)	Chained equations
Random intercepts	Yes	Yes	Yes
Random slopes	Complete variables only	No	Yes
Contextual effects	Yes for incomplete variables. User must introduce the cluster means of complete predictors	Yes	No for incomplete variables. User must introduce cluster means of complete predictors
Level-2 variables	User may need to employ separate imputation routines at each level	Yes	Yes for SAS macro. MICE requires separate imputation routines at each level
Categorical variables	Yes	Yes	No. Requires post-imputation rounding of continuous imputes
Software	Mplus, MLwiN, PAN, MLMMM, Stata, SAS macro	Mplus	MICE, SAS macro

Note. JM-SY and JM-AM are the joint models from Schafer and Yucel (2002) and Asparouhov and Muthén (2010), respectively.

Using notation from Scott, Shrout, and Weinberg (2013), the random intercept model is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + \varepsilon_{ij} \quad (1)$$

and the corresponding random slope model is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}. \quad (2)$$

In both models, y_{ij} is the outcome score for observation i in cluster j, and x_{ij} and z_{ij} are the corresponding predictors. Turning to the fixed effects, β_0 is the intercept, and β_1 and β_2 are slope coefficients for x and z, respectively. At level-2, u_{0j} is a residual that captures between-cluster intercept differences, and the u_{1j} residual in the random slope model allows the association between x and y to vary across clusters. Finally, ε_{ij} is a level-1 residual that captures within-cluster residual variation. The linear model typically assumes that the level-2 residuals are multivariate normal with zero means and a covariance matrix $\boldsymbol{\Sigma}_u$, and it assumes that the within-cluster residuals (i.e., the ε_{ij} terms) are normally distributed with a zero mean and constant variance σ^2_ε.

Throughout this paper, we follow Scott et al.'s (2013) notation because it naturally accommodates the matrix formulation from the mixed modeling literature. Because much of missing data literature uses matrix notation (e.g., Schafer, 2001; Schafer & Yucel, 2002; van Buuren, 2011; Yucel, 2008), we occasionally represent the previous models in matrix form as follows

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{Z}_j \mathbf{u}_j + \boldsymbol{\varepsilon}_j \quad (3)$$

where $\mathbf{y}_j$ is the vector of outcome scores for cluster j, $\mathbf{X}_j$ is the matrix of predictor variables (including a unit vector for the intercept), $\boldsymbol{\beta}$ is a column vector of regression coefficients, $\mathbf{Z}_j$ contains the level-1 variables in $\mathbf{X}_j$ that exert a random effect on the outcome (including a unit vector), $\mathbf{u}_j$ is a column vector that contains the random effects for cluster j (e.g., u_{0j} and u_{1j}), and $\boldsymbol{\varepsilon}_j$ is the column vector of level-1 residuals. As before, $\boldsymbol{\Sigma}_u$ and σ^2_ε quantify the dispersion of the residuals in $\mathbf{u}_j$ and $\boldsymbol{\varepsilon}_j$, respectively.

Bayesian estimation of multilevel models often employs an iterative computational algorithm known as the Gibbs sampler. The Gibbs sampler is a member of a broader class of simulation-based Markov Chain Monte Carlo (MCMC) methods that have enjoyed widespread use in recent years (Gelman et al., 2013; Jackman, 2000). Space limitations preclude a detailed description of the Gibbs sampler and Bayesian estimation more generally, so we provide a brief sketch of the major ideas. Detailed descriptions of the Gibbs sampler steps for multilevel models are widely available in the literature (Browne & Draper, 2000; Goldstein, Bonnet, & Rocher, 2007; Goldstein, Carpenter, Kenward, & Levin, 2009; Kasim & Raudenbush, 1998; Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008).

To illustrate the computational steps, reconsider the analysis models in Equations (1) and (2). The Bayesian paradigm views the parameters of these models as random variables having a distribution (a posterior). A joint distribution describes the relative probability of different combinations of parameter values, given the data. The Gibbs sampler breaks a complex estimation problem involving the joint distribution of the data and the parameters into a series of simpler computational steps, each of which focuses on a single unknown quantity. With a normally distributed outcome, a single iteration t of the Gibbs sampler cycles through four major computational steps: (a) sample level-2 residuals, (b) sample a level-2 covariance matrix, (c) sample regression coefficients, and (d) sample the within-cluster covariance matrix (or residual variance). More formally, the Gibbs sampling steps for a linear mixed model are

$$\begin{aligned}
\mathbf{u}^{(t)} &\sim \text{MVN}\left(\mathbf{u} \,|\, \mathbf{y},\ \boldsymbol{\Sigma}_u^{(t-1)},\ \boldsymbol{\beta}^{(t-1)},\ \boldsymbol{\Sigma}_\varepsilon^{(t-1)}\right) \\
\boldsymbol{\Sigma}_u^{(t)} &\sim \text{IW}\left(\boldsymbol{\Sigma}_u \,|\, \mathbf{y},\ \boldsymbol{\beta}^{(t-1)},\ \boldsymbol{\Sigma}_\varepsilon^{(t-1)},\ \mathbf{u}^{(t)}\right) \\
\boldsymbol{\beta}^{(t)} &\sim \text{MVN}\left(\boldsymbol{\beta} \,|\, \mathbf{y},\ \boldsymbol{\Sigma}_\varepsilon^{(t-1)},\ \mathbf{u}^{(t)},\ \boldsymbol{\Sigma}_u^{(t)}\right) \\
\boldsymbol{\Sigma}_\varepsilon^{(t)} &\sim \text{IW}\left(\boldsymbol{\Sigma}_\varepsilon \,|\, \mathbf{y},\ \mathbf{u}^{(t)},\ \boldsymbol{\Sigma}_u^{(t)},\ \boldsymbol{\beta}^{(t)}\right)
\end{aligned} \quad (4)$$

where MVN and IW denote the multivariate normal and inverse Wishart distributions, respectively. Note that, for a univariate model with a single outcome variable, the final sampling step would draw a within-cluster residual variance rather than a residual covariance matrix, the value of which could be obtained from either an inverse Wishart or inverse gamma distribution. We use a covariance matrix here because it is general enough to accommodate the multivariate model employed by the joint imputation framework, noting that $\boldsymbol{\Sigma}_\varepsilon^{(t)}$ reduces to σ^2_ε with chained equations imputation, which uses a univariate mixed model.

Notice that each sampling step from Equation (4) conditions on (i.e., treats as known) values from previous steps. For example, the level-2 residuals in the first step depend on parameter values from the previous iteration, the covariance matrix in the second step depends on the level-2 residuals from the previous step, regression coefficients and within-cluster covariance matrix from the previous iteration, and so on. Iterating these steps generates an empirical (posterior) distribution of each parameter. Multilevel imputation algorithms apply these same steps, but they augment the procedure with an additional step that draws imputations from a normal distribution, conditional on quantities from Equation (4). We focus on this imputation step for the remainder of the manuscript because it differentiates the joint model from chained equations imputation.

Single-Level Fixed Effect Imputation

Classic imputation approaches such as Schafer's (1997) data augmentation algorithm can produce substantial bias when applied to multilevel data because they generate filled-in values with no between-cluster variation (Andridge, 2011; Taljaard, Donner, & Klar, 2008; Reiter et al., 2006; van Buuren, 2011). Before the advent of specialized multilevel imputation routines, researchers could impute the data separately within each cluster (sample size permitting), or they could preserve between-cluster variation by dummy coding level-2 units and including the code variables in a single-level imputation scheme (this contrasts multilevel imputation procedures, which use random effects to preserve between-cluster variation). This section briefly describes the latter approach, so-called fixed effect imputation, and we later present simulation results that illustrate its performance.

To illustrate fixed effect imputation, reconsider the random intercept analysis model from Equation (1). To keep things simple, suppose that only x is incomplete. After sampling the necessary parameter values (e.g., from an MCMC algorithm such as the Gibbs sampler or data augmentation), a normal distribution generates imputations

$$x_{ij(\text{mis})} \sim \text{N}\left(\sum_{j=1}^{J} \beta_j d_j + \beta_{J+1} z_{ij} + \beta_{J+2} y_{ij},\ \sigma^2_\varepsilon\right) \quad (5)$$

where J is the number of clusters, d_j is the dummy code for cluster j, the β_j are the intercepts of the J clusters (we omit the standard regression intercept and instead include J dummy codes) capturing between-cluster mean differences, β_{J+1} and β_{J+2} represent the influence of z and y, respectively, and σ^2_ε is the within-cluster residual variance from the regression of x on the complete variables and dummy codes.[1] In words, Equation (5) says to draw imputations from a normal distribution with a mean and variance equal to a predicted score and within-cluster residual variance, respectively. An equivalent interpretation defines each imputation as the sum of a predicted value and a normally distributed residual term (i.e., an ε_{ij}), the variance of which is σ^2_ε. Importantly, the dummy variable coefficients preserve between-cluster variation in x, producing imputations with a nonzero intraclass correlation. Fixed effects imputation can also accommodate random slope analyses such as that in Equation (2) by introducing one or more sets of product variables to the imputation model. For example, introducing product terms involving d_j and y can preserve random slope variation when x is missing, and the products of d_j and x do the same when y is incomplete.

Although the fixed effect approach can yield accurate parameters estimates in some situations (Reiter et al., 2006), it has noteworthy limitations. For example, analytic work suggests that fixed effect imputation inflates sampling variance, particularly when the intraclass correlation is low (Andridge, 2011), and simulation studies suggest that the method may exhibit low coverage (van Buuren, 2011). From a practical perspective, fixed effect imputation is computationally demanding and often requires an enormous number of parameters. For example, suppose that a researcher applies the imputation model in Equation (5) to a sample of $J = 100$ clusters with $n_j = 20$ observations per cluster (a configuration that would be common in daily diary studies, e.g.). Schafer's (1997) data augmentation algorithm requires a sampling step that draws a mean vector and covariance matrix with 102 variables, or 5,355 parameters (102 means and 5,253 unique covariance matrix elements). Introducing product terms to the same scenario (e.g., to preserve random slope variation) requires 20,502 unique parameter values. Multilevel imputation is considerably more parsimonious because it preserves between-cluster variation using random effects and a level-2 covariance matrix; computationally, sampling the necessary residual terms is dramatically faster than sampling a high-dimensional covariance matrix. For computational reasons alone, we prefer specialized multilevel imputation routines to fixed effect imputation. Nevertheless, the procedure is worth mentioning because it has enjoyed some use in the substantive literature and is a historical precursor to the multilevel strategies we describe below.

Joint Imputation for Normally Distributed Variables

Schafer (1997) popularized joint imputation for single-level data, and methodologists have since extended the approach to the multilevel context (Asparouhov & Muthén, 2010; Goldstein et al., 2009; Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008). Joint imputation routines for multivariate normal variables are now widely available in statistical software, including Mplus (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), among others. This section describes two variants of joint imputation: a model proposed by Schafer and Yucel (Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008), and an alternative method outlined by Asparouhov and Muthén (2010). We henceforth refer to these methods as JM-SY and JM-AM, respectively. The term "joint model" stems from the fact that the incomplete variables are assumed to follow a common distribution. We restrict our attention to incomplete normal variables in this section and later show how the joint model can accommodate categorical variables by assuming an underlying normal latent trait for discrete responses.

JM-SY is based on a multivariate version of the linear mixed model from Equation (3), whereby incomplete level-1 variables serve as outcomes in the $\mathbf{Y}_j$ matrix, and complete variables function as predictors in $\mathbf{X}_j$ (level-1 or level-2) and possibly $\mathbf{Z}_j$ (level-1 only). To illustrate, reconsider the random intercept analysis model from Equation (1). Further, suppose that x and y have missing values and z is complete. After performing the sampling steps from Equation (4), the Gibbs sampler draws imputations from a multivariate normal distribution

$$\begin{matrix} x_{ij(\text{mis})} \\ y_{ij(\text{mis})} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + \beta_{1(x)} z_{ij} + u_{0j(x)} \\ \beta_{0(y)} + \beta_{1(y)} z_{ij} + u_{0j(y)} \end{bmatrix},\ \boldsymbol{\Sigma}_\varepsilon\right) \quad (6)$$

where $\beta_{0(x)} + \beta_{1(x)} z_{ij} + u_{0j(x)}$ and $\beta_{0(y)} + \beta_{1(y)} z_{ij} + u_{0j(y)}$ are the predicted values of x and y, respectively, for case i in cluster j, and $\boldsymbol{\Sigma}_\varepsilon$ is the level-1 residual covariance matrix. For clarity, we add alphanumeric subscripts to the parameters and level-2 residuals to emphasize that these quantities differ across incomplete variables. Like its single-level counterpart, the joint model essentially defines each replacement score as the sum of a predicted value and a normally distributed residual term (i.e., an ε_{ij} value). The key difference is that predicted values from the multilevel model include level-2 residual terms that preserve between-cluster variation in x and y. Because the Gibbs sampler treats $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\varepsilon$ as unstructured covariance matrices, the u_{0j} terms freely correlate, as do the within-cluster residuals. We discuss the practical implications of these correlations later in the section.

JM-AM treats all variables as outcomes, regardless of missing data pattern (Asparouhov & Muthén, 2010). The JM-AM imputation model for the current example is as follows

$$\begin{matrix} x_{ij(\text{mis})} \\ y_{ij(\text{mis})} \\ z_{ij} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + u_{0j(x)} \\ \beta_{0(y)} + u_{0j(y)} \\ \beta_{0(z)} + u_{0j(z)} \end{bmatrix},\ \boldsymbol{\Sigma}_\varepsilon\right) \quad (7)$$

where the level-2 covariance matrix determines the spread of the u_0 terms.[2] Notice that this model decomposes all variables into within- and between-cluster components, such that $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_u$ are unstructured level-1 and level-2 covariance matrices, respectively, rather than residual covariance matrices (i.e., all u_{0j} terms freely correlate, as do all ε_{ij} values).

[1] Because dummy codes preserve between-cluster variation, the residual variance from the fixed effect model estimates within-cluster variability. Thus, we continue using σ^2_ε to denote this quantity, as in Equations (1) to (3).

[2] Mplus denotes the models in Equations (6) and (7) as H0 and H1 imputation, respectively.

Beginning with Rubin's (1987) seminal work, methodologists have emphasized the importance of specifying an imputation model that preserves associations from the analysis model, as omitting important effects can bias parameter estimates toward zero, regardless of the missing data mechanism (Collins, Schafer, & Kam, 2001; Meng, 1994; Rubin, 1996; Schafer, 1997, 2003). When this important goal is met, the imputation and analysis models are said to be congenial. Joint modeling for single-level data can generate imputations for a wide range of statistical analyses because imputation typically assumes an unrestricted covariance matrix (Schafer, 1997). However, the same is not necessarily true for multilevel imputation, so it is important to consider whether the imputation model adequately preserves features of the analysis model.

The random intercept analysis in Equation (1) models two sources of variation, and it posits common regression slopes at level-1 and level-2 (i.e., no contextual effects; Raudenbush & Bryk, 2002). Both imputation models preserve level-1 associations, but they do so in a slightly different manner. The JM-SY model in Equation (6) uses regression coefficients to capture the relations between z and the incomplete variables, and it uses the residual covariance in $\boldsymbol{\Sigma}_\varepsilon$ to preserve the portion of the x and y relation not explained by z. In contrast, the JM-AM approach from Equation (7) uses an unstructured within-cluster covariance matrix to preserve these associations. Although not immediately obvious, both approaches implement a level-2 covariance structure that is more general than that of the analysis model, meaning that the imputation models include additional effects not present in the analysis. In particular, the correlated random effects (e.g., $u_{0j(x)}$ and $u_{0j(y)}$) allow between-cluster associations to vary independently of those at level-1. Consequently, JM-SY is congenial with a more complex contextual effects analysis that includes the x cluster means

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + \beta_3 \bar{x}_j + u_{0j} + \varepsilon_{ij} \tag{8}$$

and JM-AM is congenial with an analysis that additionally includes the z cluster means.[3] Although both imputation methods are inconsistent with the random intercept analysis in Equation (1), incompatibilities that arise from introducing additional effects not present in the analysis (e.g., imputing from a model that allows for contextual effects) are typically viewed as beneficial (Collins et al., 2001; Meng, 1994; Schafer, 2003). In this particular example, the cluster means effectively serve as auxiliary variables that appear in the imputation model but not the analysis model.

The examples thus far suggest that joint imputation can accommodate missing data in random intercept models, but the same is not necessarily true for random slope models. To illustrate, consider the random slope analysis in Equation (2). As before, suppose that x and y have missing values and z is complete. The JM-SY model in Equation (6) can incorporate random slopes for complete variables such as z, but it cannot preserve random associations among incomplete variables.[4] The JM-AM model in Equation (7) is strictly limited to random intercept analyses. Uncongeniality that arises from omitting effects that are present in the analysis is damaging and can introduce bias, even under an MCAR or MAR mechanism (Collins et al., 2001; Meng, 1994; Schafer, 2003). In this particular example, joint imputation preserves random intercept variation but ignores the random slope in the analysis model. Omitting this random association from the imputation model would attenuate the slope variance estimate, and other parameters could also be affected.

Thus far we have focused on incomplete level-1 variables because this problem has received the most attention in the literature and because software programs differ in their treatment of level-2 variables. As outlined in the original work, the JM-SY procedure does not readily accommodate incomplete level-2 variables, but applying separate imputation routines at each level can address this limitation (Gelman & Hill, 2007; Yucel, 2008).[5] Under this stepwise scheme, the user first applies single-level imputation (e.g., data augmentation; Schafer, 1997) to a cluster-level data set with J records. After the Gibbs sampler has converged, the researcher saves the imputations and uses the filled-in level-2 variables as predictors in a level-1 imputation model such as that in Equation (6). Iterating these steps is ideal because it allows the imputed values at one level to carry forward as complete-data predictors at the other level.

The JM-AM model readily accommodates missing data at level-2. To illustrate, suppose that an incomplete level-2 variable w is added to the random intercept analysis in Equation (1). The imputation model for this problem is as follows:

$$\begin{matrix} w_{j(\text{mis})} \\ x_{ij(\text{mis})} \\ y_{ij(\text{mis})} \\ z_{ij} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(w)} + u_{0j(w)} \\ \beta_{0(x)} + u_{0j(x)} \\ \beta_{0(y)} + u_{0j(y)} \\ \beta_{0(z)} + u_{0j(z)} \end{bmatrix}, \boldsymbol{\Sigma}_\varepsilon \right) \tag{9}$$

As before, correlations among the level-2 residuals preserve between-cluster associations, including those between w and the level-1 variables. Because w's variance is strictly at level-2, the variance and covariance terms for this variable are constrained to zero in $\boldsymbol{\Sigma}_\varepsilon$, but the level-1 relations are otherwise unrestricted.

Chained Equations Imputation for Normally Distributed Variables

Chained equations imputation is an alternative imputation strategy for single-level and multilevel data (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren, 2007, 2011, 2012; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Whereas joint imputation draws replacement values from a multivariate distribution, the chained equations algorithm cycles through incomplete variables one at a time, drawing imputations from a series of univariate conditional distributions. At each imputation step, missing values for a particular variable are drawn from a distribution that conditions on all other variables, including filled-in variables from a previous step. To illustrate, consider a set of $r = 1$ to q incomplete variables. The imputation scheme at iteration t is

[3] JM-SY and JM-AM are equivalent if cluster means of complete variables such as z are introduced as level-2 predictors in the JM-SY model.

[4] Yucel (2011) proposed a method that uses random level-1 covariance matrices, but software routines do not implement this approach.

[5] The variant of JM-SY implemented in MLwiN and Stata (Carpenter et al., 2011) can accommodate incomplete level-2 variables.

$$
\begin{aligned}
y_{1(mis)}^{(t)} &\sim \mathrm{P}\left(y_1 \mid y_2^{(t-1)}, \ldots, y_q^{(t-1)}, \mathbf{X}, \boldsymbol{\theta}_1\right) \\
y_{2(mis)}^{(t)} &\sim \mathrm{P}\left(y_2 \mid y_1^{(t)}, y_3^{(t-1)}, \ldots, y_q^{(t-1)}, \mathbf{X}, \boldsymbol{\theta}_2\right) \\
y_{r(mis)}^{(t)} &\sim \mathrm{P}\left(y_r \mid y_1^{(t)}, \ldots, y_{r-1}^{(t)}, y_{r+1}^{(t-1)}, \ldots, y_q^{(t-1)}, \mathbf{X}, \boldsymbol{\theta}_r\right) \\
y_{q(mis)}^{(t)} &\sim \mathrm{P}\left(y_q \mid y_1^{(t)}, y_2^{(t)}, \ldots, y_{q-1}^{(t)}, \mathbf{X}, \boldsymbol{\theta}_q\right)
\end{aligned}
\tag{10}
$$

where y_r is an incomplete variable, P denotes a probability distribution (e.g., a normal distribution), $\mathbf{X}$ is a set of complete variables, and $\boldsymbol{\theta}_r$ contains the parameters defining the distribution of replacement values for a particular variable (e.g., regression coefficients, residual terms, etc.). Gibbs sampler steps such as those from Equation (4) generate the necessary parameter values and residual terms for each of the q imputation models. Chained equations imputation is equivalent to the joint model in single-level data sets with multivariate normal variables (Hughes et al., 2014), but the methods are not necessarily equivalent in the multilevel context.

To illustrate chained equations imputation, reconsider the random intercept analysis model from Equation (1). As before, suppose that x and y have missing values and z is complete. Whereas joint imputation draws replacement values from a multivariate distribution, chained equations imputation draws values from univariate conditional distributions. The random intercept model requires two imputation steps. In the first step, the Gibbs sampler from Equation (4) generates the necessary parameter values and residual terms, and it draws the x imputations from a univariate normal distribution

$$
x_{ij(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(\beta_{0(x)} + \beta_{1(x)} y_{ij}^{(t-1)} + \beta_{2(x)} z_{ij} + u_{0j(x)},\ \sigma^2_{\varepsilon_j(x)}\right) \tag{11}
$$

where $y_{ij}^{(t-1)}$ denotes the filled-in y variable from the previous iteration. The incomplete variables switch roles in the second step, such that the imputed x variable becomes a predictor in the y imputation model. Another set of sampling steps generates the necessary parameter values and residual terms, and a normal distribution generates replacement values for y.

$$
y_{ij(\mathrm{mis})}^{(t)} \sim \mathrm{N}\left(\beta_{0(y)} + \beta_{1(y)} x_{ij}^{(t)} + \beta_{2(y)} z_{ij} + u_{0j(y)},\ \sigma^2_{\varepsilon_j(y)}\right) \tag{12}
$$

Whereas joint imputation assumes constant within-cluster variation, existing chained equations algorithms sample heterogeneous residual variances (Kasim & Raudenbush, 1998). van Buuren (2011) suggests that this modification can produce "a considerable advance in imputation quality" (p. 185) with incomplete predictor variables.

As with the joint model, it is important to consider whether chained equations imputation adequately preserves features of the analysis. Recall that joint imputation allows level-1 and level-2 relations to differ, such that the resulting imputations are consistent with a contextual effects analysis. In contrast, chained equations imputation assumes common associations at level-1 and level-2. To understand why this is the case, consider the model-implied covariance structure at level-2. Because separate Gibbs sampler steps generate $u_{0j(x)}$ and $u_{0j(y)}$, the chained equations approach effectively implements a block-diagonal residual covariance matrix, whereby level-2 residuals from different equations are conditionally independent. Consequently, between-cluster associations do not vary apart from the level-1 relations, and the resulting imputations are congenial with the random intercept analysis in Equation (1). Although it is possible to modify the chained equations algorithm to update the cluster means after each imputation step and include these aggregates in subsequent imputation steps (Carpenter & Kenward, 2013), software packages do not currently offer this option.[6]

Whereas the joint model cannot accommodate random slopes for pairs of incomplete variables, the chained equations approach is ideally suited for this situation. To illustrate, consider the random slope analysis in Equation (2). As before, suppose that x and y have missing values and z is complete. The imputation steps for this problem are as follows:

$$
\begin{aligned}
x_{ij(\mathrm{mis})}^{(t)} &\sim \mathrm{N}\left(\beta_{0(x)} + \beta_{1(x)} y_{ij}^{(t-1)} + \beta_{2(x)} z_{ij} + u_{0j(x)} + u_{1j(x)} y_{ij}^{(t-1)},\ \sigma^2_{\varepsilon_j(x)}\right) \\
y_{ij(\mathrm{mis})}^{(t)} &\sim \mathrm{N}\left(\beta_{0(y)} + \beta_{1(y)} x_{ij}^{(t)} + \beta_{2(y)} z_{ij} + u_{0j(y)} + u_{1j(y)} x_{ij}^{(t)},\ \sigma^2_{\varepsilon_j(y)}\right)
\end{aligned}
\tag{13}
$$

Importantly, the mean vectors now include level-2 slope residuals, so the resulting imputations should preserve the random association between x and y. Although we could find no published simulations that compare the two imputation frameworks with incomplete random slope predictors, we expect chained equations imputation to provide a substantial advantage over the joint model in this situation. The simulations in the next section investigate this issue.

Finally, the MICE implementation of chained equations does not automatically accommodate incomplete level-2 variables, but the program includes a suite of functions that allow the user to apply distinct imputation models to each level (van Buuren et al., 2014), as suggested by Gelman and Hill (2007) and Yucel (2008). The stepwise imputation scheme is identical to that of JM-SY: apply single-level imputation to a cluster-level data set with J records, then use the filled-in level-2 variables as predictors in a level-1 imputation routine. Iterating these steps is ideal because it allows the imputed values at one level to carry forward as complete-data predictors at the other level. For readers who do not use R, we provide a custom SAS macro that implements two-level chained equations imputation (see the Data Analysis Example section and the corresponding syntax in Appendix D). Unlike MICE, our program fully automates level-2 imputation.[7]

Simulation Study 1: Normally Distributed Variables

The previous sections highlight two important differences between the joint model and chained equations imputation: (a) the joint model allows associations between pairs of variables to differ at level-1 and level-2, whereas chained equations imputation assumes common relations; and (b) the joint model cannot accommodate random slopes for pairs of incomplete variables, whereas

[6] The model can accommodate contextual effects for complete variables such as z, but researchers must explicitly specify the cluster means as a level-2 predictor.

[7] At each iteration, our algorithm cycles through all level-1 and level-2 variables, using aggregates of the filled-in level-1 variables as predictors for level-2 imputation and filled-in level-2 variables as predictors for level-1 imputation. Yucel (2008) suggests this as an alternative to the stepwise approach.

chained equations can readily do so. To investigate the practical impact of these differences, we performed a series of simulation studies with three population models: the (a) random intercept analysis from Equation (1), (b) random slope analysis from Equation (2), and (c) contextual effects analysis in Equation (8). We chose parameters that mimic a daily diary study because these data structures are exceedingly common in the social science literature and because we wanted to focus on a scenario that maximally differs from single-level data. For all simulations, we set the intraclass correlations of x and y at .50 and defined z as a pure level-1 covariate. For the random intercept and random slope models, a single regression slope conveyed the influence of x, and we chose a value for the coefficient that reduced the y variance by 16% at both levels. The population model for the contextual effects analysis additionally included a slope for the cluster means of x. We again chose coefficients that reduced the variance by 16% at each level, setting the level-2 slope equal to the negative of the level-1 slope. Finally, we defined z as a pure level-1 variable that had a within-cluster correlation of .30 with x and y.

For each population model, we wrote a SAS IML program to generate 500 samples of $N = 2000$ with $J = 100$ clusters and $n_j = 20$ observations per cluster (e.g., 20 daily diary observations from 100 persons). We then manipulated the missing data mechanism (MCAR and MAR) and missing data rate (20% and 40% missingness on x and y), treating both as within-subjects factors to minimize simulation error (i.e., a common collection of complete data sets produced the replicates for each design cell). For the MCAR condition, we generated binary missing data indicators for x and y (0 = observed, 1 = missing) by sampling random numbers from a Bernoulli distribution with success probability equal to the missing data rate. For the MAR condition, we sampled missing data indicators from a Bernoulli distribution where the success probabilities increased with the values of z (the complete covariate). We used a logistic regression model to generate the predicted probabilities of missingness, choosing coefficients that produced a .50 correlation between z and the binary missing data indicators. Note that we deleted x and y independently, so either (or both) could be missing.

We investigated four missing data handling techniques: single-level imputation (Schafer, 1997), fixed effect imputation (Andridge, 2011; Reiter et al., 2006), multilevel joint imputation[8] (Asparouhov & Muthén, 2010; Schafer, 2001; Schafer & Yucel, 2002), and multilevel chained equations (van Buuren, 2011, 2012). To our knowledge, no simulation to date has compared these methods, much less compared them across the variety of models and conditions that we implement here. We used the MI procedure in SAS 9.3 to implement single-level and fixed effect imputation, Mplus 7 for joint imputation, and a custom Gibbs sampler for chained equations imputation (Keller & Enders, 2014). For each artificial data set, we generated 10 imputations with 2,000 burn-in and 500 thinning iterations; we chose the burn-in and thinning intervals after examining trace plots (Schafer, 1997) and potential scale reduction factors (Gelman & Rubin, 1992) from several data sets.

After generating imputations, we used Mplus 7 to estimate the analysis models in Equations (1), (2), and (8), and we used relative bias and confidence interval coverage to evaluate the missing data handling methods. Relative bias is a proportion computed as the difference between an average estimate and its true value relative to the true value. Methodologists have suggested that relative bias values less than ±.10 are acceptable (Finch, West, & MacKinnon, 1997; Kaplan, 1988). We used the average complete-data estimates as true values, thereby eliminating biases that might arise from complete-data maximum likelihood estimation (e.g., biased contextual effects; Lüdke, Marsh, Robitzsch, & Trautwein, 2011). Confidence interval coverage is the proportion of replications where the 95% confidence interval contained the true population parameter. If parameter estimates and standard errors are accurate, coverage values for an alpha level of .05 should equal .95. Values below .95 indicate elevated Type I error rates, whereas values greater than .95 reflect conservative inference.

Simulation 1 Results

Table 2 gives the results from the 40% MAR simulation, which is perhaps the most challenging set of conditions. Perhaps not surprisingly, decreasing the missing data rate produced smaller biases and better coverage, and the mechanism made no difference. Results for the remaining conditions are given in Appendix Tables F through H.

To begin, single-level imputation generally produced problematic biases and poor coverage.[9] With the exception of the contextual effects model, which produced large biases for all parameters, regression coefficients were relatively accurate. However, variance estimates exhibited large biases, with values between .60 and .80 (i.e., bias equal to 60% to 80% of the true value). This result is not surprising because single-level imputation ignores clustering and generates imputations from a model that assumes zero intraclass correlations. Fixed effect imputation produced mixed results across the three analysis models. The method was generally adequate for a random intercept analyses, but it attenuated slope variance estimates. Because the imputation model did not incorporate dummy variable product terms (this would have required more than 20,000 parameters), it was incapable of preserving the random association between x and y.

Turning to multilevel imputation, the results were consistent with the ideas we established in previous sections. Specifically, (a) joint modeling and chained equations imputation are appropriate (and equivalent) for random intercept analyses; (b) the joint model is superior for analyses that posit distinct within- and between-cluster associations (e.g., contextual effects); and (c) chained equations imputation provides a dramatic improvement over joint modeling in random slope analyses. To some extent, the simulation results are self-evident because they demonstrate a basic tenant of multiple imputation theory: the imputation model should preserve all associations in a given analysis model. Thus, it is no surprise that joint modeling attenuates the slope variance because it omits this random effect from imputation. Similarly, chained equations imputation attenuates a contextual effect (bias values were modest at about 10%) because it assumes common regressions at level-1 and level-2. Although the simulation results are consistent with statistical theory, they are useful for demonstrating the practical impact of specifying an imputation model that is uncongenial with

[8] JM-SY and JM-AM are equivalent in our simulations because both x and y are incomplete and z has only level-1 variation.

[9] With 500 replications, the simulation error for the coverage values is approximately $\sqrt{(.95)(.05)/500} = .01$.

Table 2
Simulation Results With Normally Distributed Variables, MAR Mechanism, 40% Missing

Parameter	True value	Single-level			Fixed effects			Joint model			Chained equations		
		Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis													
Intercept	.00	.00	NA	81.4%	.00	NA	95.8%	.00	NA	94.6%	.00	NA	94.4%
x slope	.34	.33	−.04	83.8%	.35	.02	90.6%	.34	−.01	92.2%	.35	.01	91.0%
z slope	.20	.20	.02	94.4%	.19	−.01	92.8%	.20	.01	94.6%	.20	.00	93.6%
Intercept variance	.83	.28	−.66	.0%	.89	.08	97.0%	.83	.00	94.2%	.83	.00	94.0%
Residual variance	.81	1.35	.68	.0%	.77	−.04	81.4%	.81	.00	95.8%	.81	.00	95.4%
Contextual effects analysis													
Intercept	.00	.00	NA	85.2%	.00	NA	97.2%	.00	NA	96.8%	.00	NA	96.8%
x slope	.34	.04	−.88	.0%	.35	.03	92.2%	.34	−.01	95.2%	.31	−.09	81.0%
z slope	.20	.29	.45	51.6%	.20	−.02	94.4%	.20	.00	94.2%	.21	.04	94.0%
Contextual slope	−.74	−.41	−.45	8.8%	−.72	−.03	97.0%	−.74	−.01	96.4%	−.67	−.10	90.6%
Intercept variance	.83	.30	−.64	.0%	.91	.09	96.8%	.83	.00	94.0%	.87	.05	97.0%
Residual variance	.80	1.53	.91	.0%	.77	−.04	82.4%	.81	.00	95.2%	.82	.02	93.2%
Random slope analysis													
Intercept	.00	.00	NA	81.4%	.00	NA	95.4%	.00	NA	95.2%	.00	NA	94.5%
x slope	.35	.34	−.01	87.4%	.36	.03	83.6%	.35	.00	85.4%	.35	.01	92.2%
z slope	.20	.20	.02	93.8%	.20	−.01	93.0%	.20	.01	94.8%	.20	.00	92.7%
Intercept variance	.86	.28	−.68	.0%	.89	.04	96.6%	.82	−.04	91.2%	.81	−.05	89.8%
Slope variance	.09	.02	−.80	.0%	.02	−.78	.0%	.02	−.80	.0%	.07	−.21	82.9%
Covariance	.00	−.02	NA	93.8%	.00	NA	99.4%	−.01	NA	99.8%	.00	NA	98.0%
Residual variance	.81	1.41	.75	.0%	.84	.04	85.4%	.88	.09	63.0%	.84	.04	89.2%

Note. MAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

the statistical analysis. To our knowledge, no existing studies have discussed congeniality in the context of multilevel models, nor have previous studies compared joint imputation and chained equations in these conditions.

Incomplete Categorical Variables

Thus far, we have considered imputation methods that assume multivariate normality. This is often a strong and unrealistic requirement for behavioral science data where mixtures of categorical and continuous variables are the norm. Researchers have at least four options for dealing with incomplete variables with mixed response types: (a) impute categorical variables as though they are normally distributed and round the continuous imputations to discrete values, (b) employ a chained equations imputation scheme based on the generalized linear mixed model, (c) use a joint imputation strategy that treats categorical variables as normally distributed latent variables, and (d) apply single-level fixed effect imputation with a chained equations algorithm that can accommodate categorical variables (e.g., MICE, BaBoon; van Buuren et al., 2014; Meinfelder, 2011). This section provides a brief overview of the first three strategies, and we subsequently restrict our attention to the latent variable formulation because this approach is available in software packages.

Perhaps the most straightforward way to deal with incomplete categorical variables is to simply impute them as though they are normally distributed. In the context of single-level data, applying a multivariate normal imputation model tends to work well with non-normal continuous variables (Demirtas, Freels, & Yucel, 2008), but this strategy is problematic for discrete variables because it yields fractional imputations (e.g., a dummy variable with an imputed value of −.23). Methodologists have proposed a variety of ad hoc rounding approaches (Allison, 2002, 2005; Bernaards, Belin, & Schafer, 2007; Horton, Lipsitz, & Parzen, 2003; Yucel, He, & Zaslavsky, 2008), but these routines often introduce substantial bias. Rounding is particularly problematic for nominal variables with more than two categories because imputation can produce a code of one on multiple dummy variables (Allison, 2002). Although methodologists have yet to examine rounding in the context of multilevel data, we do not expect this procedure to perform well in a broad range of situations.

A second strategy for incomplete categorical data is to generate imputations from a generalized linear model (or generalized linear mixed model in case of multilevel data), such that replacement values are drawn from a distribution that is tailored to the scale of each variable. Although this strategy is inherently incompatible with joint imputation, chained equations can address this problem because each imputation step can draw replacement values from a different distribution. For example, one step can use a linear mixed model to generate imputations from a normal distribution, the next step can employ a multilevel logistic regression to sample discrete imputations from a multinomial distribution, and so on. The primary disadvantage of this strategy is that the conditional distributions from which imputations are drawn are incompatible in the sense that their joint distribution may not exist. Although this feature makes it difficult to evaluate convergence (van Buuren, 2007, 2012), simulation studies of single-level data suggest that the resulting parameter estimates are accurate (Raghunathan et al., 2001; van Buuren et al., 2006).

To date, attempts to marry the generalized linear mixed model with chained equations imputation have met with limited success. For example, Zhao and Yucel (2009) examined chained equations imputation in a simple random intercept model with one continuous and one binary variable. The method worked well when the intraclass correlation was very close to zero but produced unacceptable coverage rates in other conditions (coverage values ranged between .40 and .80). Performance aside, the procedure is computationally intensive and prone to convergence failures because the Gibbs sampler requires an iterative optimization step that fits a linear mixed model to the filled-in data. Zhao and Yucel (2009) reported that convergence failures were common as the intraclass correlation increased, and our own attempts to apply chained equations imputation to a random intercept model with a binary outcome produced convergence failures over 40% of the time. Collectively, these findings cast doubt on the use of generalized linear mixed models for categorical variable imputation; if the simplest random intercept models produce estimation failures and poor coverage rates, it is unlikely that the method will work in realistic scenarios involving random slopes or complex mixtures of categorical and continuous variables. Given these difficulties, we provide no further discussion of this approach.

The third and perhaps most promising strategy for categorical imputation is to adopt a probit regression model that posits an underlying normal latent trait for each discrete variable (Johnson & Albert, 1999). This model defines a normal distribution of latent trait scores for each individual, the mean of which varies as a function of a set of predictor variables. In the case of binary and ordinal variables, one or more threshold parameters divide the continuous distributions into discrete categories. The latent variable formulation is convenient because it accommodates the following imputation strategy: (a) sample plausible latent variable scores from an underlying normal distribution; (b) perform the Gibbs sampler steps (including imputation) on the latent metric; and (c) convert the continuous imputes to discrete observed scores. From a computational perspective, latent variable imputation is straightforward because the Bayesian computational machinery applies standard Gibbs sampling steps for multivariate normal data. Latent variable imputation is currently available in the joint model framework and thus has the same limitations as joint imputation for normal variables (e.g., no random slopes for incomplete variables). The remainder of this section describes the latent variable approach in more detail, with an emphasis on binary and ordinal variables.

The latent variable formulation for ordinal (or binary) variables is based on a cumulative probit model that views categorical responses as arising from an underlying normal latent variable (Agresti, 2012; Finney & DiStefano, 2013; Johnson & Albert, 1999). We henceforth denote these latent variables with a star after the variable name (e.g., y^* is the underlying latent variable for a discrete y). A predicted value from a multilevel model defines the mean of each case's latent variable distribution, and the variance is fixed at one for identification. Discrete scores arise from one or more threshold cutoffs that divide the latent variable distribution into segments. For example, consider a binary variable y that takes on values of 0 and 1. A binary outcome requires a single threshold, such that a discrete score of $y = 1$ occurs when the measurement process yields a latent variable score above the threshold, and $y = 0$ occurs when y^* falls below the threshold. Figure 1 depicts the underlying latent variable distributions for a binary y variable at three values of x. The model readily accommodates ordinal responses by incorporating additional threshold parameters, such that the number of thresholds is one fewer than the number of response options (e.g., an ordinal questionnaire item with four response options requires three threshold parameters).

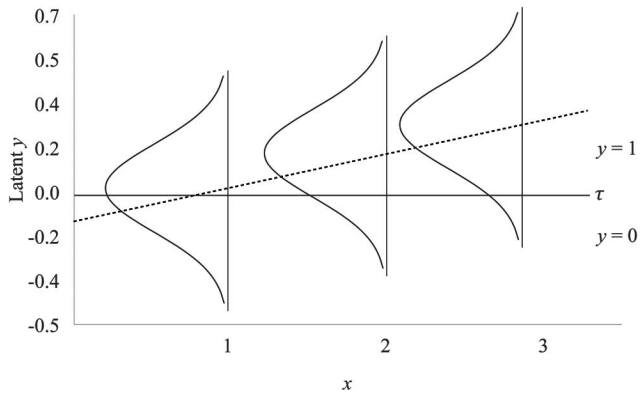

Figure 1. Latent variable distributions for a binary y variable at three values of x. A discrete score of y = 1 occurs when the measurement process yields a latent variable score above the threshold τ, and y = 0 occurs when the latent score falls below the threshold.

As mentioned previously, the latent variable formulation is attractive because it readily integrates with established multilevel imputation procedures for normally distributed variables. Because latent variable scores are completely missing, the procedures for sampling plausible latent variable scores and imputing missing values are actually one in the same. To illustrate the process, consider a multilevel logistic regression with random intercepts, where y is a binary outcome, and x and z are normally distributed predictors. Further, suppose that x and y have missing values and z is complete. With an incomplete categorical variable, the JM-AM imputation model from Equation (7) is as follows:

$$\begin{matrix} x_{ij(\text{mis})} \\ y^*_{ij(\text{mis})} \\ z_{ij} \end{matrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_{0(x)} + u_{0j(x)} \\ \beta_{0(y^*)} + u_{0j(y^*)} \\ \beta_{0(z)} + u_{0j(z)} \end{bmatrix}, \boldsymbol{\Sigma}_\varepsilon \right) \tag{14}$$

Note that the variance of y^* is fixed at unity in $\boldsymbol{\Sigma}_\varepsilon$ to provide a metric for the latent scores, but the model is otherwise identical to that for multivariate normal variables.

The sampling procedure that generates latent variable scores differs depending on whether y is observed or missing. For cases where y is observed, the sampling step restricts latent variable scores to a particular area of the distribution (e.g., for $y = 0$, the range of plausible y^* values falls in the area of the curve below the threshold parameter), whereas y^* scores for the incomplete cases are unbounded. Categorical imputes are then generated by comparing the sampled y^* to the threshold parameter(s) (e.g., imputed values of 0 and 1 are assigned to y^* values below and above the threshold, respectively). With categorical variables, the Gibbs sampler from Equation (4) requires an additional step that generates the necessary threshold parameters (Albert & Chib, 1993; Cowles, 1996), but the algorithm is otherwise the same, as the computational steps simply use the latent variable versions of each categorical variable.

The latent variable approach can accommodate nominal variables via the multinomial probit model (Agresti, 2015; Aitchison & Bennett, 1970; Albert & Chib, 1993), but this application is thus far limited to the MLwiN program (Goldstein et al., 2009). Briefly, this model defines a normally distributed latent variable (a so-called response utility) for each response option. Unlike the cumulative probit model, the multinomial model does not require threshold parameters. Rather, the relative magnitude of the underlying latent scores determines category membership. For example, reconsider a binary variable y with values of zero and one. In the multinomial framework, a discrete value of $y = 1$ occurs when the latent variable score for the high code is greater than the corresponding utility value for $y = 0$. We refer interested readers to work by Goldstein and colleagues for additional information on the multinomial probit model for missing data (Carpenter et al., 2011; Goldstein et al., 2007; Goldstein et al., 2009).

Simulation Study 2: Categorical Variables

To examine the performance of categorical imputation procedures, we again performed a series of simulation studies mimicking a typical daily diary study. To maintain consistency, we reused the artificial data sets from the previous simulations and converted x and y to categorical variables. We recoded x into an ordinal variable with category proportions of .10, .20, .40, .20, and .10, and we converted y into a binary variable with equal proportions. We examined two missing data handling procedures: joint imputation treating categorical variables as latent variables, and chained equations imputation treating categorical variables as normally distributed, followed by naïve rounding (i.e., rounding to the nearest observed score). Although methodologists have proposed sophisticated rounding schemes for categorical imputation, these approaches have not been extended to multilevel data and are not available in general use software packages. Consequently, we rounded the continuous imputes to the nearest observed score because we felt that most researchers would gravitate toward this procedure. For brevity, we restrict our attention to random intercept and random slope models with a binary outcome. We estimated multilevel logistic analysis models with Mplus 7, and we again used relative bias and confidence interval coverage to summarize the simulation results.

Simulation 2 Results

Table 3 gives the results from the 40% MAR simulation, and the results for the remaining conditions are given in Appendix Tables I through K. As before, decreasing the missing data rate produced somewhat lower biases and better coverage, and the mechanism made no difference. The joint model results are largely consistent with those from the simulation with normally distributed variables: The method yields accurate estimates and honest coverage rates with random intercept analyses but causes severe biases with random slope models. Again, the latter result is not surprising given that the joint model ignores level-2 slope variation. In the previous simulations, chained equations and joint imputation were equivalent in random slope analyses, but chained equations gave a substantial improvement with random slope models. With categorical variables, rounding the normal imputes to the nearest observed score proved detrimental, as fixed effect estimates were uniformly attenuated. Interestingly, rounding had less impact on level-2 variance estimates. Despite its bias, normal-theory chained equations imputation with rounding appears to be a better option than joint modeling for random slope models. However, it is important to note that a different configuration of category proportions could (and likely would) alter our results.

Data Analysis Example

This section uses data from a longitudinal investigation of a novel math problem-solving curriculum (Montague, Krawec, Enders, & Dietz, 2014) to illustrate multilevel imputation. Briefly, researchers assigned $J = 35$ schools to either an intervention (novel problem-solving curriculum) or a comparison condition (standard curriculum), with an average of $n_j = 30.71$ students per school ($N = 1{,}056$). To ensure that the joint model and chained equations are equally capable of preserving the associations in the data, we chose a random intercept analysis that evaluates the impact of the intervention on math problem-solving scores while

Table 3
Simulation Results With Categorical Variables, MAR Mechanism, 40% Missing

		Joint model			Chained equations		
Parameter	True value	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis							
Intercept	1.92	1.91	.00	96.0%	1.77	−.08	94.0%
x slope	.64	.64	.00	94.4%	.56	−.12	86.0%
z slope	.40	.40	.01	93.4%	.31	−.22	82.8%
Intercept variance	3.07	3.10	.01	93.0%	3.37	.10	88.4%
Random slope analysis							
Intercept	1.99	1.86	−.06	92.0%	1.88	−.05	94.6%
x slope	.66	.62	−.06	91.4%	.60	−.10	91.2%
z slope	.40	.39	−.02	94.4%	.31	−.22	82.2%
Intercept variance	5.54	3.40	−.39	64.6%	5.31	−.04	94.8%
Slope variance	.28	.05	−.83	15.2%	.28	.00	98.0%
Covariance	−.84	−.16	−.81	35.2%	−.70	−.17	95.2%

Note. MAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

controlling for five student-level covariates: a lunch assistance indicator, gender, standardized math test scores, math self-efficacy scale scores, and pretest problem-solving scores.

$$probsolv_{ij} = \beta_0 + \beta_1(lunch_{ij}) + \beta_2(gender_{ij}) + \beta_3(stanmath_{ij}) + \beta_4(efficacy_{ij}) + \beta_5(pretest_{ij}) + \beta_6(txgroup_{ij}) + u_{0j} + \varepsilon_{ij} \quad (15)$$

The gender, lunch assistance, and treatment indicators are complete, and the remaining variables have missing data rates between 2% and 23%.

We used Mplus to implement JM-SY and JM-AM, and we used MICE and a custom SAS macro to implement chained equations imputation. In addition to the analysis variables, the imputation model included academic ability (three nominal categories coded as two dummy variables) as an auxiliary variable. After examining convergence diagnostics, we generated 20 imputations (Graham, Olchowski, & Gilreath, 2007) from a Gibbs sampler that saved a data set after 1,000 iterations. We chose this interval after examining trace plots (Schafer, 1997) and potential scale reduction factors (Gelman & Rubin, 1992). Appendixes A and B give the Mplus inputs for JM-SY and JM-AM, respectively and Appendixes C and D give the MICE and SAS code for chained equations imputation. To keep the example simple, we use the Mplus analysis file from Appendix E to analyze the imputations and pool the estimates. For brevity, we do not discuss the syntax here but instead refer readers to Mplus (Asparouhov & Muthén, 2010; Muthén & Muthén, 1998–2012) and MICE (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren et al., 2014) reference materials. The SAS macro and its instructions are available at www.appliedmissingdata.com.

Table 4 gives the parameter estimates and standard errors from the data analysis example. Perhaps not surprisingly, the imputation methods produced similar estimates, with the largest differences equal to about a fifth of a standard error unit, on average. This result is expected given that both imputation frameworks are equally capable of preserving effects from a random intercept analysis. It is important to reiterate that methods differ in their ability to preserve random slope variation. In this example, the JM-SY program in Appendix A could incorporate random slopes for the complete predictors (lunch assistance or gender), whereas the JM-AM program in Appendix B is limited to random intercepts. Chained equations imputation can accommodate random slopes for any level-1 predictor in the analysis model.

Table 4
Parameter Estimates From the Real Data Analysis Example

	Imputation methods					
	JM-SY		JM-AM		CE	
Parameter	Est.	*SE*	Est.	*SE*	Est.	*SE*
Intercept	55.20	.98	55.39	.99	55.00	.95
Lunch slope	−.67	.84	−.42	.76	−.72	.82
Gender slope	.53	.68	.58	.65	.58	.66
Math slope	.09	.01	.08	.01	.09	.01
Efficacy slope	.12	.09	.17	.09	.18	.11
Pretest slope	.29	.04	.29	.04	.27	.04
Intervention slope	4.27	1.46	4.34	1.54	4.53	1.43
Intercept variance	12.46	3.60	13.10	3.69	11.98	3.62
Residual variance	80.38	4.76	80.26	5.43	83.01	6.39

Note. JM-SY = joint imputation, Schafer & Yucel (2002); JM-AM = joint model, Asparouhov & Muthén (2010); CE = chained equations.

Discussion

Multilevel imputation procedures are now widely available in software packages, but comprehensive comparisons of these methods are virtually nonexistent, and contemporary missing data texts generally devote only a few paragraphs to the topic. In this paper we sought to provide researchers with a detailed review of two principal imputation frameworks for multilevel data, joint modeling, and chained equations imputation. Using statistical theory and computer simulation, we demonstrated four major conclusions: (a) joint modeling and chained equations imputation are appropriate for random intercept analyses; (b) the joint model is superior for analyses that posit distinct within- and between-cluster relations (e.g., a contextual effects model that includes a level-1 predictor and its cluster means, a multilevel structural equation model with different path values at level-1 and level-2); (c) chained equations imputation provides a dramatic improvement over joint modeling in random slope analyses; and (d) the latent variable formulation for categorical variables is quite effective, particularly when compared with multivariate normal imputation with rounding. These conclusions are consistent with theoretical expectations and underscore the importance of specifying an imputation model that preserves all associations in a given analysis model. To assist researchers in selecting an imputation method that best suits their analytic goals, Table 1 summarizes the major features of the two imputation frameworks.

Throughout the paper we used random intercept and random slope analysis models to demonstrate key ideas. Although these analysis models are relatively simple, the conclusions that we derive readily generalize beyond conventional multilevel models. To illustrate, consider a multilevel confirmatory factor analysis (MLCFA). Normal-theory maximum likelihood estimation can readily address missing data in these models, but multiple imputation might be preferable with categorical indicators because weighted least squares estimators (e.g., the WLSMV estimator in Mplus) require the strict missing completely at random mechanism (Enders & Baraldi, in press). MLCFAs—and multilevel structural equation models in general—are similar to contextual effects analyses in the sense that they allow within- and between-cluster covariance matrices to differ (the MLCFA accommodates different parameter values and even different factor structures at level-1 and level-2). As seen in Table 1, joint imputation is the best choice for this analysis because it assumes unique covariance matrices at each level.[10] In contrast, chained equations imputation is appropriate only when the factor structure and parameters are invariant across levels and will introduce bias otherwise. This example highlights the fact that Table 1 can guide imputation decisions for a range of scenarios beyond conventional multilevel models.

Our review of multilevel imputation suggests a number of future avenues for methodological research. To begin, our illustrations and simulation results clearly suggest that joint modeling and

[10] JM-AM automatically preserves all level-specific associations among all variables, whereas JM-SY requires the user to explicitly introduce the cluster means of any complete indicators.

chained equations imputation are equivalent in some situations but not others. In single-level data with normally distributed variables, methodologists have established equivalencies between the two methods (Hughes et al., 2014), but formal linkages have not been developed for multilevel data. This is an interesting avenue for future analytic work. On a related point, chained equations imputation is currently limited because it assumes common associations at level-1 and level-2. This limitation is more a function of software rather than the algorithm, per se. Specifically, it is possible to modify the algorithm to use the filled-in variables and their cluster means as predictors in subsequent imputation steps. Carpenter and Kenward (2013) allude to this possibility, but software packages do not currently offer this option. Chained equations imputation is currently restricted to multivariate normal variables, but this too is a function of existing software, as the method could readily accommodate the latent variable formulation for categorical variables. We are currently developing a software package that combines latent variable imputation with a chained equations algorithm, and preliminary tests suggest that this approach holds great promise. Finally, relatively few empirical studies have examined multilevel imputation routines, and future simulation studies could investigate a number of issues that we did not consider (e.g., imputation for multilevel structural equation models, guidelines for determining the optimal number of imputations, imputation with small cluster sizes or a small number of clusters, and the imputation of interaction effects, to name a few).

In sum, this paper highlights similarities and differences between joint imputation and chained equations imputation for multilevel data. Both approaches have unique strengths, and both have important limitations; no single method is appropriate for every situation. Relative to single-level methods, multilevel imputation techniques are still in their infancy and offer a number of interesting avenues for methodological research.

References

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

Agresti, A. (2015). *Foundations of linear and generalized linear models*. 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. (2005, April). *Imputation of categorical variables with PROC MI*. Paper presented at the SAS Users Group International.

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.

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

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

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

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

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

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiple imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation, 78,* 69–84. http://dx.doi.org/10.1080/10629360600903866

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

Enders, C. K., & Baraldi, A. N. (in press). Missing data methods. In P. Irwing, T. Booth, & D. Hughes (Eds.), *Handbook of psychometrics*. Oxford, UK: Wiley-Blackwell.

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. (2013). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall.

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

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–511. 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

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206–213. http://dx.doi.org/10.1007/s11121-007-0070-9

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

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. C. (2014). Joint modeling rationale for chained equations. *BMC Medical Research Methodology, 14,* 1–10. http://dx.doi.org/10.1186/1471-2288-14-28

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov Chain Monte Carlo. *American Journal of Political Science, 44,* 375–404. http://dx.doi.org/10.2307/2669318

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

Keller, B. T., & Enders, C. K. (2014, May). *A latent variable chained equations approach for multilevel multiple imputation*. Paper presented at the Modern Modeling Methods Conference, Storrs, CT.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119013563

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

Meinfelder, F. (2011). Package "mice." Retrieved from cran.r-project.org/web/packages/baboon/baboon.pdf on 5–16-2015

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

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

Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle school students of varying abilities. *Journal of Educational Psychology, 106,* 469–481. http://dx.doi.org/10.1037/a0035176

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

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

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology, 32,* 143–150.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473–489. http://dx.doi.org/10.1080/01621459.1996.10476908

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: 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. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57,* 19–35. http://dx.doi.org/10.1111/1467-9574.00218

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. 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 data. *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

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

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

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. K. Roberts & J. J. Hox (Eds.), *The 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. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67.

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 on 5-18-2014

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

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

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

Zhao, E., & Yucel, R. M. (2009, August). *Performance of sequential imputation method in multilevel applications*. Paper presented at the American Statistical Association Joint Statistical Meetings, Washington, DC.

Appendix A

Mplus JM-SY Joint Imputation Program (H1 Imputation)

```
DATA:
file = probsolve.dat;
VARIABLE:
names = student school txgrp lowab lrndis lunch
     female stanmath matheff pretest probsolv;
usevariables = txgrp – probsolv;
cluster = school;
within = lowab lrndis lunch female;
between = txgrp;
missing = all (–99);
ANALYSIS:
type = twolevel random;
estimator = bayes;
bseed = 83921;
MODEL:
%within%
stanmath matheff pretest probsolv on lowab lrndis lunch female;
stanmath matheff pretest probsolv;
stanmath matheff pretest probsolv with stanmath matheff pretest
probsolv;
%between%
stanmath matheff pretest probsolv on txgrp;
stanmath matheff pretest probsolv;
stanmath matheff pretest probsolv with
     stanmath matheff pretest probsolv;
[stanmath matheff pretest probsolv];
DATA IMPUTATION:
impute = stanmath matheff pretest probsolv;
ndatasets = 20;
save = probsolveimps*.dat;
thin = 500;
```

Appendix B

Mplus JM-AM Joint Imputation Program (H0 Imputation)

```
DATA:
file = probsolve.dat;
VARIABLE:
names = student school txgrp lowab lrndis lunch female stanmath
matheff pretest probsolv;
usevariables = txgrp – probsolv;
cluster = school;
missing = all (–99);
ANALYSIS:
type = twolevel basic;
bseed = 83921;
DATA IMPUTATION:
impute = stanmath matheff pretest probsolv;
ndatasets = 20;
save = probsolveimps*.dat;
thin = 500;
```

(*Appendices continue*)

Appendix C

MICE Chained Equations Imputation Program

```
# Load library
library(mice)

# Read data in
probsolve <- read.table("~/Desktop/tst_stuff/probsolve.dat",
quote="\"")

# Imputation parameters
burn <- 1000
seeds <- 083492
numimp <- 20

# Label variables
names(probsolve) <- c("school", "txgrp", "lowab", "lrndis", "lunch",
                      "female", "stanmath", "matheff", "pretest",
"probsolv")

# Recode missing values
probsolve[probsolve == -99] <- NA

# Initialize MICE
ini <- mice(probsolve, maxit = 0)

# Create pred matrix
pred <- ini$pred
pred["stanmath",] <- c(-2, 1, 1, 1, 1, 1, 0, 1, 1, 1)
pred["matheff", ] <- c(-2, 1, 1, 1, 1, 1, 1, 0, 1, 1)
pred["pretest", ] <- c(-2, 1, 1, 1, 1, 1, 1, 1, 0, 1)
pred["probsolv", ] <- c(-2, 1, 1, 1, 1, 1, 1, 1, 1, 0)

# Impute data
imp <- mice(probsolve, meth = c(",",",",",",",",'2l.norm',
            '2l.norm','2l.norm','2l.norm'), pred = pred, maxit = burn,
            m = numimp, seed = seeds)

# Write out data.
lapply(1:numimp, function(y){
     tmp <- complete(x=imp,y)
     write.table(tmp, file =
paste0("~/Desktop/tst_stuff/probsolveimps",y,".dat"),
                                  row.names=FALSE,col.names=FALSE)
})
```

Appendix D

SAS Chained Equations Imputation Program

```
%include '\\psf\home\desktop\sasmice2l.sas';

/* inputfile = file path for input text file */
/* varlist = input variable list */
/* cluster = level-2 identifier variable */
/* missingcode = missing value code */
```

(*Appendices continue*)

```
/* outputfile = file path for imputed data */
/* outputformat = separate or stack imputed data sets */
/* imputations = number of imputed data sets */
/* burnin = number of burn-in iterations */
/* thin = number between-imputation iterations */
/* complete = complete variables at level-1 and level-2 */
/* l2missing = level-2 variables to be imputed, (n) denotes normal*/
/* l1missing = level-1 variables to be imputed,(n) denotes normal*/
/* l1random = pairs of random slope variables joined by : */

%sasmice21(
inputfile = c:\temp\probsolve.dat,
varlist = school txgrp lowab lrndis lunch female stanmath matheff
pretest probsolv,
cluster = school,
missingcode = -99,
outputfile = c:\temp\probsolveimps.dat,
outputformat = separate,
imputations = 20,
burnin = 1000,
thin = 500,
complete = txgrp lowab lrndis lunch female,
l2missing =,
l1missing = stanmath (n) matheff (n) pretest (n) probsolv (n),
l1random =
);
```

Appendix E

Mplus Analysis and Pooling Program (JM-AM Imputation)

```
DATA:
file = probsolveimpslist.dat;
type = imputation;
VARIABLE:
names = stanmath matheff pretest probsolv
     lowab lrndis lunch female txgrp school;
usevariables = txgrp lunch female stanmath
     matheff pretest probsolv;
cluster = school;
within = lunch female stanmath matheff pretest;
between = txgrp;
DEFINE:
center lunch female stanmath matheff pretest (grandmean);
ANALYSIS:
type = twolevel random;
MODEL:
%within%
probsolv on lunch female stanmath matheff pretest;
probsolv;
%between%
probsolv on txgrp;
[probsolv];
probsolv;
```

(*Appendices continue*)

Appendix F

Simulation Results With Normally Distributed Variables, MCAR Mechanism, 20% Missing

		Single-level			Fixed effects			Joint model			Chained equations		
Parameter	True value	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis													
Intercept	.00	.00	NA	88.8%	.00	NA	95.6%	.00	NA	94.4%	.00	NA	94.6%
x slope	.34	.30	−.12	68.0%	.35	.01	94.6%	.34	.00	95.2%	.34	.00	94.4%
z slope	.20	.21	.07	95.8%	.20	.00	94.6%	.20	.01	95.0%	.20	.01	95.8%
Intercept variance	.83	.52	−.37	8.4%	.85	.03	95.4%	.83	.00	94.4%	.83	.00	94.0%
Residual variance	.81	1.12	.39	.0%	.79	−.02	90.6%	.80	.00	94.0%	.81	.00	94.0%
Contextual effects analysis													
Intercept	.00	.00	NA	90.6%	.00	NA	96.8%	.00	NA	96.6%	.00	NA	96.6%
x slope	.34	.15	−.58	.0%	.35	.02	90.0%	.34	.00	92.6%	.33	−.04	92.4%
z slope	.20	.26	.30	53.8%	.20	−.01	94.0%	.20	.00	94.6%	.20	.02	94.6%
Contextual slope	−.74	−.54	−.27	49.4%	−.74	.00	95.4%	−.75	.00	95.0%	−.72	−.03	95.8%
Intercept variance	.83	.53	−.36	10.6%	.86	.03	95.2%	.83	.00	95.4%	.84	.02	94.0%
Residual variance	.80	1.23	.54	.0%	.79	−.01	94.8%	.80	.00	93.6%	.81	.01	95.0%
Random slope analysis													
Intercept	.00	.00	NA	88.8%	−.01	NA	95.8%	−.01	NA	95.8%	−.01	NA	95.0%
x slope	.35	.31	−.10	80.2%	.35	.01	90.6%	.35	.00	90.2%	.34	−.01	94.8%
z slope	.20	.21	.05	95.4%	.20	.00	96.2%	.20	.00	96.2%	.20	−.01	95.8%
Intercept variance	.86	.51	−.40	7.4%	.85	−.01	95.0%	.83	−.04	92.0%	.84	−.03	93.2%
Slope variance	.09	.03	−.68	.8%	.04	−.54	6.4%	.04	−.55	7.2%	.08	−.14	86.8%
Covariance	.00	.00	NA	98.4%	.00	NA	97.8%	.00	NA	98.0%	.00	NA	96.6%
Residual variance	.81	1.17	.45	.0%	.84	.05	82.4%	.86	.06	68.8%	.82	.02	95.0%

Note. MCAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

Appendix G

Simulation Results With Normally Distributed Variables, MAR Mechanism, 20% Missing

		Single-level			Fixed effects			Joint model			Chained equations		
Parameter	True value	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis													
Intercept	.00	.00	NA	88.0%	.00	NA	95.4%	.00	NA	95.4%	.00	NA	94.8%
x slope	.34	.32	−.08	76.8%	.34	.01	94.4%	.34	.00	94.4%	.34	.00	94.8%
z slope	.20	.21	.05	93.6%	.20	.00	92.8%	.20	.00	95.0%	.20	.00	94.8%
Intercept variance	.83	.52	−.38	7.8%	.85	.03	94.0%	.83	.00	95.0%	.83	.00	94.4%
Residual variance	.81	1.12	.39	.0%	.79	−.01	95.6%	.81	.00	92.8%	.81	.00	94.2%
Contextual effects analysis													
Intercept	.00	.00	NA	92.2%	.00	NA	97.0%	.00	NA	96.0%	.00	NA	96.8%
x slope	.34	.14	−.59	.0%	.34	.01	95.4%	.34	.00	95.2%	.33	−.03	93.6%
z slope	.20	.26	.30	58.6%	.20	.00	94.6%	.20	.00	94.6%	.20	.01	95.6%
Contextual slope	−.74	−.54	−.28	46.6%	−.74	−.01	98.0%	−.74	.00	96.8%	−.72	−.03	95.8%
Intercept variance	.83	.53	−.36	9.6%	.86	.03	95.6%	.83	.00	93.6%	.84	.02	94.6%
Residual variance	.80	1.24	.54	.0%	.79	−.01	92.4%	.80	.00	94.8%	.81	.01	95.4%
Random slope analysis													
Intercept	.00	.00	NA	89.6%	.00	NA	95.6%	.00	NA	95.6%	−.01	NA	95.8%
x slope	.35	.33	−.06	89.0%	.35	.01	90.2%	.35	.00	90.8%	.35	.00	95.6%
z slope	.20	.21	.03	95.6%	.20	.00	94.4%	.20	.01	95.0%	.20	.00	94.8%
Intercept variance	.86	.51	−.41	5.8%	.85	−.01	95.0%	.83	−.04	92.8%	.84	−.03	94.0%
Slope variance	.09	.03	−.61	2.4%	.04	−.51	11.6%	.04	−.52	10.8%	.08	−.11	88.6%
Covariance	.00	−.02	NA	91.8%	.00	NA	98.4%	−.01	NA	98.6%	.00	NA	96.6%
Residual variance	.81	1.16	.44	.0%	.84	.04	85.4%	.85	.06	73.2%	.82	.02	96.6%

Note. MAR = missing at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

(*Appendices continue*)

Appendix H

Simulation Results With Normally Distributed Variables, MCAR Mechanism, 40% Missing

Parameter	True value	Single-level			Fixed effects			Joint model			Chained equations		
		Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis													
Intercept	.00	−.01	NA	78.4%	.00	NA	96.4%	.00	NA	95.6%	.00	NA	95.8%
x slope	.34	.32	−.07	82.4%	.36	.05	89.6%	.34	.00	93.0%	.35	.01	92.6%
z slope	.20	.20	.04	93.8%	.19	−.03	94.4%	.20	.00	94.8%	.20	−.01	93.2%
Intercept variance	.83	.28	−.66	.0%	.90	.08	97.4%	.83	.00	94.0%	.83	.00	94.0%
Residual variance	.81	1.35	.67	.0%	.77	−.05	80.2%	.81	.00	95.6%	.81	.00	95.0%
Contextual effects analysis													
Intercept	.00	.00	NA	85.8%	.00	NA	96.6%	.00	NA	96.0%	.00	NA	96.4%
x slope	.34	.04	−.87	.0%	.36	.05	90.0%	.34	.00	95.4%	.30	−.13	70.4%
z slope	.20	.29	.45	40.8%	.19	−.02	93.8%	.20	.00	94.6%	.21	.07	90.0%
Contextual slope	−.74	−.42	−.43	13.4%	−.74	−.01	96.2%	−.74	.00	96.0%	−.66	−.11	89.2%
Intercept variance	.83	.30	−.64	.0%	.90	.08	96.6%	.83	.00	94.6%	.86	.04	96.4%
Residual variance	.80	1.53	.90	.0%	.77	−.05	78.0%	.80	.00	93.8%	.82	.02	89.8%
Random slope analysis													
Intercept	.00	.00	NA	80.8%	−.01	NA	96.4%	−.01	NA	96.0%	.00	NA	95.2%
x slope	.35	.33	−.05	82.6%	.36	.05	84.6%	.34	.00	87.6%	.34	−.02	92.8%
z slope	.20	.21	.04	96.4%	.19	−.03	94.2%	.20	.00	95.2%	.20	−.01	95.6%
Intercept variance	.86	.28	−.67	.0%	.89	.04	96.4%	.82	−.04	91.2%	.81	−.06	89.4%
Slope variance	.09	.01	−.88	.0%	.02	−.81	.0%	.02	−.83	.0%	.06	−.29	70.2%
Covariance	.00	.00	NA	99.6%	.00	NA	99.6%	.00	NA	100.0%	.00	NA	98.2%
Residual variance	.81	1.42	.76	.0%	.84	.05	83.6%	.89	.10	55.8%	.84	.05	86.4%

Note. MCAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

Appendix I

Simulation Results With Categorical Variables, MCAR Mechanism, 20% Missing

Parameter	True value	Joint model			Chained equations		
		Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis							
Intercept	1.92	1.92	.00	97.8%	1.78	−.07	94.4%
x slope	.64	.64	.00	95.6%	.59	−.07	92.0%
z slope	.40	.40	.01	95.6%	.37	−.07	94.8%
Intercept variance	3.07	3.07	.00	92.4%	3.12	.02	91.2%
Random slope analysis							
Intercept	1.99	1.87	−.06	93.0%	1.88	−.05	94.4%
x slope	.66	.62	−.06	89.6%	.62	−.05	93.2%
z slope	.40	.39	−.02	96.0%	.36	−.08	94.8%
Intercept variance	5.54	3.83	−.31	73.2%	5.47	−.01	94.4%
Slope variance	.28	.10	−.62	53.0%	.28	.01	97.0%
Covariance	−.84	−.31	−.63	61.2%	−.80	−.04	95.4%

Note. MCAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

(*Appendices continue*)

Appendix J

Simulation Results With Categorical Variables, MAR Mechanism, 20% Missing

		Joint model			Chained equations		
Parameter	True value	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis							
Intercept	1.92	1.91	.00	95.6%	1.83	−.05	95.4%
x slope	.64	.64	.00	96.4%	.60	−.07	93.2%
z slope	.40	.40	.00	92.8%	.34	−.14	88.6%
Intercept variance	3.07	3.08	.00	91.4%	3.14	.02	90.0%
Random slope analysis							
Intercept	1.99	1.89	−.05	92.0%	1.94	−.02	96.0%
x slope	.66	.63	−.05	91.8%	.63	−.04	94.4%
z slope	.40	.39	−.02	94.4%	.34	−.14	88.8%
Intercept variance	5.54	4.04	−.27	76.2%	5.41	−.02	94.0%
Slope variance	.28	.12	−.59	57.6%	.28	.02	97.8%
Covariance	−.84	−.36	−.57	66.2%	−.79	−.05	94.8%

Note. MAR = missing at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

Appendix K

Simulation Results With Categorical Variables, MCAR Mechanism, 40% Missing

		Joint model			Chained equations		
Parameter	True value	Avg. est.	Rel. bias	95% cov.	Avg. est.	Rel. bias	95% cov.
Random intercept analysis							
Intercept	1.92	1.97	.03	95.2%	1.70	−.12	90.2%
x slope	.64	.66	.03	94.2%	.56	−.12	86.4%
z slope	.40	.40	.00	92.6%	.34	−.15	88.8%
Intercept variance	3.07	3.11	.01	91.8%	3.29	.07	87.8%
Random slope analysis							
Intercept	1.99	1.84	−.07	91.0%	1.76	−.12	92.0%
x slope	.66	.61	−.07	90.4%	.58	−.12	89.2%
z slope	.40	.38	−.03	95.8%	.33	−.16	90.6%
Intercept variance	5.54	3.27	−.41	62.4%	5.45	−.02	94.8%
Slope variance	.28	.04	−.86	8.6%	.26	−.05	98.0%
Covariance	−.84	−.12	−.86	28.2%	−.72	−.14	97.6%

Note. MCAR = missing completely at random. True values are the average estimates from 500 complete data sets. Relative bias is difference between an average estimate and its true value expressed as a proportion of the true value.

Received June 18, 2014
Revision received May 19, 2015
Accepted August 17, 2015 ■