Three-Level Multiple Imputation:

A Fully Conditional Specification Approach

by

Brian Tinnell Keller

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved October 2015 by the
Graduate Supervisory Committee:

Craig Enders, Co-Chair
Kevin Grimm, Co-Chair
Roy Levy

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

Currently, there is a clear gap in the missing data literature for three-level models. To date, the literature has only focused on the theoretical and algorithmic work required to implement three-level imputation using the joint model (JM) method of imputation, leaving relatively no work done on fully conditional specification (FCS) method. Moreover, the literature lacks any methodological evaluation of three-level imputation. Thus, this thesis serves two purposes: (1) to develop an algorithm in order to implement FCS in the context of a three-level model and (2) to evaluate both imputation methods. The simulation investigated a random intercept model under both 20% and 40% missing data rates. The findings of this thesis suggest that the estimates for both JM and FCS were largely unbiased, gave good coverage, and produced similar results. The sole exception for both methods was the slope for the level-3 variable, which was modestly biased. The bias exhibited by the methods could be due to the small number of clusters used. This finding suggests that future research ought to investigate and establish clear recommendations for the number of clusters required by these imputation methods. To conclude, this thesis serves as a preliminary start in tackling a much larger issue and gap in the current missing data literature.

For my parents, Betty and Ken, who have been there for both the highs and lows of my life and unconditionally supported me through them. And for my brother Kyle, who has always been there, not only as a brother, but as a friend, imparting his knowledge on the roads he has already traveled.

ACKNOWLEDGEMENTS

I would like to acknowledge all the people who have made a great impact in my academic career thus far. Specifically, I would like to acknowledge Christian Geiser for introducing me to and fostering my enjoyment for quantitative methodology, David MacKinnon for always leaving me thinking in his ANOVA and mediation classes, Leona Aiken for her initial class on regression and providing advice throughout my career thus far, Steve West for acting as a mentor and his advice that will serve me well throughout my academic career, and finally my advisor and friend, Craig Enders, his continued mentoring and comments made this thesis possible. I would also like to especially acknowledge Leona and Steve for developing and fostering such an amazing program in quantitative methodology. What I have learned in my years at Arizona State University will always serve as my foundation throughout my career. Finally, I would like to acknowledge all the scholars who have come before me, for, as attributed to Bernard of Chartres and famously stated by Isaac Newton, I "[stand] on the shoulders of giants."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

While the best solution to missing data is to not have missing data, this is often not feasible in realistic research settings. Over the past few decades advances have occurred in both the theory and the treatment of missing data. Previously researchers were required to resort to so-called deletion methods (i.e., listwise deletion and pairwise deletion), however, newer and more advanced techniques are capable of estimating parameters more efficiently and without bias under less restrictive assumptions. Currently there are two major techniques that are available to handle missing data: Maximum likelihood and multiple imputation. While these methods are well developed for single-level data, they are more limited in the multilevel case. Furthermore, there has been little research done in extending missing data handling techniques to three-level data structures. Thus, the aim of this thesis is to provide a method for researchers to handle missing data on all variables in three-level models.

Moving forward I will give a brief overview of maximum likelihood and multiple imputation. Next I will describe the current limitations in the context of multilevel models and my rational for choosing multiple imputation over maximum likelihood. Finally, I will summarize and give the goals of this paper.

1.1 Maximum Likelihood and Multiple Imputation

Maximum likelihood (ML) takes a likelihood-based approach towards the handling of missing data (Little & Rubin, 2002; Schafer & Graham, 2002). The goal of ML is to estimate the parameters using all available data. This can be achieved under specific assumptions about the missingness of the data. If the assumptions hold, then

the likelihood function that ML maximizes is written as:

$$\mathcal{L}\left(\theta \mid Y_{obs}\right) = \int \mathcal{L}\left(\theta \mid Y_{obs}, Y_{mis}\right) dY_{mis} \tag{1.1}$$

where θ are the unknown parameters in the model and $\mathcal{L}\left(\theta \mid Y_{obs}, Y_{mis}\right) dY_{mis}$ denotes the likelihood function based on the observed data only. In essence, ML integrates over the missing data and eliminating the dependence on Y_{mis}, basing the estimation only on observed data (Schafer, 2001).

Multiple Imputation (MI) takes a Bayesian-based approach towards handling missing data (Rubin, 1987; Schafer, 1997b). Instead of trying to estimate the parameters using all available data, MI "fills-in" the missing data from a distribution, namely the posterior predictive distribution. Researchers subsequently fit one or more analysis models to the filled-in data. MI then averages parameter estimates and standard errors across the data sets to obtain the pooled parameters. This process is usually carried out with a Markov chain Monte Carlo (MCMC) method, such as the Gibbs sampler or Metropolis-Hastings algorithm. The motivation behind MI is to treat the missing data as a source of random variability that needs to be averaged over. Similar to ML, MI accomplishes this by integration, however, MI solves the problem via simulation methods instead of analytically. Furthermore, MI differs from ML in the sense that MI employs a general model (e.g., a saturated linear regression model for single-level data) to fill-in or *impute* the missing values for a variety of analyses. Thus, MI is tailored for a family of analytic models, whereas ML is tailored for a specific model of interest.

1.2 Missing Data Handling for Multilevel Models

With single-level data there is often no reason to prefer ML or MI. Several studies have shown that the two procedures tend to produce identical results, particularly when using the same set of input variables (Collins, Schafer, & Kam, 2001; Schafer, 2003). Nevertheless, there are some advantages of MI over ML, especially with more complex data structures like multilevel data. For several reasons imputation procedures are often thought of as being more flexible than ML. They allow for mixtures of both categorical and continuous variables and allow for one general procedure to handle a variety of analyses. In addition, the maximization of the likelihoods specified by ML are often problem-specific and sometimes must be approximated in complex analyses, such as with multilevel modeling (Yucel, 2008). This is particularly true when categorical outcomes are present. Another issue with ML is the handling of incomplete predictors. In order to properly handle missing data, distributional assumptions must be made about the missing variable. Multilevel modeling software packages that implement ML tend to define predictors as fixed (i.e., do not make any assumption about the predictors distribution), necessitating the exclusion of cases with missing values on those predictors. The exclusion of cluster-level predictors (incomplete level-2 variables) can be particularly damaging because the entire set of level-1 observations are excluded for the cluster. Currently, ML solutions that do exist for missing predictors are restricted to multivariate normal data and random intercept models only (e.g., Shin & Raudenbush, 2007).

Although MI is well suited for clustered data structures, existing imputation methods have limitations. A variety of methods have been proposed to deal with two-level multiple imputation and with accompanying software packages (Enders, Mistler, & Keller, in press; Keller & Enders, 2014, May), but relatively little work has been

devoted to three-level models. Considering the existing implementations of two-level models, all techniques can handle random intercept models with normally distributed variables, only some can handle categorical variables (e.g., Mplus and MLwIN) and others can handle random slope models (e.g., MICE). Despite this, very few methods offer a complete set of options that are often seen in social science data (i.e., both continuous and categorical data with random intercepts and slopes). Furthermore, the current approaches are limited in other ways. For example, MICE's approach, as currently implemented cannot handle missing level-2 variables at the same time as missing level-1 variables, and the same applies to the original work on joint imputation (i.e., Schafer, 2001; Schafer & Yucel, 2002). More limitations to joint modeling include the inability to tailor random effects for each variable and handle cross-level interactions between incomplete variables as currently implemented.

1.3 Goals and Summary

In conclusion, there is limited methodological work for handling missing data in the context of multilevel models and social scientists do not have adequate tools to deal with missing data in multilevel structures. Furthermore, there are very limited solution for three-level models with multiple imputation (Asparouhov & Muthén, 2010; Yucel, 2008) and even more limited simulation work assessing their performance. The goal of this thesis was to develop an imputation procedure that accommodates a three-level structure. Additionally, the aim of this paper was to assess the accuracy and precision of the developed method and other current methods.

The thesis is organized as follows. First, Chapter 2 will provide background information required to understand three-level multiple imputation. This will then be followed by a brief review of the current literature for multiple imputation with three-levels. In Chapter 3, I go into the more algorithmic details of the method used in

this thesis and describe the simulation used to compare the new method for three-level imputation to available existing methods. Chapter 4 presents the results of the simulation that was performed. Finally, Chapter 5 is a discussion of the results, the implications and limitations of the simulation, and directions of future research.

Chapter 2

Literature Review

This chapter provides a brief overview of the literature thus far and rational as to some of the decisions made in Chapter 3. First, I give a brief introduction of missing data mechanisms as this lays the theoretical foundation for using imputation as a solution to some missing data problems. Next, I briefly outline the three steps of multiple imputation (imputation step, analysis step, and pooling step). Then I focus on just the imputation step, where I explain the two different frameworks used within MCMC (joint imputation and fully conditionally specified imputation). The subsequent section will go over multilevel imputation at two-levels with both frameworks. I will then provide a critique of joint imputation and my rational for choosing a fully conditionally specified model. The next section will introduce a brief overview of the three-level model. This is then followed by a summary of Yucel (2008) and the approach proposed using joint imputation for three-levels. Finally, I will give a more in-depth statement of the purpose of this thesis.

2.1 Missing Data Mechanisms

There are two fundamental concepts in current missing data theory, missing data patterns and missing data mechanisms. These are two distinct and separate concepts that are often confused. A *missing data pattern* describes location of the missing values in the data. It makes no attempt to describe how the data are missing, neither mathematically nor conceptually. On the other hand, a *missing data mechanism* provides a theoretical account for how the data are missing, be it directly via a measured variable or indirectly through a spurious relationship caused by a third variable.

Rubin (1976) proposed the decomposition of any random variable into two parts: observed scores (denoted with an *obs* subscript) and the would-be but unobserved scores (denoted with a *mis* subscript). Building on this work, Little and Rubin (2002) defined three different missing data mechanisms: missing completely at random, missing at random, and missing not at random. Missing completely at random (MCAR) can be thought of as the data being missing in a haphazard fashion or truly random in the colloquial sense. Formally, the MCAR mechanism is stated as:

$$P\left(R_Y \mid Y_{obs}, Y_{mis}, \theta\right) = P\left(R_Y \mid \theta\right) \tag{2.1}$$

where θ represents the parameters of the model and R_Y is a vector of indicator variables that denotes missing and observed values. The equation can be read as the probability of missingness of a set of variables, Y, is unrelated to the observed or missing values. For example, a researcher may be administering a survey and if respondents inadvertently skip a survey item on income for reasons that are uncorrelated with the data, the item would satisfy the MCAR mechanism. Other examples of MCAR include planned missing data designs and the classical randomized experiment.

Missing at random (MAR) is a more general case and subsumes MCAR. In order for MAR to hold, the probability of missingness of a set of variables, Y, is conditionally independent of Y_{mis} given Y_{obs} and the parameters in the model. Or stated algebraically:

$$P\left(R_Y \mid Y_{obs}, Y_{mis}, \theta\right) = P\left(R_Y \mid Y_{obs}, \theta\right) \tag{2.2}$$

To continue the example, suppose the respondents with higher education are more likely not to respond to a survey item on income. The item would be MAR if there was an equal probability of nonresponse among respondents with the same education

level. Traditionally, software implementations of ML and MI make the assumption that the data are missing under the MAR mechanism.

Finally, missing not at random (MNAR), states that the probability of missingness of a set of variables, Y, is related to Y itself, be it directly or through a spurious relationship due to unmeasured variables. For MNAR, the probability distribution does not simplify:

$$P\left(R_Y \mid Y_{obs}, Y_{mis}, \theta\right) \tag{2.3}$$

To continue the above example, if the income item was skipped such that respondents in a higher income bracket are more likely not to respond, even after conditioning on other observed variables, then the mechanism is MNAR. While there are methods to handle MNAR data, they are not the focus of this thesis and for the remainder of the paper all variables will be assumed to be missing under the MAR mechanism.

2.2 Single-Level Multiple Imputation

In this section, I describe multiple imputation in the single-level case because it lays a groundwork for understanding multiple imputation for more than one level. Multiple imputation consists of three major steps: (1) an imputation step, (2) analysis step, and (3) a pooling step. In the imputation step, the researcher imputes the missing data with plausible values and saves multiple copies of the "filled-in" data with different values imputed. Next, the analysis step is where the researcher analyzes these saved copies of the data with a specified analysis model. Finally, the pooling step is when the researcher pools the estimates from the analysis by averaging estimates in accordance to specific formulas (Rubin, 1987; Schafer, 1997b). While this is the general layout of MI, the focus of this thesis is on the imputation step.

In the imputation step, imputations for the missing variable are generated by sampling from a distribution of plausible values (a posterior predictive distribution) determined by the imputation model. An iterative simulation method (e.g., Gibbs sampler) is employed in order to generate imputed values. More specifically, there are two steps per iteration of the simulation. For step (1), the simulation treats the missing observations as known, using the imputations from the previous iteration (or starting values for the first iteration) and then draws parameters. With single-level imputation these parameters are typically (but not necessarily) a covariance matrix and mean vector. For step (2), the simulation treats the parameter values from step (1) as known and draws values for missing data based on the simulated parameters in step (1). For single-level multivariate normal data, this is achieved by taking the predicted score in a linear regression for the missing value and adding a random error term, which restores the variability back into the data.

There are two main frameworks used to approximate the posterior predictive distribution that the simulation samples from: (1) joint models approach and (2) fully conditional specification models (also known as chained equations). To better illustrate single-level imputation and the differences between the two approaches, let us suppose I have three variables, x, y, and z, where x is complete and y and z have both complete and missing values (denoted with an *obs* and *mis* subscript or superscript). For simplicity, I will assume that all three variables are normally distributed.

To begin, consider joint imputation. For this example, I specify a multivariate normal distribution with a mean vector and covariance matrix. I can then reconstitute the parameters of the joint distribution into an imputation model for each missing data pattern. Because y and z are missing, there are three potential patterns: (1) only y missing, (2) only z missing, and (3) y and z missing together. Thus, I must

specify three separate pattern specific distributions to draw the imputations from:

$$\begin{aligned} y_{mis}^{(t)} &\sim N\left(\alpha_0^{(t)} + \alpha_1^{(t)}x + \alpha_2^{(t)}z_{obs}, \sigma_{(y|xz)}^{2(t)}\right) \\ z_{mis}^{(t)} &\sim N\left(\beta_0^{(t)} + \beta_1^{(t)}x + \beta_2^{(t)}y_{obs}, \sigma_{(z|xy)}^{2(t)}\right) \\ \begin{bmatrix} y_{mis}^{(t)} \\ z_{mis}^{(t)} \end{bmatrix} &\sim MVN\left(\begin{bmatrix} \gamma_0^{(t)} + \gamma_1^{(t)}x \\ \kappa_0^{(t)} + \kappa_1^{(t)}x \end{bmatrix}, \Sigma_{(yz|x)}^{(t)}\right) \end{aligned} \tag{2.4}$$

where (t) represents the t^{th} iteration, α, β, γ, and κ are all regression coefficients (with a subscript 0 indicating an intercept), σ^2 represents a residual variance, and Σ represents a residual covariance matrix. The coefficients and variance terms that define the normal distributions are computed from the appropriate elements of the simulated covariance matrix drawn at the previous step (described previously). See Schafer (1997b) for a complete description of this process.

In contrast, fully conditional specification (FCS) algorithms take a different approach to approximating the posterior distributions of the missing data. Instead of using a joint distribution, FCS specifies multiple conditional univariate distributions. To illustrate this in a general fashion, Y denotes a set of multivariate variables, ranging from 1 to m variables. The general form for the FCS at the t^{th} iteration is as follows (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006):

$$\begin{aligned} \theta_1^{(t)} &\sim P\left(\theta_1 \mid y_1^{obs}, y_2^{(t-1)}, \ldots, y_m^{(t-1)}\right) \\ y_1^{(t)} &\sim P\left(y_1^{mis} \mid y_1^{obs}, y_2^{(t-1)}, \ldots, y_m^{(t-1)}, \theta_1^{(t)}\right) \\ &\vdots \\ \theta_m^{(t)} &\sim P\left(\theta_m \mid y_m^{obs}, y_1^{(t-1)}, \ldots, y_{m-1}^{(t-1)}\right) \\ y_m^{(t)} &\sim P\left(y_m^{mis} \mid y_m^{obs}, y_1^{(t-1)}, \ldots, y_{m-1}^{(t-1)}, \theta_m^{(t)}\right) \end{aligned} \tag{2.5}$$

where θ represents the parameters of the imputation model from 1 to m variables. Note that the imputation scheme still follows the two steps described earlier, (1) draw the parameters treating the missing observations as known and (2) draw the missing observations treating the parameters as known; however, now these two steps are applied to each incomplete variable. Furthermore, the parameters in θ are not required to be functions of the parameters of a joint distribution (e.g., if y_1 was normally distributed and y_2 was imputed with a logistic regression).

Returning to the three-variable example from earlier, in each iteration of the FCS model one specifies a distribution for each missing variable, treating the previous parameter draws as known. Therefore, I must specify two imputation steps, each of which also requires supporting steps to generate the necessary parameters. For the missing values at iteration t the imputation steps are as follows:

$$\begin{aligned} y_{mis}^{(t)} &\sim N\left(\alpha_0^{(t)} + \alpha_1^{(t)}x + \alpha_2^{(t)}z^{(t-1)}, \sigma_{(y|xz)}^{2(t)}\right) \\ z_{mis}^{(t)} &\sim N\left(\beta_0^{(t)} + \beta_1^{(t)}x + \beta_2^{(t)}y^{(t)}, \sigma_{(z|xy)}^{2(t)}\right) \end{aligned} \tag{2.6}$$

In contrast to Equation 2.4, the 'obs' subscript for both y and z are gone and replaced with superscripts of t and $t-1$, respectively, which denote the iteration number. This is because the FCS method treats the missing values of other variables as known and conditions on them. Thus, for the imputation of y at step t, FCS uses the previous iterations imputed values for z. Similarly, for the imputation of z at step t, FCS uses the newly imputed values of y. Also note that the ordering of y and z is arbitrary (and the order may be switched as long as the subscripts are also swapped). Finally, each draw for the variable requires a supporting draw of parameter values that are not shown here, but well illustrated in the literature (see van Buuren, 2007; van Buuren et al., 2006).

2.3 Two-level Multiple Imputation

Thus far I have only discussed single-level imputation, but this can be extended to a multilevel setting. The key feature with multilevel imputation is accounting for the appropriate clustering. If clustering is ignored and a single-level imputation model is used to impute the clustered data, estimates of the cluster-level variation are attenuated (Enders et al., in press; Van Buuren et al., 2011). These, so-called "flat-file" imputation methods have been studied by previous research (e.g., Cheung, 2007; Gibson & Olejnik, 2003; Roudsari, Field, & Caetano, 2008; Zhang, 2005) and are not the focus of this thesis. Instead, the paper focuses on modeling the clustering explicitly by including random effects and the associated parameters (i.e., variances of said random effects) that account for the source of variability provided by the clustering.

Recall the three-variable example from the previous section, where x, y, and z are normally distributed with three missing data patterns. To map onto this example, assume that these three variables all have clustering (i.e., level-1 variables in a multilevel model). For example, a researcher may be studying students within classrooms, where y is a student's score on a standardized math test at the end of year, z is a student's score on a standardized math test from the previous year, and x is a math pretest given at the beginning of the year. Therefore, a researcher might be interested in the following random intercept analysis model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + u_{0j} + \epsilon_{ij} \tag{2.7}$$

where subscript i represents the student, subscript j is the classroom, and u_{0j} is the random intercept.

In order to extend the joint model to this above example, one must generate imputations from the following distribution:

$$\begin{bmatrix} y_{ij}^{(mis)} \\ z_{ij}^{(mis)} \end{bmatrix} \sim MVN \left(\begin{bmatrix} \beta_{0(y)} + \beta_{1(y)} x_{ij} + u_{0j(y)} \\ \beta_{0(z)} + \beta_{1(z)} x_{ij} + u_{0j(z)} \end{bmatrix}, \Sigma_{(yz|x)} \right) \tag{2.8}$$

It is important to note that the '(y)' and '(z)' subscripts denote that these parameters are different, but the same letters are used to keep the notation concise. The algorithm for the joint model differs by treating all missing variables as dependent variables and all complete variables as predictors of the missing variables. Conceptually, the joint model shares the same two steps (i.e., draw parameters and draw imputations) with the single-level case; however, an additional draw is now required in the first step in order to sample the random effects. Furthermore, the mean vector of the distribution is defined by predicted values that now take into account the clustering via the random effects (i.e., u_{0j}) and now multilevel parameters are sampled instead of the covariance matrix and mean vector. A number of resources in the literature give the draw steps for the multilevel model parameters and residuals (i.e., W. Browne & Draper, 2000; W. J. Browne, 1998; Goldstein, Bonnet, & Rocher, 2007; Kasim & Raudenbush, 1998; Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008). The algorithmic work of this thesis builds heavily from this previous work and is reviewed later in the paper.

There are several limiting factors with joint imputation. For example, joint imputation naturally works the best at the lowest level and it struggles to accommodate missing variables at different levels (which are not in our example). Currently, the joint imputation approach is limited to only random slopes between complete and incomplete variables and cannot estimate a random slope between two incomplete variables (and some implementations of the method allow for only random inter-

cepts). This can be illustrated in the Equation 2.8, where both y and z are drawn, yet they do not have influence in the predicted score. Thus, any residual association between the variables not explained by z must be specified in the residual covariance matrix. However, because the residual covariance matrix is specified as homogenous across clusters (denoted by the lack of subscripts), the model is unable to accommodate the random slopes. Conceptually, a random slope expresses that the relationship between the two variables changes among clusters; however, a homogenous residual covariance matrix implies that the slopes do not vary. Another limiting factor is that, in some implementations, the complete variables are required to have the same random slopes for all incomplete variable. Finally, the joint imputation method must assume a common distribution (usually multivariate normal) among the variables (all variables or only the incomplete variables depending on the joint models method). This can often lead to trouble when some variables are categorical and some are continuous. A latent approach to handling categorical variables can avoid this limitation, but this assumes the latent variable is normally distributed and thus imputes on the latent metric (Carpenter & Kenward, 2013; Enders et al., in press).

The FCS method can also be extended to multilevel data. Recall the general equation for one iteration of the FCS algorithm (Equation 2.5). As with the joint model, the θ's in the equation now contain multilevel parameters and random effects. Thus, they must be drawn from their appropriate posterior distributions (which are essentially the same as the draws of the joint model). For iteration t, the univariate draw steps are in Equation 2.9 and 2.10.

$$y_{ij}^{(t)} \sim N\left(\beta_{0(y)}^{(t)} + \beta_{1(y)}^{(t)} x_{ij} + \beta_{2(y)}^{(t)} z_{ij}^{(t-1)} + u_{0j(y)}^{(t)}, \sigma_{(y|xz)}^{2(t)}\right) \tag{2.9}$$

$$z_{ij}^{(t)} \sim N\left(\beta_{0(z)}^{(t)} + \beta_{1(z)}^{(t)} x_{ij} + \beta_{2(z)}^{(t)} y_{ij}^{(t)} + u_{0j(z)}^{(t)}, \sigma_{(z|xy)}^{2(t)}\right) \tag{2.10}$$

Once again the parameters in Equation 2.9 and 2.10 are not the same parameter and the subscript reflects this, where '(y)' represents the the parameter for the distribution of 'y' at iteration 't'. While the previous equation includes the draws of the missing variables, the main draws in the FCS Gibbs sampler for multilevel models are as follows (Schafer & Yucel, 2002; Van Buuren et al., 2011; Zeger & Karim, 1991):

1. Sample fixed effects from $P\left(\beta \mid y, u, \sigma^2\right)$.
2. Sample random effects from $P\left(u \mid y, \beta, \Sigma_u, \sigma^2\right)$.
3. Sample level-2 covariance matrix from $P\left(\Sigma_u \mid u\right)$.
4. Sample level-1 residual variance from $P\left(\sigma^2 \mid y, \beta, u\right)$.
5. Sample missing values and impute them into the data set.
6. Repeat step 1 to 5 for z.
7. Repeat step 1 to 6 until convergence.

For a more technical breakdown of the Gibbs sampler see the Methods section and Appendix A.

2.4 Overview of the Three-Level Model

Due to the complexity of three-level models, I will use a consistent notational system throughout the remainder of the paper. A lowercase 'y' will always represent the level-1 dependent variable, a lowercase 'a' will always represent a level-1 independent variable, a lowercase 'b' will represent a level-2 independent variable, and a lowercase 'c' will represent a level-3 independent variable. For example, a random intercept model for the i^{th} observation within level-2 cluster j and level-3 cluster k is

represented in Equation 2.11:

$$y_{ijk} = \beta_0 + \beta_1 a_{ijk} + \beta_2 b_{jk} + \beta_3 c_k + u_{0_{jk}} + v_{0_k} + \epsilon_{ijk} \tag{2.11}$$

where subscript i represents the lowest level (i.e., level-1), subscript j represents the level-2 cluster, and subscript k represents the level-3 cluster (i.e, the highest level). β_0 is the intercept, β_1 to β_p (i.e., $p = 3$ for this example) are the 'fixed effects' for 1 to p predictors (these can be at any level, denoted by subscripting), $u_{0_{jk}}$ is the random effect for the intercept at level-2, v_{0_k} is the random effect for the intercept at level-3. Finally, ϵ_{ijk} is the residual for cluster i within level-2 cluster j and level-3 cluster k. For completeness, Equation 2.12 represents the addition of a random slope to the model for variable 'a', at both level-1 and level-2, and variable 'b', at level-2.

$$y_{ijk} = \beta_0 + \beta_1 a_{ijk} + \beta_2 b_{jk} + \beta_3 c_k + u_{0_{jk}} + v_{0_k} + a_{ijk} \cdot \left(u_{1_{jk}} + v_{1_k} \right) + b_{jk} v_{2_k} + \epsilon_{ijk} \tag{2.12}$$

Because the notational system can start to become cumbersome as the model increases in complexity, it is also useful to represent the above equation in a matrix notation. I will use the following matrix notation in the paper for when it increases clarity of the three-level models:

$$\mathbf{y}_{jk} = \mathbf{X}_{jk}\boldsymbol{\beta} + \mathbf{W}_{jk}\mathbf{u}_{jk} + \mathbf{Z}_{jk}\mathbf{v}_k + \boldsymbol{\epsilon}_{jk} \tag{2.13}$$

where $\mathbf{y}_{jk}$ is a column vector of the criterion, $\mathbf{X}_{jk}$ is a matrix of explanatory variables, $\boldsymbol{\beta}$ is a column vector of regression coefficients, $\mathbf{W}_{jk}$ is a matrix containing a unit vector and the level-1 variables that are allowed to have a level-1 random effect on the outcome, and $\mathbf{u}_{jk}$ is a column vector with the random effects for level-1 variables. $\mathbf{Z}_{jk}$ is a matrix containing a unit vector and level-2 variables that are allowed to have

a level-2 random effect on the outcome, $\mathbf{v}_k$ is a column vector with the random effects for the $\mathbf{Z}_{jk}$ matrix, and $\boldsymbol{\epsilon}_{jk}$ is a column vector with the level-1 residuals for the level-2 cluster j within level-3 cluster k. To better illustrate the matrix notation consider the subset of a level-2 cluster, j, and within level-3 cluster k with a cluster size of $n_{jk} = 3$ cases:

$$\begin{aligned}\begin{bmatrix} y_{1jk} \\ y_{2jk} \\ y_{3jk} \end{bmatrix} &= \begin{bmatrix} 1 & a_{1jk} & b_{jk} & c_k \\ 1 & a_{2jk} & b_{jk} & c_k \\ 1 & a_{3jk} & b_{jk} & c_k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 & a_{1jk} \\ 1 & a_{2jk} \\ 1 & a_{3jk} \end{bmatrix} \begin{bmatrix} u_{0_{jk}} \\ u_{1_{jk}} \end{bmatrix} \\ &+ \begin{bmatrix} 1 & a_{1jk} & b_{jk} \\ 1 & a_{2jk} & b_{jk} \\ 1 & a_{3jk} & b_{jk} \end{bmatrix} \begin{bmatrix} v_{0_k} \\ v_{1_k} \\ v_{2_k} \end{bmatrix} + \begin{bmatrix} \epsilon_{1jk} \\ \epsilon_{2jk} \\ \epsilon_{3jk} \end{bmatrix}\end{aligned} \tag{2.14}$$

It is important to note that with the matrix notation, no distinction is made between level-1, level-2 and level-3 predictors and they are all placed into the $\mathbf{X}_{jk}$ matrix.

2.5 Three-Level Joint Imputation

To date, only joint modeling has been extended to the three-level case (Asparouhov & Muthén, 2010; Yucel, 2008). However, the three-level case for joint modeling has the same limitations as previously discussed with two-level models (e.g., assuming a common distribution among the variables, inability to tailor random effects, difficulty accommodating variables at multiple levels, etc.). Thus, this thesis extends the FCS framework to three-levels in order to address the limitations of joint modeling. Because FCS's treatment of three-levels builds heavily from certain algorithmic

steps (e.g., the process of drawing parameters) a brief discussion of Yucel (2008) and Asparouhov and Muthén (2010) is needed.

Yucel (2008) proposed an extension of Schafer and Yucel's (2002) original work on two-level imputation with the joint modeling framework. As discussed earlier, one of the limiting factors of joint modeling is that it innately works best at the lowest level. In order to bypass this limitation, joint modeling essentially uses three separate Gibbs samplers (one for each level). In order to illustrate how this process works at three-levels, suppose I have the four variables in Equation 2.12: 'y' (level-1), 'a' (level-1), 'b' (level-2), and 'c' (level-3). For the purpose of this illustration, all four variables are normally distributed and incomplete. Let the analysis model I am interested in be Equation 2.12. In order to impute the four variables, the following steps would take place. Once again, in the below equations the '(a)', '(b)', '(c)', and '(y)' subscripts denote that the parameter is specific for that variable, (e.g., $v_{0_{k(a)}}$ and $v_{0_{k(y)}}$ are different values representing the level-2 random effect for the intercept of 'a' and 'y' respectively).

Step (1). Run a standard single-level joint imputation algorithm in order to impute the level-3 missing variables using the imputation model in Equation 2.15.

$$c_k^{(mis)} \sim N\left(\beta_{0(c)}, \sigma^2_{(c)}\right) \tag{2.15}$$

Run until convergence and then save the imputed values. Treat the imputed values as known values and enter them in the predictor matrix for level-2 imputation, step (2).

Step (2). Run a two-level imputation algorithm (proposed by Schafer & Yucel, 2002) in order to impute the level-2 missing variables treating the values from step (1) as known (i.e., use the filled-in level-3 variables as predictors) using the imputation

model in Equation 2.16.

$$b_{jk}^{(mis)} \sim N\left(\beta_{0(b)} + \beta_{1(b)}c_k^{(imp)} + v_{0_k(b)}, \sigma^2_{(b|c)}\right) \tag{2.16}$$

Note that the '(imp)' superscript in Equation 2.16 is a reference to the fact that these are the imputed values from Equation 2.15. Run until convergence and then save the imputed values for use in step (3), treating the imputed values as known and entering them in the predictor matrix for level-1 imputation, step (3).

Step (3). Run a three-level imputation algorithm in order to impute the level-1 missing variables, treating the values from step (1) and step (2) as known by using the imputation model in Equation 2.17.

$$\begin{bmatrix} a_{ijk}^{(mis)} \\ y_{ijk}^{(mis)} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_{0(a)} + \beta_{1(a)}c_k^{(imp)} + \beta_{2(a)}b_{jk}^{(imp)} \ldots \\ \ldots + u_{0_{jk}(a)} + v_{0_k(a)} \\ \beta_{0(y)} + \beta_{1(y)}c_k^{(imp)} + \beta_{2(y)}b_{jk}^{(imp)} \ldots \\ \ldots + u_{0_{jk}(y)} + v_{0_k(y)} \end{bmatrix}, \Sigma_{(ay|bc)}\right) \tag{2.17}$$

Run until convergence and then save the imputed values to use in step (4).

Step (4). Combine imputations from steps (1), (2), and (3) into a single imputed data set.

Step (5). Repeat step (1) to (4) until desired number of imputed data sets are generated. Alternatively, Yucel (2008) also suggest that one could run step (1) to (3) consecutively (e.g., apply each step once, then iterate the three steps, as opposed to iterating at each step before moving tot he next) and monitor convergence of the entire Gibbs sampler before generating imputations.

The limitations of the joint modeling approach become apparent upon examining the imputation models for each step (Equations 2.15 to 2.17). More specifically, Equation 2.15 lacks any conditioning on the lower levels. Thus, the lack of correlation between level-3 and the lower levels assumes that MAR is satisfied by the complete variables at level-3 only. This can be extended to level-2, where MAR must be satisfied by the level-2 or level-3 variables only. In addition, the lack of random slopes present in the model becomes apparent, which are present in the analysis model (Equation 2.12). As with joint modeling at two-levels, these random slopes cannot be specified between missing variables and the residual covariance matrix, $\Sigma_{ay|bc}$, cannot accommodate the association due to the assumption that the residual covariance matrix is homogenous across clusters. More importantly, notice the misspecification in Equation 2.16. The random slope between the level-2 variable b_{jk} and the level-3 variable y_{ijk} cannot be modeled due to the nature of how joint modeling is treating the levels (i.e., b_{jk} can affect the imputations of y_{ijk}, but y_{ijk} cannot affect the imputations of b_{jk}, through cluster means).

In addition to Yucel's approach to joint modeling, the software package Mplus (Muthén & Muthén, 1998-2015) has extended Asparouhov and Muthén's (2010) approach to single and two-level modeling to three-levels as well. Up to this point, this thesis has mainly focused on multiple imputation with joint modeling using the framework popularized by Schafer (1997a). However, there is also the joint modeling framework proposed by Asparouhov and Muthén (2010), in which all variables (complete and incomplete) are treated as dependent variables. To continue the example using y_{ijk}, a_{ijk}, b_{jk}, and c_k under Asparouhov and Muthén's (2010) proposed framework, the imputation model at iteration t is as follows:

$$
\begin{aligned}
\begin{bmatrix} y_{ijk}^{(t)} \\ a_{ijk}^{(t)} \\ b_{jk}^{(t)} \\ c_{k}^{(t)} \end{bmatrix} &\sim MVN \left(\begin{bmatrix} \beta_{0(y)}^{(t)} + u_{0_{jk}(y)}^{(t)} + v_{0_k(y)}^{(t)} \\ \beta_{0(a)}^{(t)} + u_{0_{jk}(a)}^{(t)} + v_{0_k(a)}^{(t)} \\ \beta_{0(b)}^{(t)} + u_{0_{jk}(b)}^{(t)} + v_{0_k(b)}^{(t)} \\ \beta_{0(c)}^{(t)} + u_{0_{jk}(c)}^{(t)} + v_{0_k(c)}^{(t)} \end{bmatrix}, \Sigma_{\epsilon}^{(t)} \right) \\
\mathbf{u}_{0_{jk}}^{(t)} &\sim MVN \left(\mathbf{0}, \Sigma_{u}^{(t)} \right) \\
\mathbf{v}_{0_k}^{(t)} &\sim MVN \left(\mathbf{0}, \Sigma_{v}^{(t)} \right)
\end{aligned}
\tag{2.18}
$$

where $\mathbf{u}_{0_{jk}}$ and $\mathbf{v}_{0_k}$ are vectors of the level-2 and level-3 random intercepts respectively. Furthermore, Σ_ϵ, Σ_u, and Σ_v are unstructured level-1, 2, and 3 covariance matrices that are sampled from their appropriate posterior distributions. Equation 2.18 overcomes the limitation of including variables across multiple levels by placing constraints on Σ_ϵ and Σ_u. For example, with b_{jk} the matrix Σ_ϵ would be constrained to have 0's for all variances and covariances related to b_{jk} because b_{jk} lacks level-1 variation in the variable. In the case of c_k, both Σ_ϵ and Σ_u would have similar constraints. Thus, $u_{0_{jk(c)}}$ would in essence be a constant equal to 0 and provide no variation at level-2 for c_k. Asparouhov and Muthén's (2010) approach however does not allow for random slopes due the same reasons previously described (i.e., assuming homogenous residual variances across clusters).

2.6 Purpose

To reiterate the goal of this thesis, I have extended FCS to handle three-level data structures in order to overcome some of the limitations of existing methods. As a brief overview (and more detail is presented in the Methods section), the FCS procedure that I extended begins by imputing level-1 variables, iterating through each variable

once. In a similar fashion, the Gibbs sampler imputes level-2 variables by treating the between-cluster means of the incomplete level-1 variables as known and iterating through each variable once. Finally, the highest level (level-3) is imputed, one variable at a time, by treating the cluster means of the level-1 and level-2 variables as known predictors. Each step treats the previous steps imputed values as known, placing the imputed plausible values into the appropriate matrices of subsequent variables equations (i.e., $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$ matrices in equation 2.13). This aforementioned process continues to iterate until convergence, at which point an imputation is saved. The chain subsequently continues, saving an imputation once the appropriate between-imputation thinning interval has been reached and stopping once the desired number of imputations is obtained.

As described earlier, due to FCS's ability to breakup the computational problem of obtaining a posterior predictive distribution for the incomplete variables, such that the imputation model can be tailored for each incomplete variable, one is able to overcome many of the limitations of joint imputation. For example, because each variable is handled separately and the algorithm draws from a conditional univariate distribution, one may specify a random slope with complete and incomplete variables alike. Furthermore, the flexibility of FCS allows researchers to accommodate continuous, categorical, count, and other types of data. Although this functionality can easily be added, it is beyond the scope of my thesis. Finally, FCS algorithm affords researchers with the ability to easily incorporate variables at all three-levels into the Gibbs sampler. Case in point, in the current formulation of joint modeling's Gibbs sampler, the three-levels are separate, treating the variables as orthogonal to incomplete lower level variables (i.e., the level-3 variables are imputed without any information from the incomplete level-1 and level-2 variables, and the level-2 variables are imputed without any information from the incomplete level-1 variables).

Thus a variable is assumed to be uncorrelated with the between-cluster variation in lower level variables. Furthermore, the lack of correlation between levels makes strict assumptions about the missing data mechanism (i.e., MAR is satisfied by complete variables at the current and higher levels) and a more general approach ought to be used. Due to the iterative nature of the FCS algorithm, one is able to include the cluster means at the higher levels (e.g., an incomplete level-1s cluster mean used in the prediction of a level-2 incomplete variable). This follows standard multilevel theory, where the means of lower level variables are associated with higher level variables.

Chapter 3

Methods

Thus far this thesis has investigated the theoretical background and explained some of the rational behind the choices that have been made. This section will now explicate the specific algorithm that was used. Several sources provide the distributional draws and priors that are used (i.e., W. Browne & Draper, 2000; W. J. Browne, 1998; Goldstein et al., 2007; Kasim & Raudenbush, 1998; Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008), therefore I have provided the conditional draws made in Appendix A only. Furthermore, at the end of this section, I describe the simulation that was used to compare the method to existing implemented methods.

3.1 Three-Level FCS Imputation Algorithm

The goals of this thesis were accomplished by extending the current implementation of FCS at two-levels (Keller & Enders, 2014, May). The implementation of FCS used by Keller and Enders (2014, May) is a general latent framework for two-level imputation, able to accommodate continuous, ordinal, and nominal variables. In the interest of simplicity, I only addressed the continuous case. First, I will present a brief discussion of the steps I implemented. A more detailed and technical treatment of the algorithm used to implement FCS and the extension to three-levels is presented in Appendix A.

Returning to the four variable example that was used to describe the joint modeling steps in Section 2.5. Recall that the analysis model was that of Equation 2.12 and all four variables are normally distributed and incomplete. As with the two-level case, FCS takes a much different approach to estimating the posterior predictive distribution and iterates variable by variable treating subsequent draws as known. In

order to accomplish this for the illustration, the following steps would take place in the Gibbs Sampler for iteration t.

Step (1). Run one iteration of a three-level FCS algorithm in order to draw the level-1 missing variable, a from the following distribution:

$$a_{ijk}^{(t)} \sim N \left(\begin{array}{l} \beta_{0(a)}^{(t)} + \beta_{2(a)}^{(t)} c_k^{(t-1)} + \beta_{2(a)}^{(t)} b_{jk}^{(t-1)} + \beta_{3(a)}^{(t)} y_{ijk}^{(t-1)} + \dots \\ \dots u_{0_{jk}(a)}^{(t)} + v_{0_k(a)}^{(t)} + y_{ijk}^{(t-1)} \cdot \left(u_{1_{jk}(a)}^{(t)} + v_{1_k(a)}^{(t)} \right) \end{array}, \sigma_{(a|cby)}^{2(t)} \right) \tag{3.1}$$

Next draw y from the following distribution (note the order between a and y is arbitrary):

$$y_{ijk}^{(t)} \sim N \left(\begin{array}{l} \beta_{0(y)}^{(t)} + \beta_{1(y)}^{(t)} c_k^{(t-1)} + \beta_{2(y)}^{(t)} b_{jk}^{(t-1)} + \beta_{3(y)}^{(t)} a_{ijk}^{(t)} + u_{0_{jk}(y)}^{(t)} \dots \\ \dots + v_{0_k(y)}^{(t)} + a_{ijk}^{(t)} \cdot \left(u_{1_{jk}(y)}^{(t)} + v_{1_k(y)}^{(t)} \right) + b_{jk}^{(t-1)} v_{2_k(y)}^{(t)} \end{array}, \sigma_{(y|abc)}^{2(t)} \right) \tag{3.2}$$

Step (2). Run one iteration of a two-level FCS algorithm in order to draw the level-2 missing variable, b, treating the cluster means of the imputed variables from step (1) as known (denoted with a 'bar' above them).

$$b_{jk}^{(t)} \sim N \left(\begin{array}{l} \beta_{0(b)}^{(t)} + \beta_{1(b)}^{(t)} c_k^{(t-1)} + \beta_{2(b)}^{(t)} \bar{a}_{jk}^{(b,t)} + \dots \\ \dots \beta_{3(b)}^{(t)} \bar{y}_{jk}^{(t)} + v_{0_k(b)}^{(t)} + \bar{y}_{jk}^{(t)} v_{1_k(b)}^{(t)} \end{array}, \sigma_{(b|acy)}^{2(t)} \right) \tag{3.3}$$

Step (3). Run one iteration of a single-level FCS algorithm in order to draw the level-3 missing variable, c, using the means for cluster k from the imputed values of the level-1 and level-2 variables (treating them as known).

$$c_k^{(t)} \sim N \left(\beta_{0(c)}^{(t)} + \beta_{1(c)}^{(t)} \bar{b}_k^{(t)} + \beta_{2(c)}^{(t)} \bar{a}_k^{(t)} + \beta_{3(c)}^{(t)} \bar{y}_k^{(t)}, \sigma_{(c|aby)}^{2(t)} \right) \tag{3.4}$$

Then repeat Steps (1) to (3) until one reaches the desired 'burn-in' iteration (assessed by convergence diagnostics). Once the burn-in iterations have finished, one can begin imputing the first data set, where the values are saved as the first imputation. Next one would run enough iterations until the between-imputation thinning interval is reached, thus saving another imputation. This process can be continued until the desired amount of imputations are saved.

It is important to point out how the FCS framework is allowing for all random slopes to be specified. Looking at Equation 3.1 and 3.2, one can see the inclusion of random slopes in the imputation model. Furthermore, Equation 3.3 also has a random slope between $\bar{y}_{jk}$ and b_{jk}. Contrast this to Equation 2.16 where the random slope is not present. This is just one example that exemplifies the flexibility that FCS affords. While the previous steps are a very brief overview of the extension, Appendix A highlights the conditional draws made and pseudocode for the algorithm.

For simplicity, Equations 3.1 to 3.3 have excluded the cluster means of variables that have variation at multiple levels (e.g., in Equation 3.1, the terms $\bar{y}_{jk}^{(t-1)}$, $\bar{y}_{k}^{(t-1)}$, and $\bar{b}_{k}^{(t-1)}$ along with their regression coefficients are excluded). Mistler (2015) suggests that these cluster means are required in FCS if contextual effects are present. This can be seen as a more general model and is what was implemented, however, they have been excluded in this presentation.

3.2 Simulation

A simulation was performed to test the accuracy of the FCS algorithm previously described. A 2×3 design was used with 1000 replications in each cell. The first factor was missing data rate (20% and 40% missingness). The second factor was a "within-subjects" factor, where the different levels are methods of handling missing data at three levels: listwise deletion, Asparouhov and Muthén's (2010) JM, and the

FCS approach described in Section 3.1. The reason for Yucel's (2008) JM approach being excluded from the simulation was the lack of a software package currently implementing the method. Given the aim of this simulation was to compare the FCS algorithm to what researchers currently have access to, I decided to only use the readily available methods. Finally, a random intercept model was only investigated in the simulation. For the remaining part of this chapter, I will describe the method used to generate the three-level data, how missingness was created, and how the data was imputed and analyzed.

Data generation. The data was generated using R programming language for statistical computing (R Core Team, 2015). A total of seven variables was generated with the following separation at each level:

Level-1: Two incomplete (y and a), one complete ($AV1$),

Level-2: One incomplete (b), one complete ($AV2$),

Level-3: One incomplete (c), one complete ($AV3$).

The number of observations used was constant across conditions, with 30 clusters at level-3, 50 clusters at level-2, and 10 observations per level-2 cluster. In total, there were 15,000 observations per data set. The data structure was designed to represent a longitudinal data set, where level-1 represents repeated measures, level-2 represents participants, and level-3 represents a clustering variable (e.g., hospitals, schools, etc.).

The independent variables in the data generation model (a, b, and c) were generated by summing together three orthogonal deviation scores (denoted with a 'L' superscript followed by the number that corresponds to the level). The dependent variables in the data generation model (y, $AV1$, $AV2$, and $AV3$) were generated as a linear combination of the independent variables. Table 1 shows the decomposition of

the variances used for each variable by level. The data generation model used by the simulation is presented in Equation 3.5

$$\begin{aligned} y_{ijk} =& \beta_0 + \beta_1 \left(a_i^{L1}\right) + \beta_2 \left(a_j^{L2}\right) + \beta_3 \left(a_k^{L3}\right) + \beta_4 \left(b_j^{L2}\right) + \\ & \beta_5 \left(b_k^{L3}\right) + \beta_6 \left(c_k^{L3}\right) + u_{0_{jk}} + v_{0_k} + \epsilon_{ijk} \end{aligned} \tag{3.5}$$

Table 2 through 4 give the level-specific correlation matrices used in the simulation. They describe the population correlation matrices that were used for each level. In general, each auxiliary variable (denoted with AV) was associated with all deviation scores (or residuals in the case of y) with a Pearson's $\rho = 0.40$. This is inline with Collins et al. (2001), which suggested that the cause of missingness ought to be correlated with the incomplete variable at least 0.40 in order to see bias introduced. The correlations of all other pairs of variables were equal to 0.30, which is based on Cohen's (1988) convention of a medium effect size in social science. In order to generate the specific data, the following steps were used:

Step (1). Compute the level-specific regression coefficients for the multivariate regression of level m:

$$\mathbf{Y}_m = \mathbf{X}_m \boldsymbol{\beta}_m + \mathbf{R}_{0m} \tag{3.6}$$

where $\mathbf{Y}_m$ is a n_m x 2 matrix containing the level-specific deviation scores for the dependent variables (y and AVm), $\mathbf{X}_m$ is a n_m x 3 matrix containing the level-m deviation scores for the independent variables (a, b, and c), $\boldsymbol{\beta}_m$ is a 3×2 matrix of coefficients, and $\mathbf{R}_{0m}$ are the residuals. Thus, solving for $\boldsymbol{\beta}_m$ results in Equation 3.7:

$$\boldsymbol{\beta}_m = \Sigma_{\mathbf{X}_m}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}_m} \tag{3.7}$$

where $\Sigma_{\mathbf{X}_m}$ is a 3×3 covariance matrix of the independent variables and $\Sigma_{\mathbf{XY}_m}$ is a 3×2 matrix containing the covariances of the dependent variables with the independent variables.

Step (2). Compute the level-specific residual covariance matrix for the dependent variables using Equation 3.8.

$$\Sigma_{\mathbf{r}_{0m}} = \Sigma_{\mathbf{Y}_m} - \boldsymbol{\beta}_m^{\mathsf{T}} \Sigma_{\mathbf{XY}_m} \tag{3.8}$$

The level-specific residual covariance matrix, $\Sigma_{\mathbf{r}_{0m}}$, is later used in order to draw the residuals for each level and correlate y with the auxiliary variable.

Step (3). Draw the level-specific deviation score for the independent variables from a multivariate normal distribution:

$$\mathbf{X}_m \sim MVN\left(\mathbf{0}, \Sigma_{\mathbf{X}_m}\right) \tag{3.9}$$

Step (4). Draw the level-specific residuals, $\mathbf{r}_{0m}$, from a multivariate normal distribution:

$$\mathbf{r}_{0m} \sim MVN\left(\mathbf{0}, \Sigma_{\mathbf{r}_{0m}}\right) \tag{3.10}$$

Step (5). Compute the level-specific deviation score for the dependent variable as linear combinations of the independent variables plus the residuals:

$$\mathbf{Y}_m = \mathbf{X}_m \boldsymbol{\beta}_m + \mathbf{R}_{0m} \tag{3.11}$$

where $\mathbf{R}_{0m}$ is a n_m x 2 matrix of $\mathbf{r}_{0m}$ concatenated.

Step (6). Repeat Steps (1) to (5) for all three levels.

Step (7). Combine the orthogonal level-specific deviation scores for each variable. For example for variable y_{ijk}:

$$y_{ijk} = y_i^{L1} + y_j^{L2} + y_k^{L3} \tag{3.12}$$

All the dependent and independent variables were combined in the same process and those with no variability at a specific level were given a vector of 0's for that level-specific deviation score.

Missing data. Only MAR was implemented in the simulation. As discussed earlier, imputation assumes a MAR mechanism (which includes MCAR). A MAR mechanism was simulated by having missingness determined by the complete auxiliary variable at each level. A logistic model was used to regress a missing data indicator variable, R, (where 1 is incomplete and 0 is complete) on the auxiliary variable. The parameters of the logistic model were specified such that the pseudo r-square is approximately 0.4, which is the squared approximation of the correlation between the auxiliary variable and the missing data indicator. Furthermore, there were positive relationships between the auxiliary variables and the missing data indicator, that is, higher values on the auxiliary variables increased the likelihood of missingness. Using the logistic model resulted in a predicted probability ($\hat{p}$) of being missing for each observation. The missing data indicator was drawn from a binomial distribution with one trial, (e.g., $R_{y_{ijk}} \sim Binomial\left(1, \hat{p}_{y_{ijk}}\right)$ for individual i in level-2 cluster j within level-3 cluster k on variable y). If the missing data indicator was equal to 1 for a particular case/cluster, then the observation/cluster was coded as missing. This process was repeated for each missing variable, where the cause of missingness changes depending on the level the variable is observed at (i.e., *AV1* causes a_{ijk}, *AV2* causes b_{jk}, and *AV3* causes c_k).

Imputation of data. For JM, the imputation of the data was carried out by Mplus (version 7.11 Mac; Muthén & Muthén, 1998-2015). The imputation model used by Asparouhov and Muthén's (2010) is given in Equation 2.18. Modification to these equations were made, where the auxiliary variables are included in order to satisfy the MAR mechanism. For FCS, the imputation of data was carried out using the algorithm in this thesis (see Section 3.1). The imputation model for FCS used in the simulation is given in Equation 3.1 to 3.4. As with joint modeling, modifications were made. The auxiliary variables, and subsequently their cluster means, were included into the equations in order to satisfy the MAR mechanism. Ten imputations were generated for each replication and each imputation method. Finally, with both methods of imputation, convergence of the samplers was assessed by running convergence diagnostics on several of the generated replications. These acted as a guideline to set the burn-in iterations and the between-imputation thinning interval.

Analysis of data. All estimation and pooling of the data was carried out by Mplus (version 7.11 Mac; Muthén & Muthén, 1998-2015). The analysis model that was used for the analysis phase of imputation was a random intercept model and is shown in Equation 3.13.

$$y_{ijk} = \beta_0 + \beta_1 a_{ijk} + \beta_2 b_{jk} + \beta_3 c_k + u_{0_{jk}} + v_{0_k} + \epsilon_{ijk} \tag{3.13}$$

Instead of comparing to the population parameters, I used the mean estimates of the 1000 complete-data samples as a baseline to compare bias. The rational behind this is that the imputation methods ought not be expected to perform better than maximum likelihood estimation with complete data, which itself may yield biased estimates of certain parameters. Three measurements was used in order to assess the

accuracy of the imputation algorithm: (1) proportional bias, (2) standardized bias, and (3) confidence interval (C.I.) coverage, where proportional bias is defined as:

$$\text{proportional bias} = \frac{|\text{raw bias}|}{\text{true parameter}} \tag{3.14}$$

standardized bias defined as:

$$\text{standardized bias} = \frac{\text{raw bias}}{SD \text{ of complete-data estimates}} \tag{3.15}$$

and C.I. coverage is:

$$\text{C.I. coverage} = \frac{\text{number of replications with population parameter in C.I.}}{\text{total replications}} \tag{3.16}$$

Several criteria were used to judge bias. For proportional bias, Finch, West, and MacKinnon (1997) recommends that proportional bias does not exceed 0.10. For standardized bias, Collins et al. (2001) has suggested as ± 0.4 to ± 0.5 as a rule of thumb to determine bias estimates or not. This means that the estimate on average falls roughly plus/minus half a standard error above the estimate. For 95 % C.I. coverage, a conservative inference is considered when C.I. coverage is less than 0.95 and increased type 1 error is considered when the coverage is greater than 0.95. Bradley's (1978) suggested a "liberal" criterion for 95% C.I. coverage have a lower limit of 0.925 and an upper limit of 0.975. Finally, as previously mentioned, comparisons were made between all three approaches to handling missing data (i.e., listwise deletion, Asparouhov and Muthén's JM, and FCS approach) across the conditions.

Chapter 4

Results

Table 5 gives the average estimates in each missing data condition for the complete data, listwise deletion method, joint models method (JM), and fully conditional specification method (FCS). Note the average complete data estimates presented in Table 5 were used in the computation of the bias measures. The average complete data estimates serve as a baseline as it would be unreasonable to expect the missing data methods to out perform the complete data estimates. As a reminder, the metrics used to assess the performance of the method are given in Equations 3.14, 3.15, and 3.16. One ought to note that bias measures do not always agree with one another. For example, proportional bias is not directly impacted by sample size or cluster size, whereas standardized bias is. In contrast, standardized bias is not as affected by the scale of the variables, whereas proportional bias is. This is best illustrated by the intercept, which has a true parameter value of 0, proportional bias becomes undefined and standardized bias is not.

Table 6 reports the bias measures for listwise deletion. Except for β_3, listwise deletion performed worse than both imputation methods. For example, the average estimate of β_1 had a proportional bias of -0.19 and -0.24 (20% and 40% missing data rate respectively). In contrast, the corresponding parameter between the two imputation methods had a maximum amount of proportional bias of 0.02 (in the 40% in missing data rate for FCS). Similar results were also found for standardized bias, where β_1 had the largest standardized bias for listwise deletion (-7.03 and -8.47 for 20% and 40% missing data rate respectively). Turning to 95% confidence interval (C.I.) coverage, listwise deletion values ranged from as low as 0.02 (for β_1 in the 20% missing data rate condition) to a maximum of 0.89 (for β_1 and β_2 in the 20% condition

and β_3 in the 40% condition). Recall, Bradley (1978) suggested a "liberal" criterion for 95% C.I. coverage be equal to 0.925. These low coverage values are indicative of listwise deletion's lack of accuracy and precision. Listwise deletion did perform better than both imputation methods on one parameter, β_3, this issue is discussed in more detail in Chapter 5.

Table 7 contains the bias measures for JM. Many of JM's average estimates had less than ± 0.01 proportional bias (i.e., β_1, β_2, $Var(\epsilon_{ijk})$, and $Var(v_{jk})$) for both missing data rates. In fact, in both conditions, all but one estimate, β_3, was considered unbiased following the guidelines suggested by Finch et al. (1997). Turning to standardized bias, the results follow a similar pattern, with little to no bias in all estimates across all conditions. For example, the highest standardized bias found for JM was -0.19 for β_3 in the 40% missing data rate condition. Note that β_3 is considered unbiased in terms of standardized bias when using the guidelines set forth by Collins et al. (2001). Finally, JM performed well on 95% C.I. coverage. Across both conditions, coverage was only below the 0.925 criterion three times (i.e., β_1 for 40% missing data rate and $Var(u_k)$ for both conditions). It is important to note that normal theory confidence interval coverage may not be an appropriate measure of bias for $Var(u_k)$. As a variance estimate, the sampling distribution follows a χ^2 distribution. A χ^2 distributions only approximates a normal distribution at large degrees of freedom (Box, Hunter, & Hunter, 2005).

Table 8 gives the bias measures for FCS. Similar to JM, FCS exhibited little to no bias in the proportional bias metric. The average estimates of β_1, β_2, $Var(\epsilon_{ijk})$, $Var(v_{jk})$, and $Var(u_k)$ had proportional bias less than or equal to ± 0.02 in the 20% missing data rate condition and less than or equal to ± 0.03 in the 40% missing data rate condition. As with JM, β_3 performed the worse, with -0.17 and -0.32 proportional bias in the 20% and 40% missing data rate respectively. Looking at

standardized bias, β_3 did not appear to exhibit bias. For FCS, only β_1 was considered biased under the standardized bias metric and only under the 40% missing data rate condition (Stand. Bias = 0.55). Finally, FCS performed well under the 95% C.I. coverage metric. As with JM, the same three parameter/condition combinations fell below the 0.925 criterion (i.e., β_1 for 40% missing data rate and $Var(u_k)$ for both conditions). The same issue applies for $Var(u_k)$ and this result is expected. In addition, β_1 observed slightly lower coverage in the 40% condition (C.I. = 0.90) than JM.

To facilitate the comparison between FCS and JM, Figure 1 contains trellis plots that graphically depicts the similarity between FCS and JM for proportional bias, standardized bias, and 95% C.I. coverage. The similarity is best illustrated by the proportional bias column, where across the parameters, the biases were nearly identical. Standardized bias showed a bit more disparity between the two methods. For example, the standardized bias of β_1 for FCS in the 40% missing data rate condition was equal to 0.55 and 0.00 for JM. However, it is important to note that this difference was due the small standard deviations of the complete data estimates (the denominator of the standardized bias expression, see Equation 3.15) magnifying the small disparities. In this example, β_1 estimates were based on all 15,000 observations and were very precise. When β_1 was evaluated under proportional bias, this difference was inconsequential. Furthermore, Figure 1 illustrates another similarity between FCS and JM. For both methods β_3 was biased in terms of proportional bias, but unbiased in terms of standardized bias. In contrast to β_1, β_3 was imprecise because it was based on only 30 level-3 units. Turning to 95% C.I. coverage, the only non-negligible difference was β_1 in the more extreme 40% missing data rate condition. In addition, the coverage value for $Var(u_k)$ showed poor performance in both conditions. As mentioned previous this was most likely an artifact of the fact that

the variance estimate being based on a small number of level-3 units. Taken as a whole, the tables and the trellis plots suggest that FCS and JM were equivalent in a random intercept model, on average.

The similarity of the average estimates presented in Table 5 and biases illustrated in Figure 1 leads to the question of whether, in any single data set, the two methods would lead to similar answers or are only expected to be equivalent over repeated sampling only. In order to investigate this, I treated the replications as observations and correlated the estimates obtained from JM with the corresponding estimates obtained from FCS in each condition. Correlations were estimated for each estimate across each condition. Table 9 gives the results from this analysis. The correlations ranged from $r = 0.73$ to $r = 0.98$, with the correlations for the 20% missing data rate condition (ranging $r = 0.73$ to $r = 0.79$) lower than the 40% missing data rate condition (ranging $r = 0.91$ to $r = 0.98$). Considering these correlations as a whole, the magnitude and direction suggests that FCS and JM will frequently produce similar estimates when applied to the same data set, thereby providing more evidence that the methods may be equivalent in a random intercept model.

Chapter 5

Discussion

Until now, only the joint model (JM) approach to imputation has been extended to the three-level case. Yucel (2008) took the approach of extending the two-level work done by Schafer and Yucel (2002); however, only an application of the procedure was given, and the absence of simulation evidence did not allow for an assessment of the bias associated with his method. Asparouhov and Muthén (2010) gave a closely related joint model for two-level imputation, which was extended to three-levels in the Mplus program (Muthén & Muthén, 1998-2015). Thus far, there has only been focus on the theoretical and algorithmic work required to implement three-level imputation using the JM method and there was a clear gap when it came to evaluation of three-level imputation methods. In addition, methodologists have investigated the fully conditional specification (FCS) as a method for two-level imputation (Enders et al., in press; Van Buuren et al., 2011). Despite this, there is a lack of work on using FCS to impute data with a three-level structure. Therefore, the goals of this study were to fill the gaps in the literature by developing the appropriate FCS method for three-levels, compare FCS to current software implementations of JM (Asparouhov & Muthén, 2010), and assess both models' performance with a three-level random intercept model.

In this paper I developed a Markov chain Monte Carlo algorithm by drawing on the analytic work of Browne (1998) and Yucel (2008), both of which derive the appropriate sampling steps and posterior distributions needed for the parameters of a three-level linear model. By building on their contributions, I constructed an FCS algorithm to draw from the posterior predictive distributions and form imputations for a three-level model. The procedure accomplished this by first imputing the missing level-1

variables one by one, conditioning on the previously imputed variables and complete variables at all levels. The cluster means of the level-1 variables were then computed for the level-2 clusters (treating the imputations as known). These means were then included with the level-2 and level-3 variables and were used to impute the missing level-2 variables. Consistent with level-1 imputation, the level-2 procedure cycled through the variables one at a time, conditioning on the previously imputed variables and complete variables at all levels. Finally, cluster means were computed for both level-1 and level-2 variables for the level-3 clusters. These computed cluster means were used with the level-3 variables in order to impute the missing level-3 variables one by one. Importantly, the cluster means of lower level variables were included in the imputation of higher level variables in order to preserve cross-level associations (e.g., the relation between level-3 variables and the level-3 variation in lower level variables). The notion of preserving cross-level associations has been suggested by the literature (e.g., Enders et al., in press; Gelman & Hill, 2006), but has not been fully investigated and is a direction of future research.

Simulations were conducted to evaluate the FCS approach previously described. The simulation used a random intercept model, with only the percentage of missingness being varied at 20% and 40% with a missing at random mechanism. As theoretically expected, both FCS and JM had limited bias and good coverage for most of the estimates in the conditions examined. In addition, the simulation showed imputation as a whole, under the specific conditions examined, had lower bias and better coverage than listwise deletion. This finding was in line with statistical theory (Rubin, 1976). Furthermore, the estimates, bias, and efficiency were similar for both FCS and JM, which suggests that FCS and JM were possibly equivalent for the random intercept model used in this paper. The equivalency was expected in light of Mistler (2015) demonstration of the algebraic equivalence for FCS and JM in the

context of a two-level random intercept model. A direction of future research would be the extension of Mistler's (2015) work to three-level models or more generally to any number of levels.

Although most estimates displayed little bias and good coverage, the simulations suggested that the fixed effect estimate of the level-3 predictor exhibited bias. This was an unexpected finding and the cause is unclear. One possible explanation is the number of clusters used (i.e., 30), which Hox (2010) suggests as a bare minimum for the number of clusters at the highest level in a two-level model with complete data. It could be possible that this rule of thumb does not hold in three-level models or in models with missing data at the highest level. In addition, the low number of clusters may interact with the low interclass correlation (ICC) that the simulation used at the highest level; recall the population model assigned 10% of the total variation in the variables observed at level-1 to be accounted for by level-3. The reliability of the group means of a variable is dependent on the ICC and the number of observations within a cluster (see Snijders & Bosker, 2011). This issue is a well known issue in the complete-data literature for multilevel modeling (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Lüdtke et al., 2008). Although the simulation had a large cluster size (i.e., $n_{jk} = 500$), increasing the ICC could have a positive impact on the quality of the estimates. The bias in this estimate demonstrates the lack of any clear guidelines as to how many clusters are needed at level-3 in order to obtain accurate estimates. Even though this study presented a limited set of conditions where 30 clusters appeared to be insufficient, it is the only study to examine this issue thus far, and future studies should attempt to clarify the sources of this bias.

As with any simulation study, this study has a number of limitations that ought to be noted. Most importantly, the simulation had limited conditions, implementing only two missing data rates that were constant across levels. Moreover, the cluster

sizes and ICC were kept constant and were chosen to mimic longitudinal data (with the lowest level being repeated measures). The correlation structure (i.e., effect sizes) was also kept constant, to mimic medium effect sizes in social science data (Cohen, 1988). Therefore, future studies ought to investigate the method's performance under different simulated three-level data structures, varying the number of clusters at each level, the distribution of variance across the levels, and the magnitude of the associations between the variables, among other things. Another avenue of research is to investigate random slope models in the context of FCS, which, as mentioned earlier in this paper, can potentially preserve associations. In addition, heterogeneous residual variances presents another line of important research. Within-person variability is often an outcome of interest in psychological research, and the data collected from such research (e.g., diary data) lends itself to a three-level model. The models presented in this paper are not currently equipped to preserve cluster-specific variation that is often the focus of within-person research designs (Hoffman, 2007). Methodologists have proposed approaches to investigate this (e.g. Hoffman, 2007; Kasim & Raudenbush, 1998; Yucel, 2011) and these methods could be extended to the three-level case. Finally, the simulation examined only normally distributed continuous data. Therefore, future research ought to investigate the extension of categorical and non-normal data to three-level multiple imputation.

In summary, the current literature is very sparse and not much work has been done. Researchers routinely work with data structures requiring three-level models (e.g., individuals within families within neighborhoods, children within classrooms within schools, days within people within clinics, etc.); however, missing data techniques for three-level data are still in their infancy, leaving researchers with few sophisticated options for addressing the problem. This study serves as a preliminary start in tackling a much larger issue and gap in the current missing data literature.

References

Asparouhov, T., & Muthén, B. (2010). Multiple mputation with mplus. *MPlus Web Notes*.

Box, G., Hunter, J., & Hunter, W. (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley-Interscience.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152.

Browne, W., & Draper, D. (2000). Implementation and performance issues in the bayesian and likelihood fitting of multilevel models. *Computational statistics*, *15*, 391–420.

Browne, W. J. (1998). Applying mcmc methods to multi-level models (Doctoral dissertation, University of Bath).

Carpenter, J. R., & Kenward, M. G. (2013). Statistics in practice. *Multiple Imputation and its Application*, 346–348.

Cheung, M. W.-L. (2007). Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random. *Organizational Research Methods*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330.

Enders, C. K., Mistler, S. A., & Keller, B. T. (in press). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological methods*.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87–107.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, *63*(2), 204–238.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252–286.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, *42*(4), 609–629.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

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*(2), 93–116.

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, Connecticut.

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data.

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*(4), 444.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological methods*, *13*(3), 203.

Mistler, S. A. (2015). Multilevel multiple imputation: An examination of competing methods (Doctoral dissertation). Retrieved from http://repository.asu.edu/items/29655.

Muthén, L. K., & Muthén, B. O. (1998-2015). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Roudsari, B., Field, C., & Caetano, R. (2008). Clustered and missing data in the us national trauma data bank: implications for analysis. *Injury prevention*, *14*(2), 96–100.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 543–546.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L. (1997b). Imputation of missing covariates under a general linear mixed model. *Technical report available at www.stat.psu.edu/-jls*.

Schafer, J. L. (2001). Multiple imputation with pan. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355–377). Washington, DC: American Psychological Association.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*(1), 19–35.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, *11*(2), 437–457.

Shin, Y., & Raudenbush, S. W. (2007). Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics*, *63*(4), 1262–1268.

Snijders, T., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE Publications.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219–242.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, *76*(12), 1049–1064.

Van Buuren, S., et al. (2011). Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, 173–196.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1874), 2389–2403.

Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical modelling*, *11*(4), 351–370.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association*, *86*(413), 79–86.

Zhang, D. (2005). *A monte carlo investigation of robustness to nonnormal incomplete data of multilevel modeling* (Unpublished doctoral dissertation). Texas A&M University.

APPENDIX A

ALGORITHMIC DETAILS OF FCS

In order to extend the current two-level FCS algorithm to three-levels, I will first go over the current implementation of FCS in the context of a three-level model. The rational behind this is, in terms of imputation, the first and second levels of a two-level model can be thought of as the second and third levels of a three-level model. Thus, I will refer to level-1 in a two-level model as level-2 in a three-level model.

Suppose that we are interested in one replication of the Gibbs sampler, t, with a set of Y incomplete variables, such that, $Y = \{X, Z\}$, where X is the set of level-2 incomplete variables (in a three-level model) of size M_{L2}, and Z is the set of level-3 incomplete variables (in a three-level model) of size M_{L3}. Furthermore, the subsets (i.e., X and Z) can be further decomposed into two parts, $X = \{X_{mis}, X_{obs}\}$, where $X_{mis} = \left\{x_1^{mis}, \ldots, x_{M_{L2}}^{mis}\right\}$ and $X_{obs} = \left\{x_1^{obs}, \ldots, x_{M_{L2}}^{obs}\right\}$. In addition, I define for a single variable (i.e., x_m, y_m, z_m) a set that is compliment to its super set, $X_{mis}^{\mathsf{c}(x_m)} = X_{mis} \setminus x_m^{mis}$. Finally, we have a set of variables, 'A', which contains all complete auxiliary variables. As explained previously, the FCS cycles through the incomplete variables one at a time. The pseudocode in Figure A1 describes the steps in both words and the conditional draws used for each line.

In order to extend this algorithm to three-levels, I added another step before the first for loop that deals with missing data at level-2. To illustrate, suppose that Y now can be separated into three distinct groups, $Y = \{W, X, Z\}$, where W is a set of level-1 incomplete variables of size M_{L1}. The pseudocode in Figure A2 can be added in front of the code in Figure A1 in order to extend the current FCS algorithm to impute three-levels. This pseudo-code provides an algorithmic, programmatic, and conditional probabilistic explanation of FCS and the extension to FCS for the continuous case.

1: **procedure** FCS
2: **for** $m := 1$ **to** M_{L2} **do**
3: Sample imputations for x_m from posterior predictive distribution;

$$x_{mis}^{(t)} \sim P\left(x_{mis} \mid X_{obs}, X_{mis}^{\mathsf{c}(x_m)}, A, \mathbf{V}^{(t-1)}, \Sigma_{\mathbf{V}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\right)$$

4: Sample level-3 residuals for x_m;

$$v_k^{(t)} \sim P\left(v_k \mid X_{obs}, x_{mis}^{(t)}, X_{mis}^{\mathsf{c}(x_m)}, A, \Sigma_{\mathbf{V}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\right)$$

5: Sample level-3 covariance matrix for x_m;

$$\Sigma_{\mathbf{V}}^{(t)} \sim P\left(\Sigma_{\mathbf{V}} \mid X_{obs}, x_{mis}^{(t)}, X_{mis}^{\mathsf{c}(x_m)}, Z^{(t-1)}, A, \mathbf{V}^{(t)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\right)$$

6: Sample fixed effects for w_m;

$$\sigma_{\epsilon}^{2(t)} \sim P\left(\sigma_{\epsilon}^{2} \mid X_{obs}, x_{mis}^{(t)}, X_{mis}^{\mathsf{c}(x_m)}, Z^{(t-1)}, A, \Sigma_{\mathbf{V}}^{(t)}, \mathbf{V}^{(t)}, \beta^{(t-1)}\right)$$

7: Sample level-2 residual variance for x_m;

$$\beta^{(t)} \sim P\left(\beta \mid X_{obs}, x_{mis}^{(t)}, X_{mis}^{\mathsf{c}(x_m)}, Z^{(t-1)}, A, \Sigma_{\mathbf{V}}^{(t)}, \mathbf{V}^{(t)}, \sigma_{\epsilon}^{2(t)}\right)$$

8: **end for**
9: **for** $m := 1$ **to** M_{L3} **do**
10: Sample imputations for z_m from posterior predictive distribution;

$$z_{mis}^{(t)} \sim P\left(z_{mis} \mid Z_{obs}, Z_{mis}^{\mathsf{c}(z_m)}, X^{(t)}, A, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\right)$$

11: Sample level-3 residual variance for z_m;

$$\sigma_{\epsilon}^{2(t)} \sim P\left(\sigma_{\epsilon}^{2} \mid Z_{obs}, z_{mis}^{(t)}, Z_{mis}^{\mathsf{c}(z_m)}, X^{(t)}, A, \beta^{(t-1)}\right)$$

12: Sample regression coefficients for z_m;

$$\beta^{(t)} \sim P\left(\beta \mid Z_{obs}, z_{mis}^{(t)}, Z_{mis}^{\mathsf{c}(z_m)}, X^{(t)}, A, \sigma_{\epsilon}^{2(t)}\right)$$

13: **end for**
14: **end procedure**

Figure A1: FCS two-level multiple imputation. Where $k = 1, \ldots, c$ level-2 clusters, 'P' is a general probability distribution, and '$\mathbf{B}$' is a matrix containing the random effects at level-3 with n rows and columns equal to the number of random effects per level-2 cluster. Also note that $X_{mis}^{\mathsf{c}(x_m)}$ and $Z_{mis}^{\mathsf{c}(z_m)}$ contains no iteration scripting (i.e, t). This is because $X_{mis}^{\mathsf{c}(x_m)}$ and $Z_{mis}^{\mathsf{c}(z_m)}$ vary between t and $t-1$ depending on the variables in set X and Z respectively.

1: **procedure** FCS EXTENSION
2: **for** $m := 1$ **to** M_{L1} **do**
3: Sample imputations for w_j from posterior predictive distribution;

$$w_{mis}^{(t)} \sim P\Big(w_{mis} \mid W_{obs}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \ldots \\ \ldots, \mathbf{U}^{(t-1)}, \mathbf{V}^{(t-1)}, \Sigma_{\mathbf{B}}^{(t-1)}, \Sigma_{\mathbf{C}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\Big)$$

4: Sample level-3 residuals for w_m;

$$v_{k}^{(t)} \sim P\Big(v_k \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \ldots \\ \ldots, \mathbf{U}^{(t-1)}, \Sigma_{\mathbf{U}}^{(t-1)}, \Sigma_{\mathbf{V}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\Big)$$

5: Sample level-2 residuals for w_m;

$$u_{jk}^{(t)} \sim P\Big(u_{jk} \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \ldots \\ \ldots, \mathbf{V}^{(t)}, \Sigma_{\mathbf{U}}^{(t-1)}, \Sigma_{\mathbf{V}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\Big)$$

6: Sample level-3 residuals for w_m;

$$\Sigma_{\mathbf{V}}^{(t)} \sim P\Big(\Sigma_{\mathbf{V}} \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \ldots \\ \ldots, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \Sigma_{\mathbf{U}}^{(t-1)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\Big)$$

7: Sample level-2 residuals for w_m;

$$\Sigma_{\mathbf{U}}^{(t)} \sim P\left(\Sigma_{\mathbf{U}} \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \Sigma_{\mathbf{V}}^{(t)}, \sigma_{\epsilon}^{2(t-1)}, \beta^{(t-1)}\right)$$

8: Sample level-1 residual variance for w_m;

$$\sigma_{\epsilon}^{2(t)} \sim P\left(\sigma_{\epsilon}^{2} \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \Sigma_{\mathbf{U}}^{(t)}, \Sigma_{\mathbf{V}}^{(t)}, \beta^{(t-1)}\right)$$

9: Sample fixed effects for w_m;

$$\beta^{(t)} \sim P\left(\beta^{(t)} \mid W_{obs}, w_{mis}^{(t)}, W_{mis}^{\mathsf{c}(w_m)}, X^{(t-1)}, Z^{(t-1)}, A, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \Sigma_{\mathbf{U}}^{(t)}, \Sigma_{\mathbf{V}}^{(t)}, \sigma_{\epsilon}^{2(t)}\right)$$

10: **end for**
11: **end procedure**

Figure A2: FCS three-level multiple imputation extension. Where $k = 1, \ldots, c$ level-2 clusters, $j = 1, \ldots, d_k$ level-1 clusters, '$\mathbf{C}$' is a matrix containing the random effects at level-2 with l_k rows and columns equal to the number of random effects per level-1 cluster. Also note that $W_{mis}^{\mathsf{c}(w_m)}$ contains no iteration scripting (i.e, t). This is because $W_{mis}^{\mathsf{c}(w_m)}$ varies between t and $t-1$ depending on the variables in set W.

APPENDIX B

TABLES AND FIGURES

Table 1

Variance Decomposition by Level of All Variables

Variable	Level 1	Level 2	Level 3	Total σ^2
y_{ijk}	4.0(40%)	5.0(50%)	1.0(10%)	10.0
a_{ijk}	4.0(40%)	5.0(50%)	1.0(10%)	10.0
b_{jk}	0	5.0(83%)	1.0(17%)	6.0
c_k	0	0	1.0(100%)	1.0
$AV1_{ijk}$	4.0(100%)	0	0	4.0
$AV2_{jk}$	0	5.0(100%)	0	5.0
$AV3_k$	0	0	1.0(100%)	1.0

Note. In parentheses is the percent of total variance accounted for across the row.

Table 2
Level-1 Correlation Matrix for Simulated Data.

Variable	1	2	3
1. y^{L1}	—		
2. a^{L1}	0.3	—	
3. $AV1^{L1}$	0.4	0.4	—

Table 3
Level-2 Correlation Matrix for Simulated Data.

Variable	1	2	3	4
1. y^{L2}	—			
2. a^{L2}	0.3	—		
3. b^{L2}	0.3	0.3	—	
4. $AV\mathcal{Z}^{L2}$	0.4	0.4	0.4	—

Table 4
Level-3 Correlation Matrix for Simulated Data.

Variable	1	2	3	4	5
1. y^{L3}	—				
2. a^{L3}	0.3	—			
3. b^{L3}	0.3	0.3	—		
4. c^{L3}	0.3	0.3	0.3	—	
5. $AV3^{L3}$	0.4	0.4	0.4	0.4	—

Table 5

Average Estimates for Simulation

	20% missing data rate				40% missing data rate			
	Complete	Listwise	FCS	JM	Complete	Listwise	FCS	JM
β_0	0.00	-0.61	0.00	0.00	0.00	-0.89	-0.01	-0.01
β_1	0.29	0.24	0.30	0.29	0.29	0.22	0.30	0.29
β_2	0.21	0.19	0.21	0.21	0.21	0.19	0.20	0.21
β_3	0.15	0.16	0.13	0.13	0.16	0.17	0.11	0.12
$Var(\epsilon_{ijk})$	3.64	3.48	3.64	3.64	3.64	3.45	3.63	3.64
$Var(v_{jk})$	4.32	4.19	4.32	4.32	4.33	4.18	4.34	4.32
$Var(u_k)$	0.79	0.74	0.78	0.78	0.78	0.68	0.75	0.75

Table 6

Bias Measures for Listwise Deletion

	20% missing data rate			40% missing data rate		
	Prop Bias	Stan Bias	C.I. Coverage	Prop Bias	Stan Bias	C.I. Coverage
β_0	—	-3.33	0.16	—	-5.08	0.07
β_1	-0.19	-7.03	0.02	-0.24	-8.47	0.09
β_2	-0.09	-0.79	0.89	-0.11	-0.95	0.88
β_3	0.05	0.04	0.89	0.10	0.08	0.89
$Var(\epsilon_{ijk})$	-0.04	-3.57	0.41	-0.05	-4.50	0.59
$Var(v_{jk})$	-0.03	-0.74	0.87	-0.03	-0.80	0.89
$Var(u_k)$	-0.06	-0.22	0.84	-0.12	-0.38	0.79

Note. 'Prop' and 'Stan' are abbreviations for 'Proportional' and 'Standardized' respectively.

Table 7
Bias Measures for JM

	20% missing data rate			40% missing data rate		
	Prop Bias	Stan Bias	C.I. Coverage	Prop Bias	Stan Bias	C.I. Coverage
β_0	—	-0.01	0.93	—	-0.06	0.94
β_1	0.00	-0.02	0.95	0.00	0.00	0.92
β_2	0.00	-0.02	0.94	-0.01	-0.05	0.94
β_3	-0.13	-0.11	0.93	-0.22	-0.19	0.95
$Var(\epsilon_{ijk})$	0.00	0.05	0.93	0.00	-0.01	0.94
$Var(v_{jk})$	0.00	-0.01	0.94	0.00	0.00	0.94
$Var(u_k)$	-0.01	-0.05	0.90	-0.04	-0.12	0.88

Note. 'Prop' and 'Stan' are abbreviations for 'Proportional' and 'Standardized' respectively.

Table 8
Bias Measures for FCS

	20% missing data rate			40% missing data rate		
	Prop Bias	Stan Bias	C.I. Coverage	Prop Bias	Stan Bias	C.I. Coverage
β_0	—	-0.02	0.93	—	-0.07	0.95
β_1	0.01	0.29	0.93	0.02	0.55	0.90
β_2	-0.02	-0.15	0.94	-0.04	-0.31	0.94
β_3	-0.17	-0.14	0.93	-0.32	-0.27	0.95
$Var(\epsilon_{ijk})$	0.00	0.02	0.93	0.00	-0.10	0.93
$Var(v_{jk})$	0.00	0.04	0.94	0.00	0.10	0.95
$Var(u_k)$	-0.01	-0.05	0.89	-0.03	-0.11	0.88

Note. 'Prop' and 'Stan' are abbreviations for 'Proportional' and 'Standardized' respectively.

Table 9
Correlations of FCS and JM Estimates

	Missing data rate	
	20%	40%
β_0	0.78	0.98
β_1	0.79	0.94
β_2	0.80	0.95
β_3	0.73	0.91
$Var(\epsilon_{ijk})$	0.79	0.95
$Var(v_{jk})$	0.78	0.98
$Var(u_k)$	0.79	0.98

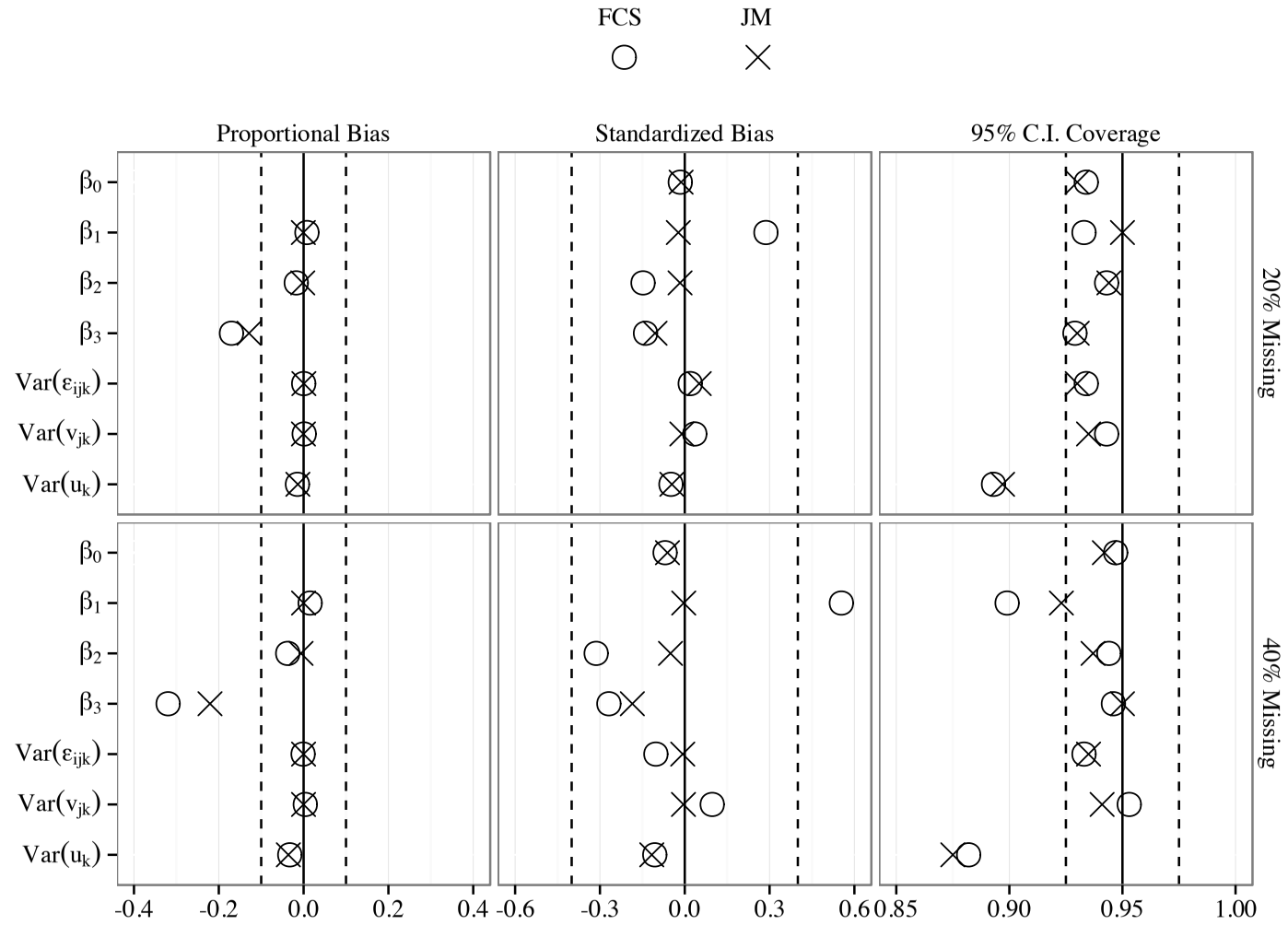

Figure 1: Trellis plot for three measures of bias. The solid black line represents the point of no bias for the measure and the dashed lines represent the acceptable limit of bias (from left to right, ± 0.1, ± 0.4, and 0.925 and 0.975 respectively). Note that β_0 for proportional bias is excluded in the figure due to the complete data estimate being close to zero, thus small deviations cause very large bias values.