Multiple Imputation for Two-Level Hierarchical Models with Categorical Variables and Missing at Random Data

by

Katie L. Kunze

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2016 by the
Graduate Supervisory Committee

Roy Levy, Chair
Craig K. Enders
Marilyn S. Thompson

ARIZONA STATE UNIVERSITY

December 2016

ABSTRACT

Accurate data analysis and interpretation of results may be influenced by many potential factors. The factors of interest in the current work are the chosen analysis model(s), the presence of missing data, and the type(s) of data collected. If analysis models are used which a) do not accurately capture the structure of relationships in the data such as clustered/hierarchical data, b) do not allow or control for missing values present in the data, or c) do not accurately compensate for different data types such as categorical data, then the assumptions associated with the model have not been met and the results of the analysis may be inaccurate. In the presence of clustered/nested data, hierarchical linear modeling or multilevel modeling (MLM; Raudenbush & Bryk, 2002) has the ability to predict outcomes for each level of analysis and across multiple levels (accounting for relationships between levels) providing a significant advantage over single-level analyses. When multilevel data contain missingness, multilevel multiple imputation (MLMI) techniques may be used to model both the missingness and the clustered nature of the data. With categorical multilevel data with missingness, categorical MLMI must be used. Two such routines for MLMI with continuous and categorical data were explored with missing at random (MAR) data: a formal Bayesian imputation and analysis routine in JAGS (R/JAGS) and a common MLM procedure of imputation via Bayesian estimation in BLImP with frequentist analysis of the multilevel model in Mplus (BLImP/Mplus). Manipulated variables included interclass correlations, number of clusters, and the rate of missingness. Results showed that with continuous data, R/JAGS returned more accurate parameter estimates than BLImP/Mplus for almost all parameters of interest across levels of the manipulated variables. Both R/JAGS and

BLImP/Mplus encountered convergence issues and returned inaccurate parameter estimates when imputing and analyzing dichotomous data. Follow-up studies showed that JAGS and BLImP returned similar imputed datasets but the choice of analysis software for MLM impacted the recovery of accurate parameter estimates. Implications of these findings and recommendations for further research will be discussed.

To my darling Jonathan.

My partner through sleepless nights, stressful days, joyful moments, and little victories.

You are so loved.

ACKNOWLEDGMENTS

When I began this journey, I started it with the reassurance and full support of my parents, Brianne and David Poole, and my now-husband (then-boyfriend), Jonathan Kunze. Without these three individuals, I would not have had the focus to begin my studies. Now that I am nearing the end of my doctoral work, I can say that I owe my completion of this process to them as well. They never doubted for a moment that I would finish and finish well.

In addition, there are several other influential individuals that I would like to thank specifically. Foremost, I would like to thank the members of my dissertation committee.

To my advisor, Dr. Roy Levy, I am forever grateful for your patient guidance at each stage, your careful correction of my missteps along the way, and your honest encouragement to continue in this process when I was not able to see a way forward. I am honored to be a branch in the Mislevy/Levy lineage, and I hope I live up to your high research standards.

To Dr. Craig Enders, thank you for sharing a small corner of your field and much of your great expertise with me. You laid the groundwork to make this work feasible, and completion of this project would not have been possible without your kindness and willingness to help see it to fruition.

To Dr. Marilyn Thompson, your applied research experience brought me back to reality whenever I wandered too far into the world of simulation studies. I appreciate your efforts to make this work accessible for educational researchers. I am also grateful

for your help throughout my studies. Both you and Dr. Green have been supportive in so many ways, and I am thankful for you both.

To my fellow students (whom I already—or will soon—call "Dr."), this process would not have been nearly as enjoyable or sustainable without you. To Nedim Yel and Yuning Xu, I am thankful to have had you both sharing in this process with me and making progress alongside me. This would have been a long series of many lonely months without you to commiserate and celebrate with me. To Erin Kube for her friendship and faith in me. To Dubravka Svetina, AAron Crawford, and Derek Fay for being the wonderful MSMSers that they are.

To the baristas at Press Coffee, you kept me both caffeinated and motivated. Thanks for not kicking me out!

To my family (Pooles and Kunzes alike), my AZ family, my dear friends, and my church family at North Bible Church, I am thankful for your faithful prayers and thoughtful encouragement throughout the years.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

2 BACKGROUND LITERATURE 4

Hierarchical/Multilevel Modeling 4

Variable Properties 5

Common Multilevel Models 6

Models without Predictors 6

Models with Predictors at L2 Only 9

Models with L1 Predictors 10

Models with Cross-level Effects 13

Interaction Effects 14

Additional Complexity 16

Longitudinal Data 16

Categorical Outcomes 16

Missing Data Analysis 17

Missing Data Patterns 18

Missing Data Mechanisms 20

Missing Data Handling 22

Traditional Missing Data Handling Procedures 22

CHAPTER Page

Modern Missing Data Handling Procedures 23

Categorical Data Analysis 29

Multilevel Modeling with Missing Data 30

MLM with Categorical Data 36

MLM with Missing Data and Categorical 37

3 METHODOLOGY 43

Modeling Decisions 44

Data Generation 45

Manipulated Variables 45

Modeling Frameworks and Software 45

Dataset Characteristics 46

Students within Classrooms/Schools 46

Interclass Correlations 47

Percent Missingness 48

Patterns of Missingness 48

Variable Types 50

Continuous Variables 50

Categorical Variables 51

Category Proportions for Discrete Variables 51

Computational Details 52

Computational Performance 52

Assessment of Results 52

CHAPTER Page

Bias 53

Root Mean Square Error 53

Confidence Interval Coverage 53

4 RESULTS 55

Convergence Issues 56

Continuous Data 61

Fixed Effects Average Bias 65

Fixed Effects CI Coverage 66

Fixed Effects RMSE 67

Random Effects Average Bias 69

Random Effects CI Coverage 70

Random Effects RMSE 72

Categorical Data with .50/.50 Split 73

Fixed Effects Average Bias 73

Random Effects Average Bias 74

Fixed Effects CI Coverage 76

Random Effects CI Coverage 78

Fixed Effects RMSE 80

Random Effects RMSE 83

Follow-up Analyses 85

Influence of Missing Data Rates 85

Fixed Effects 86

CHAPTER Page

Random Effects....87

Conclusions....89

Imputation in R/JAGS with Analysis in Mplus....90

Fixed Effects Yjags....91

Random Effects Yjags....92

Fixed Effects YX2jags....92

Random Effects YX2jags....95

5 DISCUSSION AND CONCLUSIONS....98

Continuous Data....100

Categorical Data....102

Follow-up Analyses....104

V3 Missingness Pattern....104

Yjags and YX2jags....105

Implications....106

Conclusions....108

REFERENCES....110

APPENDIX

A YX2V3 BLIMP MODEL CODE WITH CONTINUOUS DATA....115

B B YX2V3 BLIMP MODEL CODE WITH DICHOTOMOUS DATA....117

C YX2V3 MPLUS MODEL CODE WITH CONTINUOUS DATA....119

D YX2V3 MPLUS MODEL CODE WITH DICHOTOMOUS DATA....121

E YX2V3 JAGS MODEL CODE WITH CONTINUOUS DATA....123

APPENDIX Page

F YX2V3 JAGS MODEL CODE WITH DICHOTOMOUS DATA 127

LIST OF TABLES

Table Page

1. Converged Replications for R/JAGS and BLImP/Mplus Complete Data, V3 Missingness, and X2 Missingness Conditions....58
2. Converged Replications for R/JAGS and BLImP/Mplus for X2V3 Missingness, Y Missingness, Yjags, and YX2jags Conditions....59
3. Converged Replications for R/JAGS and BLImP/Mplus for YX2 Missingness and YX2V3 Missingness Conditions....60
4. Symbol Key for Interpreting Figures of Results....62

LIST OF FIGURES

Figure Page

1. Example of MLM with Varying Intercepts for Each Cluster *j* and the Average Intercept and Slope Averaged across *j*. 12
2. Comparison of Cluster Differences between Groups with High Average Exam Scores and Low Average Exam Scores 15
3. Four Typical Missing Data Patterns in the Multilevel Data with Two Levels and Three Groups 33
4. Average Bias, Confidence Interval Coverage, and RMSE of the Fixed Effects for Continuous Data....63
5. Average Bias, Confidence Interval Coverage, and RMSE of the Random Effects for Continuous Data....71
6. Average Bias of the Fixed Effects for .50/.50 Split and Continuous Data 75
7. Average Bias of the Random Effects for .50/.50 Split and Continuous Data....77
8. Confidence Interval Coverage of the Fixed Effects for .50/.50 Split and Continuous Data 79
9. Confidence Interval Coverage of the Random Effects for .50/.50 Split and Continuous Data 81
10. RMSE of the Fixed Effects for .50/.50 Split and Continuous Data 82
11. RMSE of the Random Effects for .50/.50 Split and Continuous Data 84
12. Comparison of BLImP/Mplus Fixed Effects Results with V3 Missingness for Different Methods of Imposing Missingness....88

Figure Page

13. Comparison of BLImP/Mplus Random Effects Results with V3 Missingness for Different Methods of Imposing Missingness.....89

14. Comparison of Fixed Effects Results for Y Missingness with Data Imputed/Analyzed in R/JAGS, Imputed/Analyzed in BLImP/Mplus , and Imputed in R/JAGS Then Analyzed in Mplus.....93

15. Comparison of Random Effects Results for Y Missingness with Data Imputed/Analyzed in R/JAGS, Imputed/Analyzed in BLImP/Mplus , and Imputed in R/JAGS Then Analyzed in Mplus 94

16. Comparison of Fixed Effects Results for YX2 Missingness with Data Imputed/Analyzed in R/JAGS, Imputed/Analyzed in BLImP/Mplus , and Imputed in R/JAGS Then Analyzed in Mplus 96

17. Comparison of Random Effects Results for YX2 Missingness with Data Imputed/Analyzed in R/JAGS, Imputed/Analyzed in BLImP/Mplus , and Imputed in R/JAGS Then Analyzed in Mplus 97

Chapter 1

INTRODUCTION

Modern life contains seemingly endless opportunities for data to be gathered on a plethora of human characteristics including our attitudes, opinions, experiences, interactions, skills, knowledge, and behavior. These data that are collected from individuals may be used for a wide variety of purposes, but the factors which can affect accurate analysis and interpretation of the data tend to be universal. Some of these factors include the measurement procedures used, data management practices, use of human raters or interviewers, data collection procedures, the chosen analysis model(s), the presence of missing data, and the type(s) of data collected. Although all of these influence the results and interpretability of a study, these last three factors are of great interest—especially when modeling choices, data type, and missingness intersect. For instance, if an analysis model or set of models are used which do not accurately capture the structure of relationships in the data or do not accurately compensate for different data types (e.g., categorical variables), then the assumptions associated with the model have not been met and the results may be inaccurate. Similarly, if the analysis makes specific assumptions about the presence, absence, or nature of data that are missing or unreported (i.e. missing data), the analysis results may again be affected.

One way in which a model may not accurately capture the variability in the data occurs when the data are naturally clustered in some way. Clustering may be present when individuals have shared experiences, originate from common sources, or are grouped together specifically for research purposes (e.g. treatment and control groups). When clustered data are modeled in a way which does not account for the clustering or

when non-clustered data are modeled in a way intended for clustering, the analysis model and, therefore, the results may not accurately reflect the relationships present in the data. One common form of clustered data is hierarchical or multilevel data. Data are grouped within clusters at higher/subsequent levels. Using a model which properly accounts for the clustering present in the data will lead to more accurate results.

Examples of clustered data include students (in a classroom, school, state, etc.), employees (within project teams, departments, managers, etc.), individuals in a family, some other grouping, or multiple data points collected from the same person. Often in practice, this natural hierarchy of the data is ignored during analysis and reporting. In other words, researchers work from a single-level theoretical framework when selecting statistical models, analyzing their data, and interpreting the results. This practice violates the independence assumptions (individuals/observations are independent of each other) of many common statistical analyses. If data are not independent and are clustered in some way, a more accurate assessment of the relationships among variables is possible when the clustering that is present in the data is identified and modeled in the analysis. Research has shown that choosing to ignore the multilevel structure and modeling multilevel data using a single-level analysis may produce inaccurate results which misrepresent the relationships among variables (Clarke, 2008; Garson, 2013; Guo & Zhao, 2000; Lee & Bryk, 1989; Steenbergen & Jones, 2002).

In much the same way, if data are collected in a categorical or discrete form, they cannot be analyzed using models created for continuous data. Modeling categorical data using continuous modeling procedures violates assumptions of normality (or other distributional assumptions) which are common to many analysis models, and the results

of the analysis, again, may not be accurate. In educational data, examples of categorical variables include demographic variables such as gender or ethnicity, measures of student ability (e.g., reads at grade level vs. reads below grade level), or aggregate measures of student success (e.g., National Assessment of Educational Progress; NAEP). In job satisfaction or survey data, variables measured as categorical might be ratings of job satisfaction or an indicator of whether or not the individual has been promoted in the last year.

Another key factor which may impact analysis results is the absence of data or missing data (i.e., missingness). Modeling procedures rely on assumptions about the data and its representativeness. When data are incomplete in some way, the analysis results and interpretation may not be accurate. Missingness is often conceived of as existing in a particular item or variable, but may also occur across items or across surveys or data collection instances. Some specific examples of missingness include individuals who skip items or sets of items on a survey, data points which are not collected by design, and data which are not reported or are missing in an archive. The current study focuses on the combined impact of hierarchical data and missing data on recovery of accurate analysis results.

Chapter 2

BACKGROUND LITERATURE

Hierarchical/Multilevel Modeling

One modeling framework which directly addresses nested or clustered data is hierarchical linear modeling (HLM) which has also been called multilevel linear modeling or multilevel modeling (MLM[1]). This family of models has been designed to assess the influences of variables across multiple known groups of individuals and/or across sources of data within an individual (Raudenbush & Bryk, 2002). In this framework, levels are described in terms of the unit of analysis. Using one of the previous examples, a single-level model might include variables associated with students or variables associated with schools but not both. In this example, the unit of analysis would either be students or schools. If we were to examine a two-level model of individual students who are associated with (nested within) each of their respective schools, then the unit of analysis for level-one (L1) would be students who are nested within the level-two (L2) unit of schools.

In the presence of clustered/nested[2] data, the ability of MLM to predict outcomes for each level of analysis and across multiple levels (accounting for relationships between levels) provides a significant advantage over single-level analyses. This flexibility allows researchers to examine separate research questions at each level. Alternatively, a single-level analysis would examine outcomes for either the schools or students but cannot

[1] Some make a distinction between these frameworks, but the current work will use MLM to refer to hierarchical linear modeling, multilevel linear modeling, and multilevel modeling. MLM is also used to refer to multilevel models themselves.

[2] The terms clustered, nested, hierarchical, and multilevel will be used interchangeably.

capture relationships between these levels. Typically, MLM analyses use forms of regression to capture predictive relationships, although other analysis routines and modeling frameworks can be used or altered to accommodate multilevel data (e.g. multilevel item response theory).

Variable properties. Variables in an MLM may belong to any level in the hierarchy, but commonly the dependent variable or outcome is measured at L1. There are two ways in which we can conceptualize each predictor variable (independent variable) in an MLM as: (a) belonging to a specific level (i.e., a variable collected at L1 is a separate variable from the aggregated L2 average of that L1 variable), or (b) belonging to any level but allowed to vary at any one level or at more than one of the levels. We can determine the level at which a variable varies by considering the level at which a variable is measured. Variables at the highest level in the hierarchy have one source of variability which means they only vary at that same highest level. Variables at lower levels have at least two sources of variability (depending on the total number of levels and the place of the level of interest in the overall hierarchy). For example, in a three-level model, variables at L3 would only vary at L3, variables at L2 would be allowed to vary at L2 and L3, and variables at L1 would be allowed to vary at L1, L2, and L3. It should be noted that although variables may vary across multiple levels, this does not mean that the variable is measuring the same construct (knowledge, skill, ability, opinion, attitude, etc.) at each level. Assessing a student variable such as grade point average (GPA) at L1 would measure a different construct from the L2 version of the variable which measures the average GPA for a given school. In other words, the measure of a single individual's GPA has different meaning and implications from an average taken across a group of

students, and we cannot interpret the results of both variables as measuring the same construct. Determining the sources of variability for each variable is necessary to inform our understanding of which variables can share a relationship (i.e. correlation). The level(s) at which a variable varies determines other variables to which it could possibly be correlated. MLM allows for within-level correlations as well as across-level correlations to other variables (e.g., cross-level interactions). Additionally, the sign or magnitude of associations may differ across levels. In other words, a correlation between two variables at L1 may be quite different from the correlation between the aggregate versions of these same variables at L2 or higher. Failing to recognize this property can lead to an *ecological fallacy* which occurs when inferences are made about individual units using group means or aggregate values.

Common multilevel models. MLM procedures have been developed to explore a wide range of research questions. Common practice is to use model building procedures of starting with a simple model and incrementally adding complexity. The end result is a model that reflects the characteristics of the data and provides the best model fit (as indicated by model fit statistics, convergence, distributional characteristics, interpretability, and other indicators of interest). Several common MLMs will be discussed.

Models without predictors. A simple MLM with no predictors can be used to capture individual differences and mean differences between groups[3]. If we let Y_{ij}

[3] This type of model is often called an *unconditional model*.

represent our outcome of interest (y) across individuals (i) and clusters/groups (J[4]), then a general model of this type would be

$$Y_{ij} = \beta_{0j} + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

in which β_{0j} is the intercept capturing the mean for each cluster and r_{ij} is the L1 residual which captures the individual differences for each unit (i.e., variability from the mean). In standard regression, we might use dummy coding to model groups, but using dummy codes for MLM would require a very large number of parameters to be modeled, especially as the number of clusters increases. With hierarchical data, we often have a very large number of clusters making dummy coding an unwieldy solution. Using an MLM allows the researcher to analyze data with complex clustering of individuals without overcomplicating the model.

Implementing a random effects model in MLM treats β_{0j} in a different manner from standard regression. In MLM, β_{0j} is still conceptualized as the mean outcome for a particular cluster j, but this intercept is modeled as the average across these means for each cluster (γ_0) and the variation or residual differences in these cluster means (u_{0_j}). For example, supposed we are modeling student GPA using a MLM in which students (i) are groups within schools (j). A random effects model for this scenario would model β_{0j} in Equation 2.1 as

$$\beta_{0j} = \gamma_0 + u_{0_j}, \qquad u_{0j} \sim N(0, \tau_{00}) \tag{2.2}$$

[4] J describes a set of individual groups, j.

such that the model of individual and group mean differences in student GPA (GPA_{ij}), would be

$$\text{L1}: GPA_{ij} = \beta_{0j} + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2) \tag{2.3}$$

$$\text{L2}: \beta_{0j} = \gamma_0 + u_{0_j}, \qquad u_{0j} \sim N(0, \tau_{00}).$$

In other words, the outcome GPA_{ij} is predicted by the combination of the average cluster mean value (L2 intercept; γ_0), the random effect for the L2 intercept (u_{0_j}), and the within-unit (L1, or lower levels) differences (r_{ij}).

For this two-level MLM with no predictors, four groups of parameter estimates are recovered: the L2 intercept (i.e., the average value of all of the cluster means γ_0), the L1 intercepts (β_{0j} s), the L2 random effect of the L2 intercept (u_{0j}), and the L1 residual variance of the L1 intercept (r_{ij}). If there is no variation among L2 clusters, then the unconditional MLM simplifies to a single-level regression model. The amount of variation present at higher levels is determined by calculation the intraclass correlation (ICC). The ICC ranges from 0 (indicating no variance at L2; all observed variance occurs at L1) and 1 (indicating that 100% of the variation in the data is present at L2) and is calculated using the L1 and L2 random effect components. To estimate the ICC, the L2 random effect u_{0j} is divided by the total of both the L1 random effect r_{ij} which captures the deviation of an individual's value from her average value and L2 random effect u_{0j} which captures the deviation across groups from the L2 average, as in

$$ICC = \frac{u_{0j}}{u_{0j} + r_{ij}}. \tag{2.4}$$

The ICC value also identifies the degree to which the assumption of independent samples has been violated. When cluster-level variation is present, data are no longer independent, and the effective sample size is reduced. In the presence of cluster-level variation, using the single-level formula will decrease the estimates of standard error leading to an increase in Type I errors as the value of the ICC increases. Using MLM when higher-level variation is present will produce more accurate standard error estimates (Black, Harel, & McCoach, 2011; van Buuren, 2011).

Models with predictors at L2 only. When only L2 variables are used as predictors of an L1 outcome without including any L1 predictors, the cluster means can be used to predict between-cluster differences at L1[5]. In other words, the cluster mean differences are partitioned into known and unknown elements. Using the student example, a sample MLM of this type might predict student GPA using the school level average of student exams scores ($\overline{exam_j}$). We could estimate this model using

$$\text{L1:} GPA_{ij} = \beta_{0j} + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2) \tag{2.5}$$

$$\text{L2:} \beta_{0j} = \gamma_0 + \gamma_1(\overline{exam_j}) + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00})$$

in which GPA is found by the combination of the intercept at L2 [expected value when the value of the predictor (e.g., *exam*) is equal to zero; γ_0], the slope for the mean values in each cluster of the predictor $\overline{exam_j}$ (γ_1), the L2 random effect (u_{0j}), and the random effect at L1 (r_{ij}). Fixed and random effect interpretations can be made to describe the means and deviations for the L2 intercept and slope. We would interpret these parameter

[5] This model is often called a *means-as-outcomes* model.

values as we would for single-level regression. In our example, β_{0j} is interpreted as the average GPA for students in a particular school *j*, the intercept at L2 (γ_0) would be expected average GPA for a school where $\overline{exam_j} = 0$, and the slope (γ_1) would describe the change in a school's average GPA given a one-unit increase in school average exam score. The L2 random effect (u_{0j}) describes the deviation across schools from the average student GPA after accounting for school average exam score. The L1 random effect (r_{ij}) is interpreted as the unique deviation in GPA values for person *i* in school *j* from his own average value.

In addition to these fixed and random effects, we can also describe the amount of variance explained by the model by calculating R^2 at each level of the hierarchy. One recommended R^2 calculation is a pseudo-R^2 which calculates the proportion reduction in error at L2 (higher levels) as

$$R^2_{L2} = \frac{u_{0j(baseline)} - u_{0j(full)}}{u_{0j(baseline)}} = 1 - \frac{u_{0j(full)}}{u_{0j(baseline)}}. \quad (2.6)$$

in which the baseline model is a model without predictors, and the full model is the model containing predictors. In the current scenario, the full model is the model with an L2 predictor found in Equation 2.5, and the baseline model is the model without predictors in Equation 2.3.

Models with L1 predictors. In addition to models which include predictors at higher levels, models may also include predictors at L1. Adding these L1 predictors allows us to examine whether the influence of the predictor is similar or different for each level and whether this influence varies across clusters. For instance, a variable may be a

significant predictor of L1 variance but does not significantly predict at L2, and each cluster could have a unique regression equation with different predictors. Using the ICC, we can determine which levels our predictors of interest vary across. When variation exists across L1 and L2 in a two-level model, this creates two possible sources of correlation/relationship among the variables. We may observe a correlation between two L1 variables and a correlation between the average/cluster mean versions of these same variables at L2.

MLMs which allow for predictors at all levels partition the L1 variance in the data into explained and unexplained elements[6]. Building on the previous model, the current model adds a L1 predictor of exam score allowing for L1 variation in exam (shown by subscript *i*) and for cluster-specific average exam values at L2 such that

$$\text{L1:} GPA_{ij} = \beta_{0j} + \beta_{1j}(exam_{ij}) + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2) \tag{2.7}$$

$$\text{L2:} \beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{exam_j}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad \boldsymbol{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}).$$

The slope β_{1j} is included in the model, and allowed to vary across clusters, but slope variation across clusters is not predicted. Figure 2.1 presents an example in which cluster differences in the intercept are modeled as a function of exam score as shown in the L1 model of Equation 2.7. This figure captures a difference in intercepts (group means) between group *red* and group *blue*, as well as the average of all group means (β_{0j}) for schools with high (Figure 2.1a) and low (Figure 2.1b) average exam scores. Slopes are

[6] This model is one version of a *random intercept model*.

allowed to vary but are not predicted in the model. In this example, the slopes are the same (i.e., constant) across the schools with high and low average exam scores. However, we may also observe slopes which differ across schools. In this example, it is evident that schools with a higher average exam score tend to have higher intercept values (averages) and schools with lower average exam scores have lower intercept values.

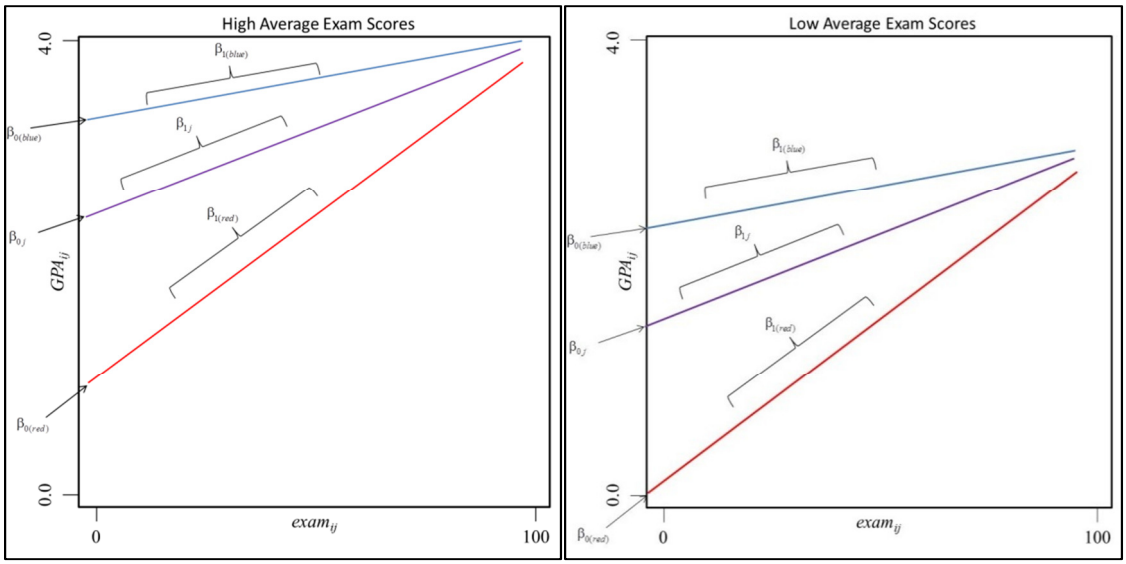


Figure A Figure B

Figure 2.1. Example of MLM with varying intercepts for each cluster *j* and the average intercept and slope averaged across *j*. Intercepts for each cluster *j* are shown in the red and blue lines and the average intercept and slope averaged across *j* is shown in purple. Slopes are allowed to vary but are not being predicted in this model. Figure 2.1a captures intercepts and slopes for three schools with relatively high average exam scores. Figure 2.1b depicts intercepts and slopes for three schools with relatively low average exam scores.

Our interpretation of this model's intercept (β_{0j}) is as the expected value of GPA for students in a particular school *j* with exam scores of 0. The interpretation of the slope (β_{1j}) is the average GPA change for every one-unit change in a student in school *j*'s exam score. The slope at L2 for β_{0j} (γ_1) describes the change in a school's average GPA, for students with an exam score of 0, given a one-unit increase in school average exam score.

When interpreting random effects, the L2 deviation from the average L2 intercept (u_{0j}), the deviation from the L2 average slope (u_{1j}), and the L1 deviation from the individual's average value (r_{ij}) capture the residual variance remaining after accounting for the predictor (e.g. average exam score).

Models with cross-level effects. MLMs can include elements of cross-level variation in which the influence of one or more predictors differs across levels. If the predictor of interest is an L1 predictor with influences shown to differ across L1 and L2, two regression coefficients must be used to capture the influence of that predictor[7]. Due to the differential influence of the predictor, interpretation of the L1 slope changes such that the L1 and L2 effects become partial regression coefficients. This means that the slope elements at L1 and L2 are partial regression coefficients, and the influence of the predictor at L2 is above and beyond the influence already shown at L1. We can test for this type of influence in an L1 predictor by modeling the cluster means as an L2 predictor of the slope. Now, the model given in Equation 2.7 is expanded to include the average exam score in each school as a predictor of the slopes which would be given by

$$\text{L1}: GPA_{ij} = \beta_{0j} + \beta_{1j}(exam_{ij}) + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2) \tag{2.8}$$

$$\text{L2}: \beta_{0j} = \gamma_{00} + \gamma_{01}(\overline{exam_j}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\overline{exam_j}) + u_{1j}, \qquad \boldsymbol{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}).$$

Having these two slope elements allows the L1 and L2 associations to vary and provides us with parameter estimates of the unique influence of L1 variability (γ_{10}) and L2

[7] This is known as a *contextual effect.*

variability above and beyond L1 (γ_{11}). The results now show the separate estimates of the L1 and L2 slopes without any cross-level influence. Random effect interpretations for L2 now describe the deviation from the L2 means (i.e., intercepts; u_{0j}), the deviation from the L2 slopes (u_{1j}), and the L1 residual within-person deviation from an individual's average value/score (r_{ij}) *after* accounting for the L1 predictor. When $\gamma_{11} = 0$, there is no difference in the influence of the predictor across levels and the L1 and L2 regressions are the same. When $\gamma_{11} \neq 0$, the cluster means for exam (i.e., schools' average exam scores) influence the outcome of GPA differently than do the L1 individual student exam scores.

If the L1 and L2 versions of the same predictor exhibit different influence across levels, then we must interpret the predictors separately. For example, if there is a positive relationship between student level exam score ($exam_{ij}$) and the outcome of GPA_{ij} and a negative relationship between school average exam score ($\overline{exam_j}$) and GPA, then γ_{11} would be unequal to zero. In this scenario, observing a cross-level influence like this would mean that as a particular student's exam score increases, the value of her GPA increases as well, but as the school level exam score increases, GPA decreases.

Interaction effects. Another way to conceptualize this differential influence across levels is as a cross-level interaction. When estimating a model with two or more levels, we can investigate interaction effects between variables across levels (cross-level interactions), between pairs of L2 or higher variables (cluster-level interactions), and between pairs of L1 variables (within-cluster interactions). At their core, these interaction effects capture differences in the influence of predictors at a given level or across levels.

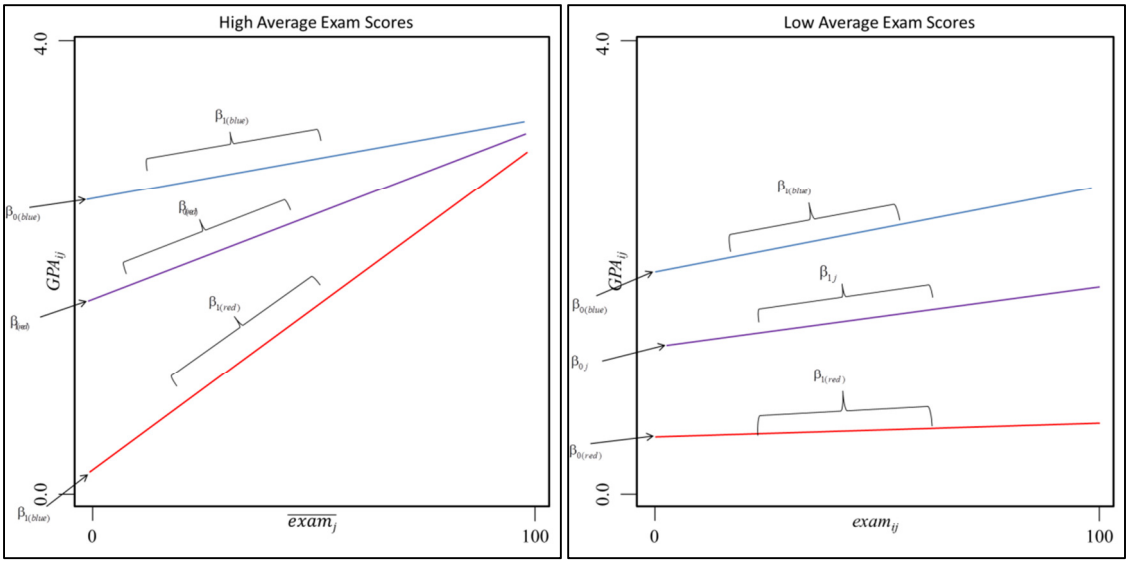


Figure A Figure B

Figure 2.2. Comparison of cluster differences between groups with high average exam scores and low average exam scores

As such, the model shown in 2.8 can also be thought of as an interaction effect of exam across levels[8]. In other words, the values of L1 differ with respect to the values of the cluster average values at L2. Figures 2.2a and 2.2b illustrate a difference in slopes due to an interaction at L2. In this case, we can see that the slopes and intercepts differ by the value of the average exam score in that cluster (high vs. low average exam score). We can see that schools with higher average exam values (i.e., intercepts) in Figure 2.2a also have steeper slopes, and schools with lower average exam scores (i.e., lower values for L2 intercepts) in Figure 2.2.b have flatter slopes. Cross-level interactions can occur when L2 averages of L1 predictors or when predictors which are solely at L2 (e.g., school size) are added to the MLM.

[8] Some researchers make the distinction between cross-level interactions and contextual effects, but the principle is the same: a predictor at L2 influences the L1 regression.

Additional Complexity. The models discussed here can be extended to model further complexity found in a given dataset. When data contain additional predictors at any level, additional parameters can be added to include those predictors and their related effects (e.g., error variances, interactions, etc.). For data in hierarchies with more than two levels, additional levels can be added to the MLM to model the higher levels (e.g., three-level models).

Longitudinal data. Another form of MLM which allows researchers to model variables that were measured at multiple time points is longitudinal growth modeling (LGM). LGM provides a framework for modeling change or growth in individuals by estimating a regression analysis across multiple time points or multiple measurements. Longitudinal MLM (L-MLM[9]) is designed to model changes or growth at each level of a multilevel analysis across time points or measurement instances. In the modeling and prediction of student success, L-MLM can be used to model changes in student success at different points in a student's education and provides researchers with the ability to model change in individuals' (or the unit of analysis) growth rates (i.e. slopes) instead of at some larger group level of average growth or change. This form of hierarchical data contain individual data points (i.e., L1 data) which exist in defined groups (L2 or higher). L2 groups may be further clustered at higher levels. Additionally, longitudinal data may be clustered with time acting as the clustering factor and individual data points being clustered by data collection times.

Categorical outcomes. Many traditional regression analyses model continuous outcomes, but, like single-level regression, procedures have been developed to handle

[9]Also called hierarchical/multilevel growth curve analysis.

outcomes that are binary or categorical. Because categorical outcomes do not follow the assumption of being normally distributed, we must use other procedures when outcomes are non-continuous. With categorical data, we become interested in describing the probability of being in a certain category. Logistic and probit regression are two common variants of regression that are used with categorical outcomes (Hedeker, 2008; Powers, 2012; Serban, Staicu, & Carroll, 2013). Both models use link functions to transform these probabilities of being in a certain category into a form of a latent variable which can be used with linear regression. Then a mean function can be used to transform the parameters from the regression to back to the probability metric. The use of categorical variables in MLM will be discussed in later sections as well.

Missing Data Analysis

As mentioned, another key factor which may impact analysis results and prevent accurate interpretation of findings is the absence of data or missing data (aka missingness). Modeling procedures rely on assumptions about the data and its representativeness. When data are incomplete in some way, the analysis results and interpretation may not be accurate. Missingness is often observed for a particular item or variable, but may also occur across items or across measures or data collection instances.

Missing data theory and analysis of missing data have gained much interest in recent years (Enders, 2010). This increased focus may be due to the implementation of new techniques to explore the patterns and mechanisms of missing data or *nonresponse* and to recent research which has explored the impact of items which were left blank by test/survey takers or which are not available by some other means (i.e. not reported). Nonresponse might refer to instances in which respondents chose not to answer

individual items (i.e. skipped items), respondents did not provide answers for sets of items or measures (e.g. missing responses for items at the end of a test or survey), individuals were not be administered certain items or sets of items (e.g. computer adaptive testing, studies with planned missing data designs), or data was not be available for certain variables (e.g. missing entries in a database).

Missing Data Patterns

When examining the presence of missing data, it is helpful to examine the patterns of provided responses and missing responses which are called *missing data patterns* (Enders, 2010). These patterns are observational in nature and allow the researcher to describe data that are present or missing for each individual and across groups of individuals with identical missing data patterns. It should be noted that missing data patterns describe whether a response has been provided by individuals but do not typically describe the type or quality of responses (e.g., we observe the presence or absence of a response, not the content of the response).

Several common missing data patterns have been described by Rubin (1976) and others (Baraldi & Enders, 2010; Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002). First, a *univariate pattern* describes the occurrence of missing data on only one variable in the dataset. This type of missing data pattern is not commonly found in practice, but an example of this pattern is a short multiple choice survey with an open-ended item at the end. Due to the increased effort necessary to complete the open-ended item, respondents might tend to answer all of the multiple choice items and choose not to answer the open-ended item which could result in a pattern of missing responses on that item only.

Second, a *unit nonresponse pattern* describes a scenario in which an individual does not provide answers to one or more groups of consecutive items. These groups of items or subtests may have been administered to certain respondents and not presented to other respondents. We might see this kind of missing data pattern with longitudinal surveys which have been designed to collect answers from different groups of respondents across prescribed time points. Another example of unit nonresponse occurs when sets of items on an exam are ordered by difficulty with the most difficult items placed at or near the end of a subtest or the overall exam. Individuals may not provide answers to these item sets if they do not have sufficient time or the required knowledge or ability to provide responses. Related to the unit nonresponse pattern is the *monotone missing data pattern*. This pattern is also common to longitudinal research, and it is observed when individuals drop out of a study and do not return (i.e. attrition). These respondents have data present at early time points, but data are not available for later time points.

One of the most prevalent missing data patterns is the *general missing data pattern* which may also be described as *item nonresponse.* Item nonresponse includes situations in which the respondent has chosen not to provide a response to specific items. Looking across respondents, we would observe missingness that is scattered across items in our dataset in what appear to be random patterns. Item nonresponse may be found in instances in which respondents accidentally skip an item, in situations in which providing an answer could be detrimental to the respondent (e.g. surveys about sensitive or personal topics),or in educational testing scenarios in which respondents may lack the necessary familiarity with, knowledge about, or ability required for the specific subject matter.

Two other types of missing data patterns are found in planned missing data survey designs and in latent variable modeling. For planned missing data designs, items of interest are included in different rounds of surveys and groups of individuals are administered the survey forms at different time points during testing. This type of design creates known patterns of unit nonresponse. Likewise, when latent variables are included in our analyses, we know that individuals' values on the latent variables are universally not observed and are therefore missing for all respondents (Bollen, 2002).

Missing Data Mechanisms

Although it might interest researchers to know the kind(s) of missing data patterns present in a dataset, it is often much more important to investigate the *mechanisms* of missingness (Enders, 2010; Little & Rubin, 2002; Rubin, 1976). We can think of missing data patterns as the "what" of missing data analysis (e.g. knowing what items individual choose to provide a response or skip) and missing data mechanisms as the "why" of missing data analysis (i.e. investigating reasons for the observed missingness). In other words, we must consider whether relationships exist between items with missing values and other variables (e.g. items, demographics, subscale scores, etc.) in the dataset. Three mechanisms of missingness have been commonly referenced in the literature describing whether data are missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR). Mechanisms of missingness are also described as *ignorable* or *nonignorable* missingness. Typically, MCAR is considered to be ignorable missingness, MNAR is nonignorable missingness, and MAR may or may not be ignorable depending on other characteristics in the dataset.

MAR describes instances in which the probability of missingness on a particular variable (*H*) is related to the values of some other variable but is not related to the values of *H* itself. In other words, missingness is observed on *H* due to differences in other variables, or there is no relationship between the values of *H* and missingness on *H* after factoring out the influence of other variables on missingness for *H*. For example, if values of *H* are missing for individuals with high values on another variable *O* then those values are considered to be MAR. It is often impossible to determine whether data are truly MAR because even if we can factor out the missingness due to relationships between missingness on *H* and other variables, we do not have all values of *H* and cannot prove that the missingness is not in some way due to whatever values might have been observed if they were not missing. Many estimation techniques assume that the missing data mechanism is (at least) MAR.

When data are MCAR, we can think of the patterns of missingness as being truly random. In other words, missingness on *H* is not due to any other variables in our dataset and is not a function of values of *H* itself. The assumptions of MCAR are much stronger than those required for MAR data. We can think of data that are MCAR as being a random sample of the values we would observe if we had complete data for the variable(s). With continuous data, we can test whether data are MCAR by splitting the dataset into observed and missing for each variable with missingness and examining tests of mean differences for the other variables in the dataset. For categorical data, we can use a likelihood ratio chi-square test for whether the data are MCAR. If no significant differences are found between the groups with observed and missing data, then we can describe the missingness as MCAR.

When the missingness on the variable is related to values of itself after controlling for other variables in the dataset, we consider data to be MNAR. For our variable *H*, values would be MNAR if individuals with low values drop out of the study or are removed from the dataset in some other way (e.g. GRE scores are used to select interviewees among a pool of applying graduate students). An example of MNAR data would be a medical study in which individuals with terminal neurological diseases undergo periodic testing to examine the density of neurons in certain parts of the brain. Because decreasing density of neurons is a result of the disease progression, very low density values would not be observed in the dataset of living patients; unfortunately, patients with this disease drop out or succumb to the disease before extremely low values can be observed. This missing data mechanism can also be described as *informative missingness* because the missingness itself influences our understanding and interpretability of the data (Goldstein, 2011a).

Missing Data Handling

Traditional missing data handling procedures. Missing data analysis includes procedures for handling missing data and procedures for modeling missing data. Many historical approaches have been developed to handle missing data. *Listwise* or *pairwise deletion* removes the affected case(s) from the dataset or from particular analyses respectively. *Mean imputation* substitutes the variable mean for cases with missing values for the variable. *Regression imputation* replaces missing values with predicted scores from a regression analysis predicting each variable with missingness by using the complete variables as predictors. *Stochastic regression imputation* also estimates regression scores but places the additional constraint of normally distributed residual

values to preserve the variability in the data. *Hot deck imputation* replaces missing values with scores from respondents which are similarly matched on characteristics. *Similar response pattern imputation* substitutes missing values with scores of other respondents with similar patterns of response and nonresponse. *Person mean imputation* averages across a single individual's observed response values and is used primarily in survey research or with rating scales. Finally, *last observation carried forward* repeats the individual's last values at a previous data collection point for later time points if the individual drops out of a longitudinal or repeated-measures study. These historical approaches have been found to negatively influence the results of subsequent analyses by increasing bias, decreasing variability, and restricting the range of the data.

Modern missing data handling procedures. More recent procedures of full-information estimation and multiple imputation have been found to be more effective than the historical approaches. These modern techniques minimize bias in parameter estimates, improve the power of the analyses by including all observed data (not just the complete data), and prevent inaccurate inferences based on datasets in which cases with missingness have been removed (Enders, 2010). Maximum likelihood estimation approaches such as full-information maximum likelihood (FIML) estimation include all data in the analysis—whether cases are missing or complete for a certain variable. This estimation procedure captures the probability of an individual's scores coming from a multivariate normal (MVN) distribution with some vector of means and some covariance matrix. When deletion techniques are used, parameter estimates tend to be inflated, but FIML estimation includes the cases with missing data which act as a correction factor to

produce estimates which are closer to actual values that would be recovered with complete data.

When FIML is used for estimates of complete data, log-likelihood values depend on all variables. If data are missing on one or more variables, then log-likelihood values depend only on the variables with observed values. Estimation procedures are performed on each case, but cases with the same missing data patterns should result in similar log-likelihood values (Enders, 2010). For example, in a dataset with five variables, if data for a particular case are present for all five variables, the log-likelihood values would depend on the values of the mean vector and covariance matrix based on all variables. If data are only present for variables 1, 3, and 5, then the log-likelihood is dependent on a subset of values of the mean vector and covariance matrix related to the three variables with data present. Explicitly, mean and variance/covariance values for variables 2 and 4 are not included in the analysis for this case. For cases with data present for two variables, the log-likelihood captures the probability that an individual's scores belong to a bivariate normal distribution with some mean vector and covariance matrix. When data are present for only one variable in the set, the log-likelihood captures the probability that an individual's score belongs to a univariate normal distribution with a certain mean and variance.

Multiple imputation (MI) procedures (e.g., Finch, 2008; Maier, 2002) include a three-step process of (1) generating multiple datasets of potential replacement values using methods which estimate the means and covariance matrices and predict the missing variables from the complete variables (similar to stochastic regression methods); (2) analyzing each now-complete dataset with the chosen analysis model; and (3) pooling the

results of the analyses across datasets to provide single-value point estimates of the parameters (Rubin, 1987). These are generally labeled as the imputation phase, analysis phase, and pooling phase, respectively.

This three-step process describes all multiple imputation procedures, but the imputation phase can be tailored to the particular data analysis which is of primary interest to the researcher. Specifically, the algorithm for estimating the missing data can be altered for particular uses, and when imputation procedures and analysis procedures are aligned, then the models are said to be *congenial*. The most widely used method for the imputation phase is the data augmentation algorithm which is based on MVN data (Enders, 2010). The imputation phase of the data augmentation algorithm is made up of two sub-routines: the imputation step (I-step) and the posterior step (P-step). The *I-step* uses a form of stochastic regression to predict the variables with missingness from the variables with complete data using estimates of the mean vector and covariance matrix for the data. A general equation for a bivariate imputation formula is:

$$Y_i^* = \left[\hat{\beta}_0 + \hat{\beta}_1 (X_i)\right] + z_i \tag{2.9}$$

in which Y_i^* is the imputed score for an individual on the variable with missing data, X_i is the observed score for the complete variable, and z_i is a normally distributed random residual with a mean of 0 and a variance which is equivalent to the value of the residual for the regression of *Y* on *X*. The addition of this residual value adds variability into the predicted scores for the imputed data, decreasing bias which is introduced by using regression to predict the imputed scores.

Bayesian estimation is a popular computational framework for estimation of the I-step and P-step. In Bayesian estimation, for each I-step, we make random draws from the *full conditional distribution* which is the conditional distribution of the observed data and values of the mean vector and covariance matrix. Due to this procedure, the estimated values for the mean vector are considered to be conditional means (expected values of the incomplete variable conditional on having a certain value of the complete variable). These imputed values are drawn randomly in a given I-step using the full conditional distribution of:

$$Y_i^* \sim p(Y_{mis} \mid Y_{obs}, \theta^*) \tag{2.10}$$

such that Y_i^* [10] are the imputed values at that I-step within an iteration, Y_{mis} are the data points with missingness, Y_{obs} are the observed data points, and θ^* captures the estimated values of the vector of means and the covariance matrix from the previous P-step.

The *P-step* is used to generate other possible predicted values in the I-step which vary randomly around the values of the mean vector and covariance matrix. Using the imputation results of a previous I-step, the P-step adds random residual values to the mean vector and covariance matrix values, which are then used in the next I-step to predict imputed values in the regression procedure. The introduction of estimates which randomly differ at each P-step creates new estimates in the I-step. These new predictions carry forward to the following P-step and the process iterates to create many datasets with estimates of the missing values. The variability among estimates that is introduced via the P-step allows for multiple estimates of the missing values.

[10] Just drawn values are indicated with an "*".

From a Bayesian perspective, we can think of these values for the mean vector ($\hat{\boldsymbol{\mu}}$) and covariance matrix ($\hat{\boldsymbol{\Sigma}}$) as being randomly drawn from their posterior distributions. As described, resulting values from the previous I-step are used to estimate values of $\hat{\boldsymbol{\mu}}$ and the sums of squares and crossproducts ($\hat{\boldsymbol{\Lambda}}$). With the newly estimated values, the full conditional distribution of the covariance matrix [$\Sigma|\hat{\boldsymbol{\mu}}$, $\boldsymbol{Y}$] is found by

$$\Sigma|\hat{\boldsymbol{\mu}}, Y \sim W^{-1}\left(N-1, \hat{\boldsymbol{\Lambda}}\right) \tag{2.11}$$

in which $\hat{\boldsymbol{\mu}}$ is the mean vector drawn from the posterior distribution, Y is the now-complete data matrix from the previous I-step, W^{-1} is an inverse Wishart distribution with the parameters of N-1 degrees of freedom (i.e. mean value) and the matrix of samples sums of squares and crossproducts $\hat{\boldsymbol{\Lambda}}$ which captures the variability of the distribution. The data augmentation algorithm mentioned previously uses this formula and Monte Carlo estimation procedures to draw new values for $\hat{\boldsymbol{\Sigma}}$ from this posterior distribution. Similarly, to create a new set of means ($\hat{\boldsymbol{\mu}}^{*}$), draws are taken from the full conditional distribution of the mean vector $\hat{\boldsymbol{\mu}}^{*}|Y, \Sigma$ as

$$\hat{\boldsymbol{\mu}}^{*}|Y, \Sigma \sim MN\left(\hat{\boldsymbol{\mu}}, N^{-1}\boldsymbol{\Sigma}^{*}\right) \tag{2.12}$$

which is distributed as a MVN distribution with a vector of sample means ($\hat{\boldsymbol{\mu}}$) and the simulated population covariance matrix, $\boldsymbol{\Sigma}^{*}$. After this P-step, the following I-step uses these newly drawn values of the means and covariance matrix in the regression equation predicting the missing values. New I-step values are passed along to the following P-step, and the process iterates. The general formula for the P-step is

$$\boldsymbol{\theta}^{*} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}^{*}\right) \tag{2.13}$$

in which $\boldsymbol{\theta}^{*}$ are estimated the values for $\hat{\boldsymbol{\mu}}^{*}$ and $\boldsymbol{\Sigma}^{*}$ from the previous P-step.

From a large number of iterations of the I-step and P-step, a few datasets are selected as the complete datasets for the analysis and pooling phases. Each dataset is analyzed using the researcher's chosen analysis. So, if *m* imputed datasets are created, then *m* analyses are conducted. Because these imputed datasets are now complete, they can be analyzed using procedures that are suitable for complete datasets. After being analyzed, the *m* analysis estimates are combined using the *multiple imputation point estimate*, $\bar{\theta}$, which is the average across *m* estimates in

$$\bar{\theta} = \frac{1}{m}\sum^{m}\hat{\theta} \tag{2.14}$$

where $\hat{\theta}$ is a parameter estimate (e.g., GPA_{ij}) in a given iteration. Furthermore, the variability of the multiple imputation point estimate can be estimated using by combining the within-imputation variance

$$V_W = \frac{1}{m}\sum^{m} SE \tag{2.15}$$

where SE^2 is the squared standard error for a given iteration with the between imputation variance

$$V_B = \frac{1}{m-1}\sum^{m}(\hat{\theta} - \bar{\theta})^2 \tag{2.16}$$

to estimate the total sampling variance

$$V_T = V_W + V_B + \frac{V_B}{m}. \tag{2.17}$$

This total sampling variance captures the combination of the variability in the complete dataset (i.e., without any missingness) and the additional error due to the presence of missing data. This can also be expressed as the MI standard error

$$SE = \sqrt{V_T} \ . \tag{2.18}$$

Categorical Data Analysis

Another characteristic of data which can affect both modeling and missing data handling procedures is the type of data that are collected. Data tend to fall into two main classifications regarding type: continuous and discrete. Continuous data can be measured in increasingly small increments. For example, we can describe the weight of a given item in increments ranging from kilograms down to fractions of a gram. Discrete data may only take on certain values. Categorical data are one form of discrete data that are often observed in survey or educational research. Data of this nature may only take on specific values and may not have partial values or be described in smaller increments of values. An example of categorical data is the binary categorical variable of gender which takes on one of two values: male or female[-]. Categorical data may be unordered (as in the example of gender), or ordered (e.g., a variable of employment with the categories of unemployed, employed part-time, and employed full-time).

Analyzing categorical data necessitates the use of more specific models designed to handle this type of data. If data are single-level, then a variant of logistic regression may be used to analyze the relationship between predictor variables and the outcome variable (Boyle & Willms, 2001; Patrick, 2001; Powers, 2012). It should be noted that in

some instances even variables which are typically conceived of as being discrete could be conceptualized as and, therefore, measured as continuous (e.g., gender measures as a continuum of masculinity/femininity).

Multilevel Modeling with Missing Data

Modern missing data handling techniques have also been applied to hierarchical and longitudinal modeling frameworks. In one sense, the use of techniques such as MI is even more important when data are clustered. Traditional techniques such as LD do not only remove cases at the initial unit level (i.e., L1), but deletion of higher-level variables with missingness includes the deletion of all lower-level cases. For example, if the value of an L3 variable of school type (e.g. public, private, charter) is missing for a certain school, then all teacher/classroom and student data is removed as well. The results of subsequent analyses may be more greatly impacted than results of similar analyses using complete cases only for single-level datasets.

Given the added complexity of the data and modeling procedures, multilevel datasets containing missingness may require special multiple imputation procedures to handle missing data (Black et al., 2011). Many of the techniques used in single-level data have variations that are suitable for data with two or more levels or data collected across two or more time points. Because both are comprised of clustered data, similar missing data handling techniques may be used for multilevel and longitudinal datasets.

A description of common multilevel missing data handling techniques was detailed by van Buuren (2011). Before selecting a method to handle missing data, the researcher must first determine the impact of the missingness by examining five factors: the presence of specific missingness and the role these variables play in the data/model; the

pattern or patterns of missingness; the mechanism of missingness; the scales used for variables with missingness; and the framework or design that the study follows (e.g. longitudinal). Regarding the role of the variable, the treatment for missing data may differ if missingness is present on the L1 outcome, the L1 predictors, the L2 predictors, group identifier variable(s), or a combination of these types of variables. Although much work has focused on methods for handling missingness on the L1 outcome and L1 predictors, little work has been done to explore methods for missingness on the L2 predictors and the grouping variable(s).

Considering missing data patterns present in the data requires the researcher to determine whether the pattern(s) are monotone/nonmonotone and univariate/multivariate. Monotone patterns often occur in data that are longitudinal and are observed when once missing values are reported for an individual, no further data values are collected. Referring to specifically longitudinal designs, this pattern is often called *drop out*. *Intermittent* (i.e. nonmonotone) missing data are observed when only certain variables are not observed or when variables at certain time points are not observed but variables administered at a later time point or at a later point in the test/survey are observed. The existence of intermittent missingness is typically an example of multivariate missingness. Figure 2.3 (van Buuren, 2011) shows the patterns of univariate/multivariate and monotone/nonmonotone missing data patters.

As with missing data theory in general, we can describe the mechanism of missingness. The common mechanisms are similarly MCAR, MAR, and MNAR. Most approaches to handling missing data and analysis models require the data are at least MAR. The concept of ignorability is applicable in MLM as it is in single-level data. If

missingness is ignorable, observed data procedures such as ML may be used. If missingness is determined to be nonignorable, we must take special steps when applying missing data handling techniques and with subsequent analyses. Similar to single-level data, we make assumptions about the mechanism of missingness present in our data, and if these assumptions are incorrect, they can affect the application of missing data handling techniques and the accuracy of results from analysis models.

The scale of the variables with missingness will determine the method of missing data handling that is most appropriate. Multilevel data may be continuous or discrete/categorical (examples include: ordered or unordered categories, binary/dichotomous, and a mixture of categorical and continuous). Additional formats include count data, data that have been censored or truncated, data that are dependent on previous items/variables, and data that are related to other variables (e.g. summed scores or means). Furthermore, the distribution of each variable may follow normal, bimodal, skewed, or kurtotic distributions. For MLM, many procedures assume that data are MVN. Violation of this assumption may impact the accuracy of some methods for handling missing data.

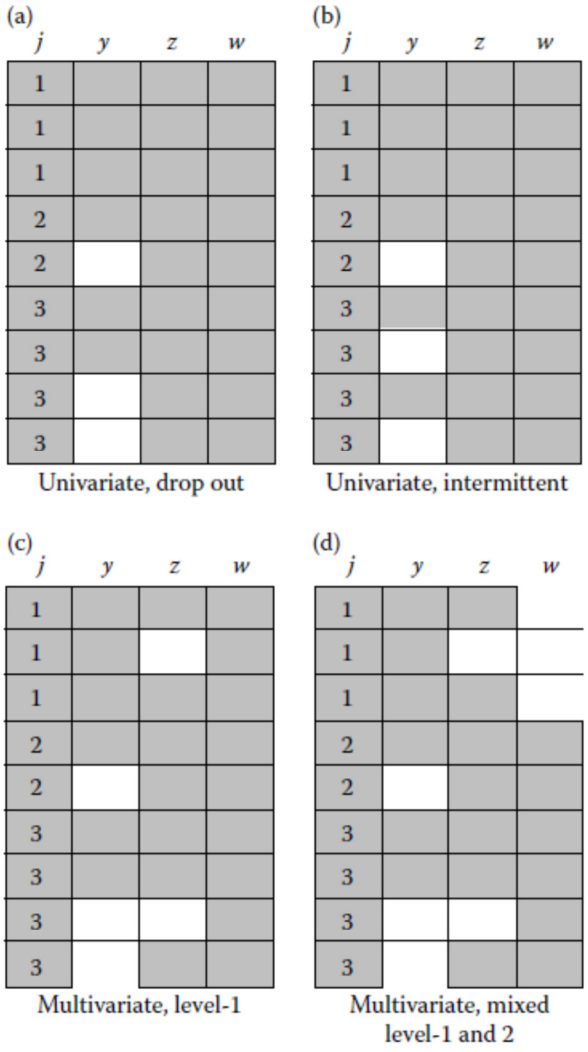


Figure 2.3. Four typical missing data patterns in the multilevel data with two levels and three groups[11]. The grey parts represent observed data, whereas the transparent cells indicate the missing data.

Van Buuren (2011) also discussed the prevalence of traditional methods of missing data handling in MLM such as listwise deletion, last observation carried forward (specific to longitudinal designs), class mean imputation (similar to mean imputation), as well as, more modern ML methods and multilevel multiple imputation (MLMI)

[11] From van Buuren (2011, p.179); ***Z*** captures the collection of L1 predictors; ***W*** captures the collection of L2 predictors.

procedures. When data are complete, MLM can be conducted using the methods described previously or a linear mixed model variation and Bayesian estimation via Gibbs sampling. If missingness is present in the L1 outcome, we can use y_{obs} to represent the observed data and y_{mis} to capture the missing data such that $y = [y_{obs}, y_{mis}]$. Common practice is to estimate the chosen MLM using only y_{obs} if missing data are assumed to be MAR or MCAR (essentially listwise deletion). Then, imputations are made for the missing data by sampling y_{mis} from $y_{mis} \mid y_{obs}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\tau_B}, \sigma^2$ [12] in which $\boldsymbol{\beta}$ is the vector of regression coefficients, $\boldsymbol{u}$ contain the collection of L2 residuals (e.g., u_{0_j}), $\boldsymbol{\tau_B}$ is the variance of the distribution of $\boldsymbol{u}$, and σ^2 is the variance of the L1 residuals.

We can then estimate the imputations using Gibbs sampling to draw values of the standard error for each cluster

$$r_{ij}^{*} \sim N\left(0, \sigma^2\right) \tag{2.19}$$

and of the outcome

$$\boldsymbol{y}_{ij}^{*} = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{W}_j\boldsymbol{u}_j + r_{ij}^{*} \tag{2.20}$$

using the matrix form of the linear multilevel model (Enders & Keller, in press; Enders, Keller, & Mistler, in press; van Buuren, 2011). In this matrix version of the two-level MLM, $\boldsymbol{y}_j^{*}$ captures the vector of outcome scores for each cluster j, $\boldsymbol{X}_j$ represents the collection of L1 predictor variables, $\boldsymbol{W}_j$ is a collection of L2 predictors which have effect on the outcome (i.e., those which are allowed to vary across levels; e.g., random slopes

[12] Notation from van Buuren (2011) has been adapted here for consistency.

for L1 predictor in a random slopes model), $\boldsymbol{u}_j$ contain the random effects[13] for each cluster *j* (e.g., u_{0_j}, u_{1_j}), and r_{ij}^* is the vector of L1 residuals for cluster *j*. Although accurate when imputing values of the outcome variable, this method is less accurate when imputing values for predictors with missingness. Instead, the prescribed method for imputing predictors allows σ_j^2 to vary for each cluster.

Much work has been done to develop procedures for handling missing data in two-level MLMs. Joint modeling (JM) is one method for imputing missing multilevel data that has been explored for use with MLMs (Andridge, 2011; Black et al., 2011; Drechsler, 2015; Enders, 2011a; Resche-Rigon, White, Bartlett, Peters, & Thompson, 2013; Shin & Raudenbush, 2007, 2011; Yucel, 2008;). For JM, individual cases are grouped by missing data pattern and missing values are imputed for each specific pattern using a joint model for all variables with missingness in that pattern. Another estimation method that has been used to impute missing data for MLMs is fully conditional specification (FCS) which is also called multiple imputation with chained equations or chained equations estimation[14] (Andridge, 2011; Liu, Taylor, & Berlin, 2000; van Buuren, 2011). In contrast with JM, FCS imputes values for each incomplete variable individually. FCS is an iterative procedure that seeks to specify the full multivariate variable distributions using conditional densities. Both JM and FCS are especially useful when missingness is multivariate (i.e. on both L1 predictors and L1outcomes).

[13] The term *random effect* is used here in the tradition of MLM frequentist language to describe the deviation from the average over *J*; under the Bayesian framework "fixed" and "random" effects are, indeed, both random by definition in that a prior of randomly drawn values is used during estimation.

[14] In this work, FCS will be used to label all of the methods involving chained equations or conditional estimation.

Recent work by Enders and Keller (in press) has shown that using FCS is advantageous for MLM with data that are MAR (or MCAR) using custom software. FCS imputes values one variable at a time conditional on the other variables in the sample such that missing values are imputed for one variable, X_1, and those imputed values are used in the next MCMC step to impute values for the next variable, X_2, based on the now-complete X_1. With the exception of Enders (2011b), the focus has been on multiple imputation for multilevel data that are MAR or MCAR. Enders (2011b) detailed the three most common MNAR models (mentioned previously) and applied these models to longitudinal data. Due to the promising initial results using FCS for MLM, the primary focus will be on this procedure and its uses for modeling hierarchical data.

MLM with Categorical data

As mentioned, when data are categorical, modeling procedures must be adjusted to compensate for the non-normality of the data. When data are also multilevel, then we must select an MLM which is appropriate for categorical data. Using a linear model (i.e., standard regression) to model discrete data will lead to inaccurate predictions (Boyle & Willms, 2001).

Two main model types have been used to model categorical data in a multilevel framework. The first set of approaches model the discrete variables using category proportions or percentages via a logit or probit (Goldstein, 2011b; Guo & Zhao, 2000) or by a hazard link function (Teachman, 2011). This is known as the proportional odds model or proportional response model (Goldstein, 2011b; Hedeker & Mermelstein, 2011) and can be used for binary variables, categorical (ordered or unordered), and count variables. The second set of approaches views the discrete variables as indicators of

normally distributed latent variables and models them as such (Goldstein, 2011b; Teachman, 2011; Serban et al., 2013) and can be used with the same types of categorical data. If data are binary, then a single threshold is applied to divide the continuous latent distribution, but if the variable contains multiple categories, thresholds are applied to divide the continuous latent distribution into the same number of sections as there are categories (Hedeker, 2008). Much research has focused on estimating MLMs with binary outcome variables (Boyle & Willms, 2001; Patrick, 2001; Powers, 2012). For binary variables, missing values are replaced with randomly drawn values from a normal distribution (Goldstein, Bonnet, & Rocher, 2007).When data are categorical, approaches for unordered or ordered categories can be implemented. These categorical approaches expand each variable with *p* categories into *p* number of variables. Similar to dummy coding, values of 0 or 1 are assigned to reflect the category chosen with a value of 1 and all other variables are assigned a value of 0. A latent Gaussian variable underlies the categorical variable and follows a MVN distribution. A probit probability function is used to calculate the probability that a value is observed in a given category. When missing data are present in L1 responses, data are drawn from the Gaussian distributions described to obtain a complete dataset of multivariate Gaussian responses.

MLM with Missing Data and Categorical Data

Recent work by Enders et al. (in press) developed software specialized software for FCS imputation for multilevel categorical and continuous variables with missingness. To compare the accuracy of various imputation procedures, single-level methods, models with random intercepts and random slopes, and multilevel forms of JM and FCS methods were examined. With the goal of demonstrating the effectiveness of newly developed

MLMI software, a series of simulation studies examined MI for multilevel data with missingness present on both normally distributed continuous variables and categorical variables in random intercept and random slope models. All simulations used two-level MLMs which mimicked diary data with L1 observations being nesting within clusters which were individual respondents. Data were generated in Mplus 7 using either the random intercept or random slope model, and a custom SAS program and was developed for FCS imputation and via C++ code. FCS can be formulated for covariance structures which are either the same or differ across clusters (i.e., individuals). Both of these structure types were used with continuous and categorical missing data. Additionally, the researchers tested another formulation of FCS using latent variables to model categorical variables.

With missing data present on the L1 predictor (x) only, the distribution for the random intercept model for FCS is

$$x_{ij(mis)} \sim N\left(\sum_{j=1}^{J}\beta_j d_j + \beta_{J+1} z_{ij} + \beta_{J+2} y_{ij}, \sigma_e^2\right) \tag{2.21}$$

in which J is again the number of L2 groups/clusters, d_j is the dummy code representing a particular cluster j, $\boldsymbol{\beta}_j$ are the collected of intercepts for the J clusters, z_{ij} is another predictor in the model, $\boldsymbol{\beta}_{J+1}$ represents the influence of z, $\boldsymbol{\beta}_{J+2}$ represents the influence of y, and σ_e^2 captures the residual variance of the regression of x on the complete variables (i.e., z and y).When missingness is present on both the L1 predictor and the L1 outcome (x and y), then the assumptions for the distributions differ given the chosen method. With missingness on both x and y, FCS models x and y separately as

$$x_{ij(mis)} \sim N\left(\beta_{0(x)} + \beta_{1(x)} y_{ij} + \beta_{2(x)} z_{ij} + u_{0j(x)}, \sigma^2_{e_j(x)}\right) \tag{2.22}$$

in which y_{ij} [15] is a filled in value from the previous iteration and the remaining parameters ($\boldsymbol{\beta}_{(x)}$, z_{ij}, $u_{0j(x)}$, and $\sigma^2_{e_j(x)}$) have already been drawn in the current iteration, and

$$y_{ij(mis)} \sim N\left(\beta_{0(y)} + \beta_{1(y)} x_{ij} + \beta_{2(y)} z_{ij} + u_{0j(y)}, \sigma^2_{e_j(y)}\right) \tag{2.23}$$

in which x_{ij} and (the remaining parameters) have already been drawn in the current iteration. Although the distributions for *x* and *y* differ when L1 units do not have the same variance values, when a common variance is modeled and data are single-level, JM and FCS produce equivalent results.

With the random slope model and missingness on both *x* and *y*, FCS assumes

$$x_{ij(mis)} \sim N\left(\beta_{0(x)} + \beta_{1(x)} y_{ij} + \beta_{2(x)} z_{ij} + u_{0j(x)} + u_{1j(x)} y_{ij}, \sigma^2_{e_j(x)}\right) \tag{2.24}$$

and

$$y_{ij(mis)} \sim N\left(\beta_{0(y)} + \beta_{1(y)} x_{ij} + \beta_{2(y)} z_{ij} + u_{0j(y)} + u_{1j(y)} x_{ij}, \sigma^2_{e_j(y)}\right). \tag{2.25}$$

The FCS procedure allows slopes to vary randomly by using these two equations, and the JM formula does not. Results from the simulation study showed that JM and FCS assuming common variances produced negatively biased estimates of the slope variance and residual variances. When FCS was modeled using different variance values across L1 clusters, returned parameter estimates and CI coverage that were accurate (compared to complete data). This finding emphasizes the importance of modeling random slope

[15] If this is the first iteration, then the value for y_{ij} will not come from a previous iteration but from a user-provided set of start values.

variances when conducting MI procedures, and highlights the advantages of the FCS procedure. When slopes varied in the data but were not modeled, bias in estimates of the slope variance was extreme and CI coverage was near zero. Results of FCS with common variance values also returned negatively biased estimates of the slope variance, and JM (which does not model random slopes) for L1 residual variances was slightly positively biased in some instances. When incomplete variables are categorical in nature, researchers typically treat categorical data as if they were normally distributed and continuous (i.e. rounding values to nearest category), use FCS and a general linear mixed model, or use a form of JM which considers categorical variables to be indicators of latent continuous variables.

Work by Enders & Keller (in press) further explored imputation for categorical variables modeled as normally distributed latent continuous variables using FCS. Modeling categorical variables in this way permitted the use of a probit regression model to impute a continuous variable. After the continuous variables were imputed, the categorical imputed values were determined by applying a set of threshold parameters (τ) such that imputed values above (or below) a given threshold were assigned the related category value. This process was applied to both binary and ordered categorical variables.

For nominal (unordered) categorical variables, a multinomial regression model used the relative magnitude of the latent variable scores to determine the category values. For example, when imputing a four-category ($C = 4$) variable x in this way, a value in the third category would show latent propensity scores such that the value for the third category x_3 would be greater than the values for x_1 x_2, or x_4.Using MCMC via FCS, imputations are drawn for the sample at L1, these drawn values are used to sample

residual and parameter values at higher levels, and these values are carried forward to the next iteration, and discrete values are determined from the latent scores (via thresholds for ordered categorical variables or values of the latent propensity scores for nominal/unordered categorical variables).

For an ordered categorical variable with $C = 4$ categories, we can link the categorical and latent scores for this variable *x* via

$$x = c, \text{if } \tau_{c-1} < x^* \leq \tau_c . \tag{2.26}$$

in which $\tau_0 = -\infty$, $\tau_C = \infty$, and x^* is the underlying normal variable for *x*. This means that the discrete variable *x* is assigned a certain category value *c* if the underlying normal variable x^* is above a given threshold and below the next threshold. For the example variable *x*, there would be three possible threshold parameters by which to partition the underlying continuous normal variable. Conversely, a set of C-1 latent difference scores can be calculated using the multinomial probit model to define the normally distributed variable z^* for each response option. For our sample variable *x* with four response categories, the set of latent difference scores would be calculated as

$$\begin{aligned} x_1^* &= z_1^* - z_4^* \\ x_2^* &= z_2^* - z_4^* \\ x_3^* &= z_3^* - z_4^* \end{aligned} \tag{2.27}$$

in which $c = 4$ is the reference category.

Enders and Keller (in press) explored the recovery of accurate imputations via simulation using FCS for varying numbers of clusters, cluster sizes, ICC values, and MAR missingness rates. The model of interest was a two-level MLM which included contextual effects and random slope variation. Variables were continuous, binary, or 6-

category ordinal, and all were permitted to have missingness. Categorical variables were imputed as normally distributed latent variables and categorical values were assigned using threshold cutoffs as described. Imputation was conducted using BLImP software developed by the researchers, and analysis and pooling phases were completed in Mplus 7.

Measures of bias in the estimates showed that FCS produced minimal bias (less than ± .10) in recovered estimates for the regression model. Some bias was observed for estimates of the slope variance with greater biases observed for small cluster sizes and high (25%) rates of missingness. Additionally, bias was observed for the condition with 200 clusters and a low ICC of .20. Reasons for this observation were not fully understood, and the researchers hesitated to make strong statements about the implications. Further simulations are needed to confirm that bias increases with a large number of clusters and moderate to low ICC values.

Coverage rates for conditions with missingness were comparable to conditions with complete data. Trends in coverage rates for the slope variation showed a decrease in coverage as the rate of missingness increased. The number of clusters did not influence coverage rates, and, again, as missingness rates increased the coverage for the slope variance decreased.

Chapter 3

METHODOLOGY

Previous research has focused on simulation studies to demonstrate general procedures for MLMI via JM and FCS. Due to the promising results of Enders et al. (in press) and Enders and Keller (in press), FCS should be further pursued as a method for imputing MAR missing data under a MLM framework. These efforts have examined the effects of the amount of missingness, ICC values, and both continuous and categorical variables. In addition, working from the traditionally frequentist MLMI framework, this work developed custom BLImP software using Bayesian estimation to perform the FCS procedure on a two-level model with categorical and continuous variables.

Due to the relative novelty of applying FCS procedures to two-level MLMs with categorical variables and MAR missingness, further work is needed to determine the separate and combined influences of these data characteristics and modeling choices on the recovery of accurate imputed datasets needed to support correct inference. The current work seeks to explore the impact of data characteristics, missingness, and variable type (i.e., categorical or continuous) on the recovery of accurate imputed datasets using BLImP software for FCS and a more formally Bayesian estimation procedure via the JAGS software. Although both the BLImP software and JAGS make use of the Gibbs sampling routine, comparing results of these two approaches will illuminate whether interpreting Bayesian results within a frequentist MLMI framework or a fully Bayesian framework would lead to different conclusions about the analysis model or the data being modeled.

A series of simulation studies examined the impact of estimation software, hierarchical data characteristics, missing data, and measured variables which are continuous or categorical on the estimation of a two-level MLM. Data for this study were generated from a two-level MLM with random intercepts and random slopes. Values were deleted based on a MAR missing data mechanism using the chosen percentages of missingness. Multilevel multiple imputation of missing values at both levels was conducted by : (1) using the fully conditional specification (FCS) procedure described in Enders & Keller (in press) and (2) modeling the FCS model using fully Bayesian estimation and the Gibbs sampling routine via JAGS. With the focus of this work on comparison between frameworks, manipulated variables common in previous research were chosen.

To test the recovery of the random slope relationship, analysis models with random slopes were implemented. The model contains two L1 independent variables and three L2 independent variables.

Modeling Decisions

MLMs for educational research with complete data are often complex with multiple variables at each level of analysis (Cosgrove & Cunningham, 2006; Du, 2009). However, software for MI with multilevel data may not be able to accommodate such complex models. Simulations studies using FCS have focused on one to three L1 predictors and zero to two L2 predictors (Enders & Keller, in press; Zhao & Yucel, 2009). The MLM for the current work contains two L1 predictors and three L2 predictors (two of which capture the cluster means of the L1 predictors and one L2 only predictor).

Data Generation

In each replication in a given condition, a dataset with *I* L1 data points (e.g., students) and *J* L2 clusters (e.g., schools) were generated using a normal distribution [N(0,1)] for each normally distributed variable. Values for categorical variables were assigned based on the proportions of responses for each category. Missing data were implemented using a variable-by-variable strategy (such that values could be missing according to one of the proposed missingness scenarios) and the chosen percentage of missing data. Each condition contained 500 replications.

Manipulated Variables

Modeling Frameworks and Software

Multiple imputation and analysis of the MLM of interest was conducted using two analysis frameworks: multiple imputation using FCS and frequentist analysis of the MLM of interest and a comparable fully-Bayesian model for imputation and analysis in R/JAGS. Multiple imputation using FCS was conducted using BLImP (Enders & Keller, in press) software for which was developed to impute the missing data via FCS (See Appendix A for sample continuous BLImP model code and Appendix B for sample dichotomous BLImP model code). Subsequent analysis of the MLM of interest was conducted in Mplus (see Appendices C and D for sample Mplus code for continuous and categorical models respectively), and results were pooled across analyses using R. For the fully-Bayesian estimation routine, JAGS (via R) was used (see Appendix E for a sample continuous JAGS model file and Appendix F for a sample dichotomous JAGS model file).

In order to be best able to compare across FCS and fully-Bayesian estimation via Gibbs sampling, prior distributions in JAGS were selected to reflect the common selections made in the MI literature: flat/diffuse/non-informative priors (Enders, 2010; Enders & Keller, in press; Gelman et al., 2013; Rubin, 1987)[16]. This means that the prior for each slope coefficient (β) was a very diffuse normal distributions [N(0, 1000)], the prior for each error variance (σ^2) followed a diffuse gamma distribution [G(1,1)], and the prior for the entire covariance matrix (Σ) was drawn from an inverse Wishart distribution with dimension which matched the given condition. These values were chosen to most closely approximate the choices in the MI literature for FCS imputation.

Dataset Characteristics

Given that the current work is couched in the field of educational research, characteristics of the simulated datasets are based upon real-world examples of MLM analyses that have been conducted. This is especially relevant for decisions regarding sample/cluster size, ICCs, and type(s) of variables being modeled. The multilevel scenario in educational research of students nested within classrooms or schools was selected as a motivating example.

Students within classrooms/schools. Several studies in educational research have focused on students nested by the classroom/teacher to which they are assigned and/or by the school which they attend (Black et al., 2011; Cosgrove & Cunningham, 2006; Du, 2009; Frempong, Reddy, & Kanjee, 2011; Goldstein et al., 2007; Guo & Zhao, 2000; Patrick, 2000; Muñoz & Chang, 2008). Research on multilevel educational data

[16] Some instances may arise in the presence of categorical data in which JAGS may not be able to model prior distributions that are exactly identical to those modeled in BLImP software. In these cases, the closest approximation to the prior used in the MI literature and the BLImP software will be used.

with missingness, categorical variables or both has used L1 sample sizes ranging from approximately 1,500 to 34,000 students with a median value of 8,186 students. This survey of previous research included both grant-funded classroom-based studies and complex international/national educational surveys. These students were grouped into 20 to 1,800 classrooms or schools with between 4 and 134 students per cluster/group. When classroom was the L2 grouping variable, an average (median) of 20 students were grouped into each classroom, and with school as the grouping variable, each cluster contained an average (median) of 28 students. Based on these values, the current study explored the measurement scenario of a cluster size of 25 students nested within either 25 or 50 groups. The total number of students at L1 was 625 and 1,250 respectively. This value of 25 students was chosen to be representative of a typical classroom size, and setting the number of clusters to 25 or 50 reflects the typical number of classrooms/schools commonly found in grant-funded, classroom-based studies.

It should be noted that previous applied research tended to focus on an additional layer of nesting at level-three, but the current work is solely concerned with two-level MLMs. Examining scenarios in which the number of clusters is manipulated and the cluster size is small to moderate—but constant—provides an opportunity to measure the impact of the number of clusters (with values of 25 and 50) on the recovery of accurate parameter estimates after MLMI and analysis.

Interclass correlations. Much of the applied educational research using MLM did not report recovered ICC values. However, several of the simulation studies did report values for ICCs. These values ranged from .00 to .7 (Andridge, 2011; Drechsler, 2015; Enders & Keller, in press; van Buuren, 2011; Zhao & Yucel, 2009). To reflect

plausible ICC values, the current work used ICC values of .1 and .3 to represent a weak and moderate effect of cluster-level (e.g., school-level) variance.

Percent missingness. In past simulation studies of MLMs, rates of missingness have ranged from less than 20% to extremes above 75% with some as high as 100% missingness on a given variable[17] (Enders, 2011a; Lui, Taylor, & Belin, 2000; Schafer, 2001; Schafer & Yucel, 2002; Shin & Raudenbush, 2007; Shin & Raudenbush, 2011). More commonly observed rates of missingness explored range from 20% to 60% missingness (Andridge, 2011; Black et al., 2011; Drechsler, 2015; Goldstein, 2011a; Gottschall, West, & Enders, 2012; Kadengye, Ceulemans, & Van den Noortgate, 2013; Resche-Rigon et al., 2013; van Buuren, 2011; Yucel, 2008). The current study imposed MAR missingness rates (MAR) at 20% and 40% levels. Missingness was imposed on each variable individually such that it is possible for an individual to have missing values on any combination of Y_{ij}, V_{3j}, and X_{2j}.

Patterns of missingness. For the two-level MLM of interest, missingness could be present for the elements of the regression equation in Equation 3.1[18]. This model uses X_{ij} to represent the L1 predictors, $\bar{X}_j$ to represent the cluster means (aggregated values)

[17] Although a rate of 100% missingness may be controversial when imputation is used with applied data, these simulation studies may have chosen rates of missingness that were more extreme than would be recommended for imputation procedures to demonstrate the effects of extreme missingness rates on parameter recovery.

[18] JAGS uses the precision metric to capture variability; τ_r is the inverse of σ_j^2.

for the L1 predictors at L2, and V_j [19] to capture a L2 predictor that is not an aggregate of a variable collected at L1.

$$\text{L1:} Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{1j}\right) + \beta_{2j}\left(X_{2j}\right) + r_{ij} \qquad \begin{matrix} r_{ij} \sim N(0, \tau_r) \\ \tau_r \sim Gamma(1,1) \end{matrix} \tag{3.1}$$

$$\text{L2:} \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_{1j} + \gamma_{02}\bar{X}_{2j} + \gamma_{03}V_{3j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\boldsymbol{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}_B)$$

$$\boldsymbol{\tau}_B \sim W\begin{pmatrix} 3 & 0 & 3 \\ 0 & 3 & 0, \\ 0 & 0 & 3 \end{pmatrix} 3$$

For the series of simulations, X_{1j} is fully observed, and missingness will be implemented via MAR mechanism with three covariates of missingness of a_1, a_2, and a_3 which correlate to missingness on Y_{ij}, X_{2j}, and V_{3j} as the covariate of missingness for all variables. The following scenarios were explored with missingness implemented on the following variables/combinations of variables:

1. No missingness on any variables (for all other manipulated variables of interest)
2. Missingness on Y_{ij} (as a limited run of for the condition with *ICC* = .30, *J* = 50, and *MAR* = 20% with to check for adequate performance)

[19] Although standard notation uses W_j to indicate predictors at L2, the current work uses W_j to indicate L2 cluster means of L1 variables, and V_j to represent an L2 predictor that does not have an L1 counterpart. Examples include school or classroom characteristics that are not also defining characteristics of the students.

3. Missingness on X_{2j} (as a limited run of *ICC* = .30, *J* = 50, and *MAR* = 20% to check for adequate performance)
4. Missingness on Y_{ij} and X_{2j} (for all other manipulated variables of interest)
5. Missingness on Y_{ij}, V_{3j}, and X_{2j} (for all other manipulated variables of interest)
6. Missingness on V_{3j}
7. Missingness on V_{3j} and X_{2j}

As a check for potential differences in results based on analysis software, one condition was run with missingness on Y, *ICC* = .10, *J* = 50, and *MAR* = 20% in which data were imputed in R/JAGS and 20 imputed datasets were saved out. Analysis of these imputed dataset was then conducted in Mplus using the same analysis model used for Model 2 in the previous list.

Variable types

In order to compare the impact of the type of variable on the recovery of accurate imputations, both continuous and categorical variables will be used. Many variables of interest in educational research can be measured as either continuous or categorical (e.g., student achievement), but some variables are by nature discrete and can only be measured as such (e.g., gender; see Burstein, 1980 for a fairly comprehensive list of possible variables). The current work included continuous and categorical variables with missingness at both L1 and L2.

Continuous variables. Variables of interest that have been measured continuously at L1, L2, or both levels include scores measures of student achievement (Carpenter & Goldstein, 2005; Cosgrove & Cunningham, 2006; Frempong et al., 2011; Goldstein et al., 2007; Muñoz & Chang, 2008), SES (Black et al., 2011; Du, 2009; Frempong et al., 2011), class size (Frempong et al., 2011), health indicators (Guo &

Zhao, 2000; Shin & Raudenbush, 2011), and teacher effectiveness (Du, 2009). For the current study, continuous variables were drawn from a N(0,1) distribution, and imputations of continuous variables with missingness are used to assess the recovery of imputations for categorical variables.

Categorical variables. A number of previously researched categorical variables include student variables such as gender and ethnicity (common variables across almost all of the studies but not always modeled directly as categorical), eligibility variables (such as qualifying for free/reduced lunch; Carpenter & Goldstein, 2005; Cosgrove & Cunningham, 2006; Du, 2009; Goldstein et al., 2007; Guo & Zhao, 2000; Muñoz & Chang, 2008; Shin & Raudenbush, 2011), educational outcomes for the student (e.g., pass/fail, advance to next grade/repeat current grade; Black et al., 2011; Cosgrove & Cunningham, 2006; Du, 2009; Guo & Zhao, 2000), parental education level (Cosgrove & Cunningham, 2006; Du, 2009), and teacher education (Cosgrove & Cunningham, 2006; Du, 2009). Given the nature of the variables that are commonly found in educational research, the current work focused on binary variables. This means that the L1 outcome, the L1 predictor, and the L2 predictor were modeled as continuous or dichotomous.

Category proportions for discrete variables. The proportion of responses to categories in a categorical variable has been explored in work by Enders et al. (in press), but the impact of differing proportions of responses has not been further discussed in the educational literature for MLMs with categorical data. The current work examined whether having equal or unequal category proportions will influence the accuracy of the imputation procedures. For the current study, binary variables were explored with the category proportions of .50/.50, .70/.30, or .90/.10.

Computational Details

The imputation and analysis procedures for this study were performed using supercomputing clusters available to students at ASU and in partnership with colleagues at UCLA. The analysis using R/JAGS were performed on the ASU Saguaro computing cluster. Analyses using BLImP and Mplus were conducted using the UCLA Hoffman2 computing cluster. Results from BLImP/Mplus and R/JAGS were housed in an SQL database. Further analysis and plotting was conducted using R.

Computational performance. In addition to the manipulated variables, the speed at which analyses can be completed and convergence can be reached was also of interest. Due to long runtime for the dichotomous conditions, computational choices were made in Mplus to limit the number of integration (i.e., quadrature) points used to estimate the MLM of interest to 10 points per dimension. Reps in which convergence issues were present using 10 integration points were rerun using 15 points. If convergence was still not reached, a final attempt was made for a using 20 integration points. Convergence issues were also present in R/Jags, and subsequent analyses of the results were based upon the replications that did converge.

Assessment of Results

Bias, root mean square error (RMSE), 95% confidence interval coverage, and descriptive statistics for the distributions of fixed effects (e.g., β_{0j}) and random effects (e.g., L1 residuals) found in the complete data case are used to assess the accuracy of multiply imputed datasets. Parameter estimates for each replication come from the posterior means in the R/JAGS procedure. In the BLImP/Mplus procedure, parameter

estimates come from the complete data analysis in Mplus for complete data and from the pooled analysis results across imputed datasets for the imputed data.

Bias

Using the average values across replications from the complete data as the true population values, the average bias was calculated across replications within a given condition. Equation 3.2 (adapted from Meng, 2007) gives an example of how bias was calculated for the values of the intercept

$$BIAS_{AVG} = \frac{1}{R}\sum_{r=1}^{R}, \hat{\beta}_{0jr} - \beta_{0jr}, \tag{3.2}$$

such that R is the total number of replications, r represents a given replication, β_{0jr} is the value of the intercept for the complete data in that replication, and $\hat{\beta}_{0jr}$ is the estimate of the intercept for that replication.

Root Mean Square Error

To capture the differences between parameter estimates recovered with complete data and parameter estimates recovered with each of the missingness patterns examined, RMSE was calculated for each parameter condition as

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_r - \beta_r\right)^2}. \tag{3.3}$$

Confidence Interval Coverage

In order to describe the recovery of accurate imputations, coverage of the 95% confidence interval (CI; or credible interval where relevant) will be estimated. First, the 95% confidence intervals will be constructed for each of the focal parameter estimates for each replication. Coverage will then be calculated as the proportion of replications in

which the estimated confidence interval contained the true value (i.e., complete data average value) of the estimate.

Chapter 4

RESULTS

The description of the results for the current study will primarily focus on the comparison between R/JAGS and BLImP/Mplus for the continuous data and the dichotomous data with a .50/.50 split. During the analysis process, several issues were encountered using both analysis routines when attempting to estimate the dichotomous data with the .70/.30 and .90/.10 splits. The specifics of these convergence issues for the more extreme dichotomous splits will be in the next section. For the continuous data, all conditions converged in both sets of analysis software[20]. However, some convergence issues were encountered for the dichotomous .50/.50 split data. Table 4.1 (in the section on Convergence Issues) shows in the number of replications that converged (i.e., 500 successful replications) in each condition. In BLImP/Mplus, convergence issues were initially observed for some replications when using 10 integration points (i.e., quadrature points) during the analysis phase. Subsequently, these replications were rerun using a larger number of integration points (15 or 20 integration points). After increasing the number of integration points in Mplus, all conditions for the .50/.50 split data converged. Performance of each of the analysis procedures was assessed using confidence interval (CI) coverage, average bias ($BIAS_{AVG}$), and RMSE for each condition by comparing the parameter values estimated with complete data to those estimated from imputed data. This chapter will describe convergence issues and detail which conditions did and did not run successfully in each analysis framework, explain the continuous data results, describe

[20] Due to the nature of the process for imposing missingness, five reps from the V3 missingness data pattern did not have any missing data. Calculations used to assess results in these cases adjusted the total number of reps to reflect this lower number. These conditions are included in Table 4.1.

the comparison of the continuous data results to the dichotomous .50/.50 data, and highlight a special case in which imputations from R/JAGS were analyzed in Mplus for two specific conditions.

The primary method for display results will be via plots with some tables used to supplement when necessary. Except where noted, results will be presented in two groups: (a) fixed effects (e.g., γ_{00}, γ_{01}, γ_{02}, and γ_{03}; γ_{10} and γ_{20}) which are values associated with the L2 intercept β_{0j} and the L2 random slopes β_{1j} and β_{2j} and (b) variances for the random effects (e.g., $\sigma_r^2, \sigma_{u_0}^2, \sigma_{u_1}^2, \sigma_{u_2}^2$ labeled as *R*, *U0*, *U1*, and *U2* respectively, and covariances between L2 random effects, labeled as *U0U1*, *U0U2*, and *U0U3*) which capture the deviation from the related fixed effect estimate. Dividing results into these groups will aid in clarity and interpretation.

Convergence Issues

As mentioned previously, convergence issues were encountered for the dichotomous data conditions. Table 4.1 shows the number of replications that converged for each condition. Successful initial convergence during model testing for R/JAGS was defined as completion of the adaptation phase and burn-in phase in JAGS with 100,000 or fewer total cycles. For example, if a model successfully adapted after 50,000 cycles and was successfully adapted after 50,000 burn-in cycles, then further cycles were taken for the analysis estimation routine. Convergences was tested for a few replications of each group of conditions (given by the combinations of manipulated variables of ICC, *J*, and MAR) to check that the model converged across types conditions/data types. Conditions that converged based on these criteria for these test replications were then analyzed using the ASU cluster with complete sets of 500 replications for each condition.

In some instances, conditions which produced converged results during the testing phase produced individual replications which did not converge during the analysis phase due to incomplete model adaptation.

Conditions which were analyzed on the cluster but had individual replications which did not converge are indicated in Table 4.1 by cells with values that are less than 500. Conditions that were not able to run successfully during the model testing phase and were not analyzed on the cluster are marked with "-". As Table 4.1 shows, some individual replications for conditions with .50/.50 split dichotomous data did not converge during estimation with R/JAGS. In these instances, calculations of CI coverage, average bias, and RMSE included only the replications that had converged. With the more complex missing data patterns of YX2 and YX2V3, the testing phase for models in R/JAGS proved to be unsuccessful across the manipulated characteristics of interest. Although R/JAGS did converge for some of the most simple condition of $ICC = .10$, $MAR = 20\%$, and $J = 25$ during testing, a majority of replications failed when attempting to run on the ASU cluster. The issues with JAGS seem to occur within the model adaptation phase in which the number of cycles needed for adaptation was upwards of 100,000. Conditions with .50/.50 split that adapted in 100,000 iterations or fewer were analyzed using R/JAGS on the ASU cluster. For the other missing data patterns, a majority of problems with convergence were observed using R/JAGS with the more extreme splits of .70/.30 and .90/.10.

Table 4.1

Converged replications for R/JAGS and BLImP/Mplus for complete data, V3 missingness, and X2 missingness conditions

Missing Data Pattern	ICC value	Number of Groups	Missingness Rate	R/JAGS Converged Replications				BLImP/Mplus Converged Replications			
				Continuous	.50/.50	.70/.30	.90/.10	Continuous	.50/.50	.70/.30	.90/.10
Complete Data	0.1	25	20%	500	500	500	500	500	500	500	-
	0.1	25	40%	500	500	500	500	500	500	500	-
	0.1	50	20%	500	500	500	500	500	500	500	-
	0.1	50	40%	500	500	500	500	500	500	500	-
	0.3	25	20%	500	500	500	500	500	500	500	-
	0.3	25	40%	500	500	500	500	500	500	500	-
	0.3	50	20%	500	500	500	500	500	500	500	-
	0.3	50	40%	500	500	500	500	500	500	500	-
V3	0.1	25	20%	500	500	-	-	**496**	**496**	-	-
	0.1	25	40%	500	500	-	-	500	500	-	-
	0.1	50	20%	500	498	-	-	500	500	-	-
	0.1	50	40%	500	500	-	-	500	500	-	-
	0.3	25	20%	500	499	-	-	**499**	**499**	-	-
	0.3	25	40%	500	494	-	-	500	500	-	-
	0.3	50	20%	500	500	-	-	500	500	-	-
	0.3	50	40%	500	500	-	-	500	500	-	-

Note. The symbol "-" indicates conditions that were not analyzed due to issues during model testing. **Bold** text indicates conditions in which fewer than 500 replications contained missingness due to the process for imposing missingness. "N/A" indicates conditions that were not planned or attempted.

Table 4.2

Converged replications for R/JAGS and BLImP/Mplus for X2V3 missingness, Y missingness, Yjags, and YX2jags conditions

Missing Data Pattern	ICC value	Number of Groups	Missingness Rate	R/JAGS Converged Replications				BLImP/Mplus Converged Replications			
				Continuous	.50/.50	.70/.30	.90/.10	Continuous	.50/.50	.70/.30	.90/.10
X2V3	0.1	25	20%	500	468	-	-	500	500	500	-
	0.1	25	40%	500	500	-	-	500	500	500	-
	0.1	50	20%	500	500	-	-	500	500	500	-
	0.1	50	40%	500	500	-	-	500	500	500	-
	0.3	25	20%	500	500	-	-	500	500	500	-
	0.3	25	40%	500	499	-	-	500	500	500	-
	0.3	50	20%	500	500	-	-	500	500	500	-
	0.3	50	40%	500	490	-	-	500	500	500	-
X2	0.3	50	20%	500	500	-	-	500	500	460	-
Y	0.3	50	20%	500	500	-	-	500	500	467	-
Yjags	0.3	50	20%	N/A	N/A	N/A	N/A	496	N/A	N/A	N/A
YX2jags	0.1	25	40%	N/A	N/A	N/A	N/A	500	N/A	N/A	N/A

Note. The symbol "-" indicates conditions that were not analyzed due to issues during model testing. **Bold** text indicates conditions in which fewer than 500 replications contained missingness due to the process for imposing missingness. "N/A" indicates conditions that were not planned or attempted.

Table 4.3

Converged replications for R/JAGS and BLImP/Mplus for YX2 missingness and YX2V3 missingness conditions

Missing Data Pattern	ICC value	Number of Groups	Missingness Rate	R/JAGS Converged Replications				BLImP/Mplus Converged Replications			
				Continuous	.50/.50	.70/.30	.90/.10	Continuous	.50/.50	.70/.30	.90/.10
YX2	0.1	25	20%	500	-	-	-	500	500	481	-
	0.1	25	40%	500	-	-	-	500	500	479	-
	0.1	50	20%	500	403	-	-	500	500	488	-
	0.1	50	40%	500	-	-	-	500	500	488	-
	0.3	25	20%	500	497	-	-	500	500	454	-
	0.3	25	40%	500	498	-	-	500	500	459	-
	0.3	50	20%	500	415	-	-	500	500	470	-
	0.3	50	40%	500	427	-	-	500	500	471	-
YX2V3	0.1	25	20%	500	485	-	-	500	500	483	364
	0.1	25	40%	500	-	-	-	500	500	472	217
	0.1	50	20%	500	-	-	-	500	500	486	468
	0.1	50	40%	500	-	-	-	500	500	486	418
	0.3	25	20%	500	-	-	-	500	500	459	332
	0.3	25	40%	500	-	-	-	500	500	449	203
	0.3	50	20%	500	-	-	-	500	500	469	402
	0.3	50	40%	500	-	-	-	500	500	474	412

Note. The symbol "-" indicates conditions that were not analyzed due to issues during model testing.

BLImP/Mplus was able to handle many of the .70/.30 conditions, but was unable to run a majority of the .90/.10 conditions as Mplus was not able to converge to a solution. Additional complexity was added due to the analysis in Mplus in that individual replications for more complex models with dichotomous data took more than an hour to run on the UCLA computing cluster –even after cutting down the number of integration points that Mplus was using to assess convergence. By utilizing the computing power of the UCLA cluster to split analyses into multiple separate runs, BLImP/Mplus analyses for the .50/.50 were successfully completed. It should be noted that for the .50/.50 split conditions, analyses that did not converge with the initial run of 10 integration points were rerun using 15 integration points (or 20 integration points if needed). Numbers of replications for the .50/.50 conditions in Tables 4.1 through 4.3 reflect convergence after attempts with up to 20 integration points. After convergence issues were encountered with the .70/.30 and .90/.10 conditions that were attempted, analysis of the remaining conditions and missing data patterns was not carried out.

Continuous Data

To allow for comparison across statistics used to assess results, each plot in this section contains values for $BIAS_{AVG}$, CI coverage, and RMSE for each continuous condition. Shape, color, and fill are used to depict the different levels of the manipulated variables, as well as which analysis procedure was used. Table 4.4 contains a key describing each symbol, color, and fill combination used. Dashed vertical guidelines are used as reference lines to highlight average bias values between -.1 and .1, coverage rates of .95 and above, and an RMSE value of 0. Each plot captures values from both R/JAGS and BLImP/Mplus (represented by the red/yellow and green/blue color groups

respectively). Columns of plots indicate the type of statistic that was used to assess the results, each row of plots divides results by the variable(s) with missingness in the data/model, and each individual tick mark/row inside of a given plot captures values for a particular parameter.

Table 4.4

Symbol key for interpreting figures of results.

MAR	J	ICC	Software	
			R/JAGS (red or yellow)	BLImP/Mplus (green or blue)
20% (outlined)	25 (red or green)	.1 (upward-facing triangles)	△	△
		.3 (downward-facing triangles)	▽	▽
	50 (yellow or blue)	.1	△	△
		.3	▽	▽
40% (filled)	25	.1	▲	▲
		.3	▼	▼
	50	.1	▲	▲
		.3	▼	▼

Note. This key describes what each of the markers represent in each of the figures in this chapter. Direction of the marker is used to capture ICC value with upward-facing triangles for *ICC* = .1 and downward-facing triangles for *ICC* = .3. Color captures the number of groups with *J* = 25 shown in red for R/JAGS and green for BLImP/Mplus and *J* = 50 is shown by yellow for R/JAGS and blue for BLImP/Mplus. The rate of missingness is shown by whether the marker is outlined (*MAR* = 20%) or filled (*MAR* = 40%).

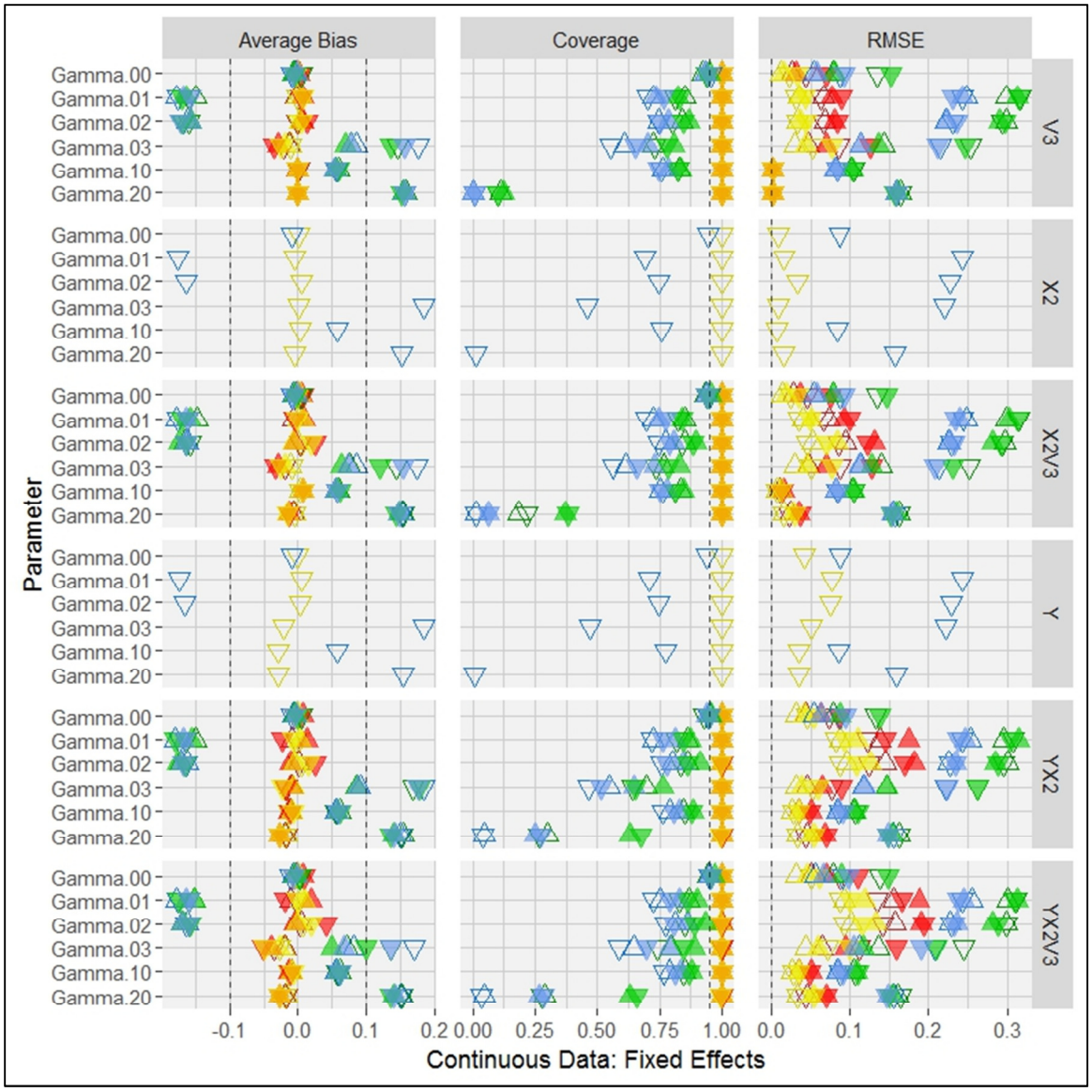


Figure 4.1. Average bias, confidence interval coverage, and RMSE of the fixed effects for continuous data. Statistics used for assessing the accuracy of results are shown in each column. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼ . A single ★ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across ICC, J, or MAR rate.

When looking at the plots in this section, two general trends are key for interpretation of the results: (a) is there a difference in results across manipulated variables of ICC, number of groups, and missingness rate, and (b) do the observed values

fall above or below the chosen guideline values. Regarding the difference in results across manipulated variables, an analysis procedure which performs consistently across all of the manipulated variables would present as a star shape in which both filled and outlined shapes are overlapping. For R/JAGS this star shape would appear to be orange in color (overlap of yellow and red) as seen for the $BIAS_{AVG}$ of γ_{20} for V3 missingness (top left plot panel); for BLImP/JAGS, this star would be a turquoise color (overlap of green and blue) such as the BLImP/Mplus values for $BIAS_{AVG}$ for γ_{10} in the same (top left) panel for V3 missingness. To the extent that results for the two missingness rates differ from each other but are similar across other manipulated variables, a filled-in star and an outlined star would be evident in the plot. An example of this kind of result is seen for the CI coverage rate value of γ_{20} for X2V3 missingness with BLImP/Mplus in which an outlined blue star of *MAR* = 20% is located next to the filled in blue star for *MAR* = 40% (for *J = 50).* Differences between conditions with 25 or 50 L2 groups would be shown by separation between either the red and yellow (R/JAGS) or green and blue (BLImP/Mplus) markers as shown for RMSE values for γ_{01} with V3 missingness (top right plot panel) in which neither the red/yellow markers nor the green/blue markers are overlapping. Disparity between conditions with an ICC of .1 and an ICC of .3 is evident as the two triangle shapes of a given color spread and a star is no longer formed. An example of this would be CI coverage for BLImP/Mplus for γ_{03} with 20% missingness on X2V3 and *J* = 50 (blue, outlined triangles) in which there is a shifting between the upward- and downward-facing triangles so that a star shape is no longer visible.

The second trend of interest is whether these markers fall within the lines used to guide interpretation. For this same top left plot cell for $BIAS_{AVG}$ and V3 missingness in Figure 4.1, several markers fall between the guidelines of -.1 and .1; others do not. Looking across the columns and rows of plots in Figure 4.1, a couple of additional characteristics are of interest. First, having plots paneled in this way allows us to compare across missingness patterns for a chosen statistic by following down a column of plots and across statistics for a given missingness pattern by looking across the three columns in a row of plots. Second, looking across the rows and columns, note that for MAR patterns in which missingness is imposed on only X2 or only Y, only one condition of 500 replications was analyzed. For these two rows of plots, only downward-pointing triangles, only yellow and blue markers, and only outlined shapes are visible which indicates that this condition had *ICC* = .30, *J* = 50, and rate of missingness of 20%.

Fixed Effects Average Bias

Again, looking down the column for $BIAS_{AVG}$ in Figure 4.1, repeated patterns of results are evident in each row. First, it appears that all of the R/JAGS conditions (red and yellow markers) fall within the guidelines of +/- .1. The blue and green markers for BLImP/Mplus fall between these guidelines for γ_{00} and γ_{10}, below the -.1 guideline for γ_{01} and γ_{02}, within the guidelines for *ICC* = .1 conditions but beyond the upper guideline for *ICC* = .3 conditions for γ_{03}, and entirely beyond the .1 guideline for γ_{20}. Star patterns (or near star patterns) are evident for most parameters when R/JAGS or BLImP/Mplus were used (meaning that there is no difference in estimation of parameter values across the manipulated factors of ICC, *J*, and MAR rate)—with the exception noted for γ_{03}.

These patterns are generally repeated for all of the missingness patterns with more variation in $BIAS_{AVG}$ shown for conditions with missingness on larger number of variables.

Fixed Effects CI Coverage

Turning to the column for CI coverage, the difference in patterns between R/JAGS and BLImP/Mplus becomes more noticeable. As mentioned previously, CI coverage captures the percent (or proportion) of replications in a condition in which the value of the parameter estimate for the complete data fell between the lower and upper values of the estimated CI coming from analysis of the same dataset with missingness imposed (which was imputed and analyzed via R/JAGS or BLImP/Mplus). In this column of plots, the guideline is placed at .95 to indicate the threshold for 95% CI coverage. A marker falling to the right of this line indicates that the value of the parameter estimate from the complete data fell within the CI from the dataset with missingness 95% or more of the replications in that condition. A marker to the left of this line at .95 indicates that the value of the parameter estimate from the complete data fell within the estimated CI from the data which had missingness in less than 95% of replications for that condition.

Looking down this column, all estimates for R/JAGS fall near/at 1.0 and are well to the right of the cutoff of .95. For BLImP/Mplus, CI coverage estimates for γ_{00} fall on or very near the cutoff of .95 (shown by the turquoise star –or blue marker for Y or X2 patterns—which overlap the guideline at .95 for all rows of CI coverage plot panels). For the other parameters in the model, BLImP/Mplus coverage rates range from as low as near zero for the random slope fixed effect γ_{20} (as shown by the blue outlined stars near

a value of 0)to close to .95 for γ_{00} (shown by the turquoise star shapes mentioned previously).

Many of these parameters show some disparity between conditions on the basis of ICC value with conditions with higher ICCs showing lower coverage rates. An example of this pattern is the blue, outlined, downward-pointing triangle for γ_{03} in the V3 missingness pattern plot (center, top row). Differences are also noticeable between numbers of clusters with the blue markers for $J = 50$ showing lower coverage rates than conditions with $J = 25$ (green markers). Additionally, an examination of the differences across missing data rates show an interesting pattern of the higher MAR rate of 40% resulting in better coverage than the lower missingness rate of 20%. Again, the variability in coverage estimates increases as more variables had missingness that was imputed.

Fixed Effects RMSE

Finally, RMSE is a measure of the average difference (i.e., error) between the estimated parameter values using the missing/imputed data and the estimated parameter values from the complete data. RMSE values near 0 indicate little difference between these two sets of data for a given condition. This is the reason for drawing a guideline at 0 on the left side of the RMSE plots to highlight whether or not RMSE values are close to 0. Looking down the column, similar patterns are again visible across the patterns of missingness.

Overall, R/JAGS shows smaller RMSE values than BLImP/JAGS with completely overlapping star shapes evident when looking at the L2 slopes for V3 missingness. The components of these star shapes start to stretch or spread for other missingness patterns indicating differences in RMSE values for the different levels of

ICC (e.g., the green markers for γ_{02} with X2V3 missingness). RMSE values for γ_{00} are again closer between R/JAGS and BLImP/Mplus than for the other parameters in this group. Furthermore, a more noticeable effect of number of groups and ICC is evident for RMSE than for the other statistics that were measured. This is evidenced by the shifting apart of the upward- and downward-facing triangles and the distance between red and yellow markers for R/JAGS and blue and green markers for BLImP/Mplus.

Across all missing data patterns and parameters, yellow R/JAGS and blue BLImP/Mplus markers indicating conditions with $J = 50$ returned smaller RMSE values than RMSEs observed in conditions with $J = 25$ (red R/JAGS and green BLImP/Mplus markers). As the number of variables with missingness increases, we see increased differences between conditions with differing values of ICC (shown by spread/shifted individual triangle shapes). In particular, conditions with lower ICC values tend to have smaller RMSE values, and again, this pattern is more extreme as more variables had missingness that was imputed.

Many outlined and filled star patterns are visible indicating that RMSE values for conditions with either ICC value were similar even though these values differed when other manipulated variables such as number of L2 clusters or missingness rate were taken into account. Additionally, a number of different patterns are visible when looking across the rates of missingness; however, in general, little difference between the two rates or some slight variation is evident between the two rates with some conditions showing smaller RMSE values for the lower missingness rate of 20% (shown by outlined markers) and other conditions displaying slightly smaller RMSE values for the higher rate of

missingness (shown by filled markers). It should be noted that these differences between conditions with different missingness rates are on average less than .25.

Random Effects Average Bias

Figure 4.2 shows the values of the variances of the random effect parameters from both L1 and L2[21]. Again, looking down the column for average bias values, we see similar patterns across each row in this figure. For missingness on V3, X2, X2V3, and Y, we see that all of the $BIAS_{AVG}$ values for R/JAGS (red and yellow markers) are between +/- .1 and, in general, tend to be very near 0. The same can be said for most of the parameters when looking at $BIAS_{AVG}$ values for BLImP/Mplus (green and blue markers) with the exception of the variance $\sigma^2_{u_1}$ for the random slope β_{1j} and the covariance estimate between u_{0j} and u_{1j} in which case conditions with an $ICC = .3$ showed a negative bias value below the cutoff of -.1. Overall, more conditions using BLImP/Mplus show smaller $BIAS_{AVG}$ values for the random effects variances and covariances than were observed for the fixed effects estimates.

Looking at the more complex missingness of YX2 and YX2V3, the general pattern is primarily the same as was observed in other missingness pattern with the exception the R/JAGS returns average bias values that are below the lower guideline of -.1 with $BIAS_{AVG} \approx -.2$ for σ^2_r (shown by the red and yellow star shapes on the left side of these plot panels). This is also true for σ^2_r when missingness is present only on Y. Similarly to the patterns observed with the fixed effects, a larger disparity in average bias

[21] Note that X-axis scales differ slightly across figures. The plotting package used to create these paneled plots allows the axes to vary by column and row but not allow axis limits and tick marks to be set for each axis.

estimates for both R/JAGS and BLImP/Mplus are observed for YX2 and YX2V3 than for other missingness patterns. For $\sigma^2_{u_0}$ especially, a pattern of differences between levels of ICC is more clearly depicted for R/JAGS but also evident for BLImP/Mplus with the upward- and downward-facing triangles being more spread apart than was observed for the fixed effect parameters with positive BIAS_{AVG} value for *ICC* = .1 (upward-facing triangles toward the upper guidelines) and negative BIAS_{AVG} values for *ICC* = .3 (downward-facing triangles located near the lower guideline).

Random Effects CI Coverage

Next, examining CI coverage shows that for V3 and X2V3 missingness patterns, both analyses procedures produced coverage rates at or below the threshold of .95 for σ^2_r and $\sigma^2_{u_0}$ associated with L2 intercept β_{0j} (for these two parameters, markers for R/JAGS and BLImP/Mplus are overlapped). For the Y, YX2, and YX2V3 missingness patterns, the coverage rates for R/JAGS are very low with values near 0 as shown by the red and/or yellow markers on the left side of the CI coverage plots in the rows for Y, YX2, and YX2V3. Patterns are similar for BLImP/Mplus across missingness patterns in that coverage rates for σ^2_r and $\sigma^2_{u_0}$ associated with L2 intercept β_{0j} are above .95 and at/near 1.0 (shown by turquoise star shapes on the right side of each plot panel). For the covariances between L2 random effects and variance $\sigma^2_{u_1}$ on the random slope β_{1j} coverage rate are quite low ranging between near 0 and .5 for BLImP/Mplus.

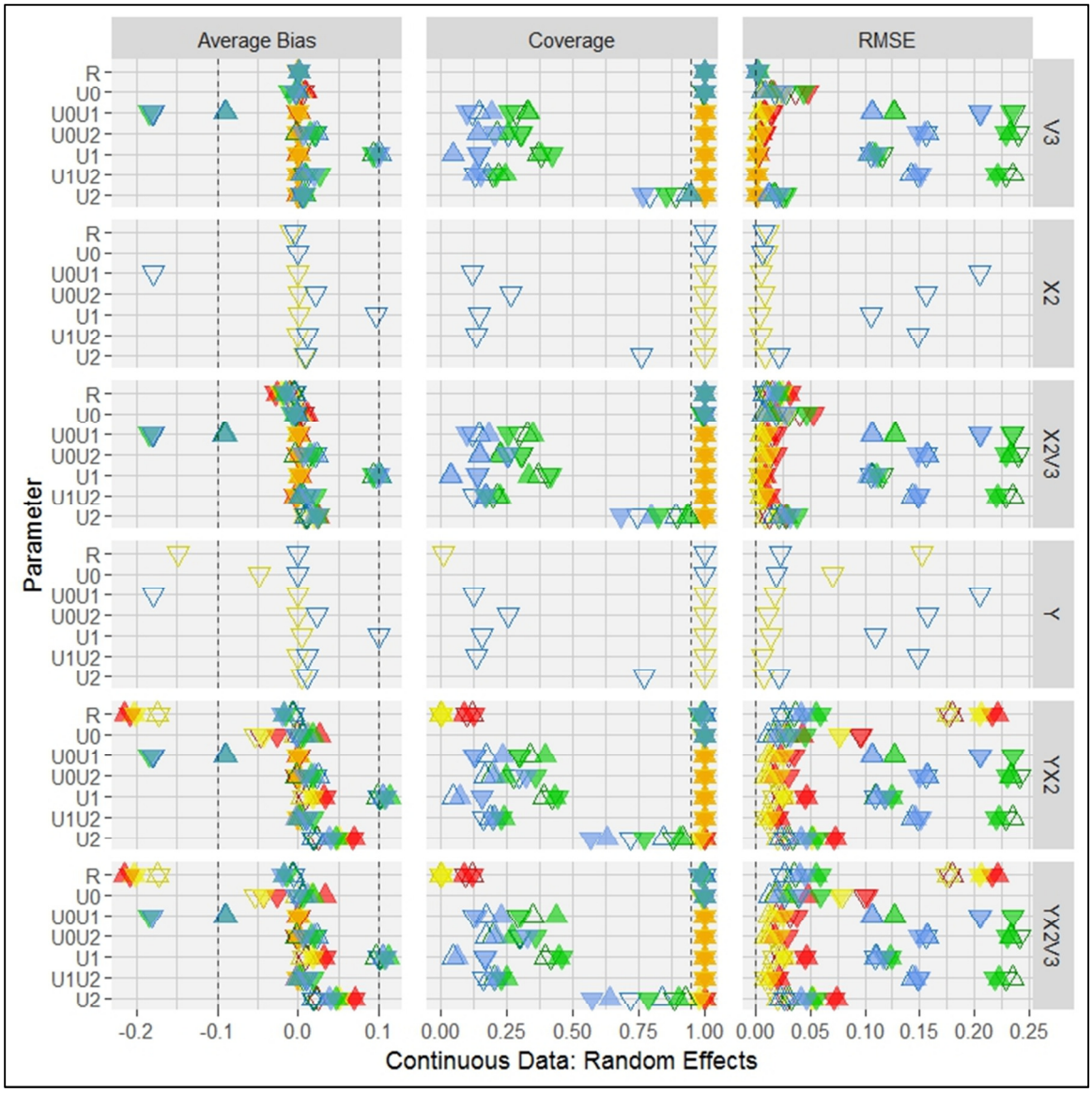


Figure 4.2. Average bias, confidence interval coverage, and RMSE of the random effects for continuous data. Statistics used for assessing the accuracy of results are shown in each column. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼ . A single ★ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across ICC, J, or MAR rate.

For the variance $\sigma^2_{u_2}$ on the random slope β_{2j} (labeled as *U2* in Figure 4.2)

coverage values were higher than .5 and in some instances at/above .95 for

BLImP/Mplus. In these cases, coverage rates for *ICC* = .1 are higher than those for *ICC* =

.3 as shown by the upward-facing triangle being located closer to the guideline of .95 than the downward-facing triangles. Overall, coverage rates for the random effects did not show the pattern of the MAR rate of 20% returning drastically lower coverage values than those observed for the 40% missingness rate. Some differences are present (e.g., $\sigma^2_{u_1}$ for YX2 missingness, shown by a slight shift between filled and outline blue or green triangles), but these differences are slight compared to what was observed with the fixed effects in Figure 4.1. For R/JAGS, coverage for parameters other than σ^2_r was at/near 1.

Random Effects RMSE

Similar to patterns observed in other statistics in Figure 4.2, RMSE values for the random effects with missingness on V3 and X2V3 are much smaller than those observed for the fixed effects with agreement between analysis procedures and across manipulated variables evident in the presence of many star-shaped markers. Much like the pattern observed with CI coverage, RMSE values for R/JAGS and BLImP/Mplus are similar for σ^2_r, $\sigma^2_{u_1}$, and $\sigma^2_{u_2}$ with values below .05 for V3, X2, X3V3, and Y missingness. With the exceptions of σ^2_r and $\sigma^2_{u_0}$ for Y missingness and σ^2_r, $\sigma^2_{u_0}$, $\sigma^2_{u_1}$, and $\sigma^2_{u_2}$ for YX2 and YX2V3 missingness patterns, RMSE values for R/JAGS are at/near 0 for most parameters and across missingness patterns (shown by orange overlapping stars or nearly-overlapping sets of triangles). Again, with missingness on Y, YX2, or YX2V3, σ^2_r produces the worst values of RMSEs with values between .15 and .25 for R/JAGS. Values for BLImP/Mplus range from .1 to .25 for the remaining parameters across missingness patterns. When looking at RMSE, we see less separation between markers which differ based on ICC (meaning more star shapes), but the few instances in which markers do have distance

between upward- and downward-pointing triangles show again that conditions with an ICC values of .30 (downward-facing triangles) return more extreme RMSE values.

Categorical Data with .50/.50 Split

The assessment of the results for the .50/.50 split dichotomized data are presented in Figures 4.3 through 4.8 which are similar to Figures 4.1 and 4.2, but now each plot captures one of the three statistics of interest and highlights the comparison between the two types of data used for either the fixed or random effects parameters. Figure 4.3 shows the comparison of fixed effects $BIAS_{AVG}$ values. $BIAS_{AVG}$ results for the .50/.50 split dichotomous data in the left-hand column and the continuous results in the right hand column (shown previously in Figure 4.1). The organization of this series of plots is similar with the missing data patterns listed in the rows. The same schema for shapes, colors, and fill continues here. Panels capturing the .50/.50 split data for YX2 and YX2V3 only contain a few of the expected shapes, colors, and filled/unfilled markers for R/JAGS conditions due to convergence issues during the model testing phase (meaning that these models were not analyzed on the ASU cluster). Looking at the panels for YX2V3, the upward-pointing, red, outlined triangle markers show that only a single condition with $ICC = .1$, $J = 25$, and $MAR = 20\%$ with this pattern of missingness converged in R/JAGS.

Fixed Effects Average Bias.

As mentioned, Figure 4.3 shows the comparison of $BIAS_{AVG}$ values across data types and missingness patterns for fixed effects parameters. Similar to comparisons made between data type, we see that $BIAS_{AVG}$ values are more variable for dichotomous data than values observed for continuous data. This is true for both analysis procedures. For

R/JAGS, $BIAS_{AVG}$ values continue to be within the interval of -.1 to .1; however, more estimates of $BIAS_{AVG}$ are closer to these guidelines for dichotomous data than were observed for continuous data. R/JAGS conditions with larger $BIAS_{AVG}$ values tend to have 50 L2 groups and an ICC value of .3 as shown by the red, downward-pointing triangles visible in Figure 4.3 with γ_{00} for YX2 missingness as an example of this; note the orange upward-pointing triangle for the X2V3 missingness pattern is an exception that indicates that values are similar for conditions with *ICC* = .1 and *MAR* = 40% for both *J* = 25 (red markers) and *J* = 50 (yellow markers) resulting in an the orange upward-facing marker for γ_{02} with .50/.50 split data

Random Effects Average Bias

$BIAS_{AVG}$ for random effects parameters are presented in Figure 4.4[22]. Note that the scale of the x-axis for this plot differ from what was previously shown in Figure 4.3. Figure 4.4 shows that for the V3, X2V3, X2, and Y patterns of missingness the bias values are similar across data types when R/JAGS is used to perform the imputation and analysis. When the missingness pattern becomes more complex, the estimates of $BIAS_{AVG}$ become much more extreme with some $BIAS_{AVG}$ values greater than 1.5.

[22] In R/JAGS, estimates of r_{ij} are not available for the dichotomous conditions as the Bernoulli model used does not include this parameter in the estimated model; therefore, estimates are not drawn and sampled values cannot be saved out for comparison to the continuous data results.

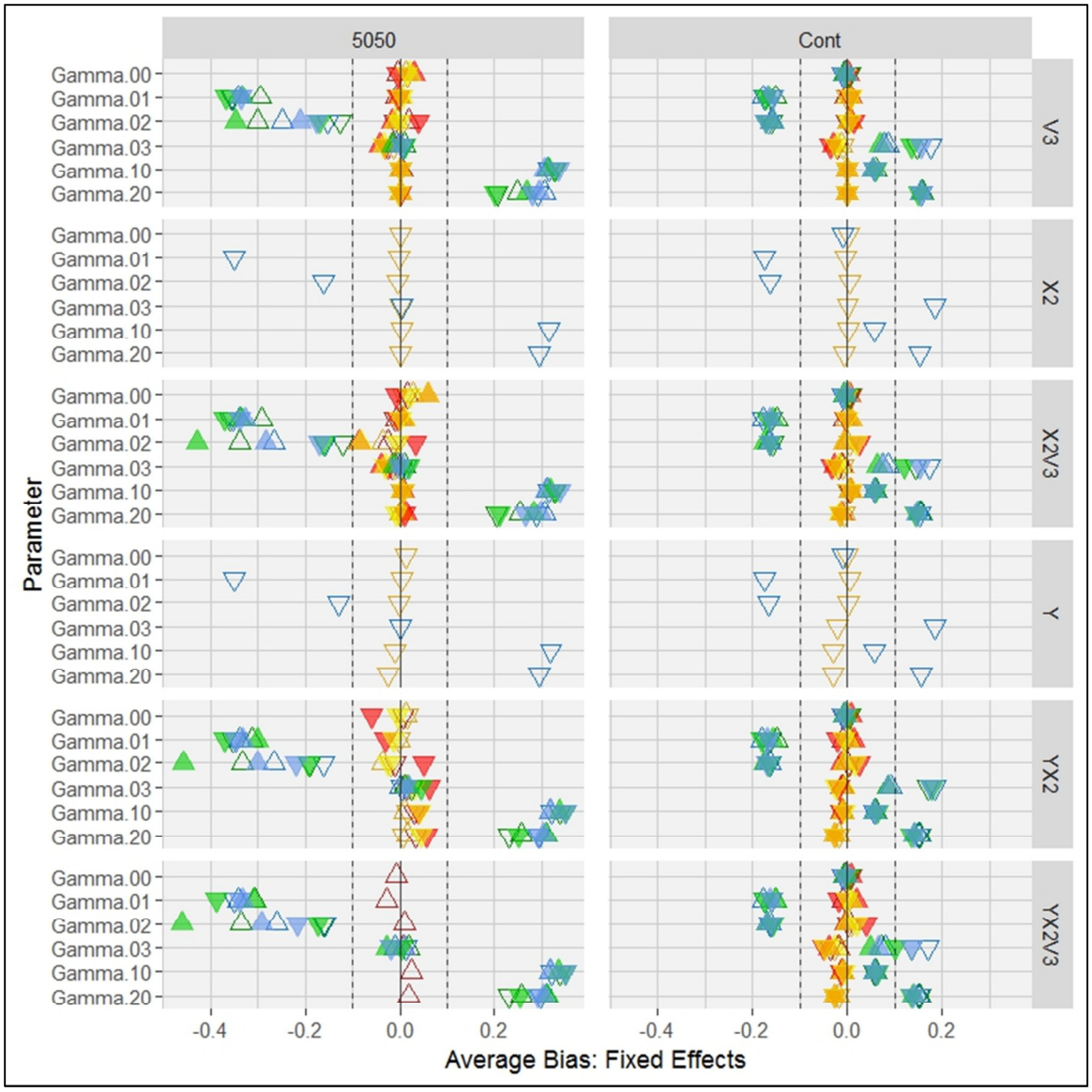


Figure 4.3. Average bias of the fixed effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ✸ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across ICC, J, or MAR rate.

Looking at BLImP/Mplus, Figure 4.4 shows that $BIAS_{AVG}$ for one of the covariances between L2 random effects [cov(u_{0j}, u_{1j})] is lower for dichotomous data than was observed with continuous data (shown by the orange and turquoise star shapes

in many of the left column plot panels for .50/.50 data). Otherwise, $BIAS_{AVG}$ values are much more extreme for the dichotomous data conditions than were observed with continuous data (shown by a wider spread between markers in general). Whereas almost all parameters has $BIAS_{AVG}$ values between +/- .1 for continuous data, the only parameters with any $BIAS_{AVG}$ near +/- .1 for dichotomous data are σ_r^2, $\sigma_{u_0}^2$, and cov(u_{0j}, u_{1j}). For dichotomous data, the $BIAS_{AVG}$ for BLImP/Mplus range from +/- 1.

Fixed Effects CI Coverage

Comparing patterns across data types for R/JAGS for conditions which were able to converge, coverage rates are similar when either dichotomous or continuous data were used. As is evident in Figure 4.5, all of the coverage rates for R/JAGS for the .50/.50 split data are higher than the cutoff of .95 and are at/near a value of 1. Looking at the dichotomous results for YX2 and YX2V3 missingness patterns again shows that some conditions are missing due to lack of convergence (i.e., the star shapes are missing some of their components or are simply a single shape as mentioned for the YX2V3 pattern).

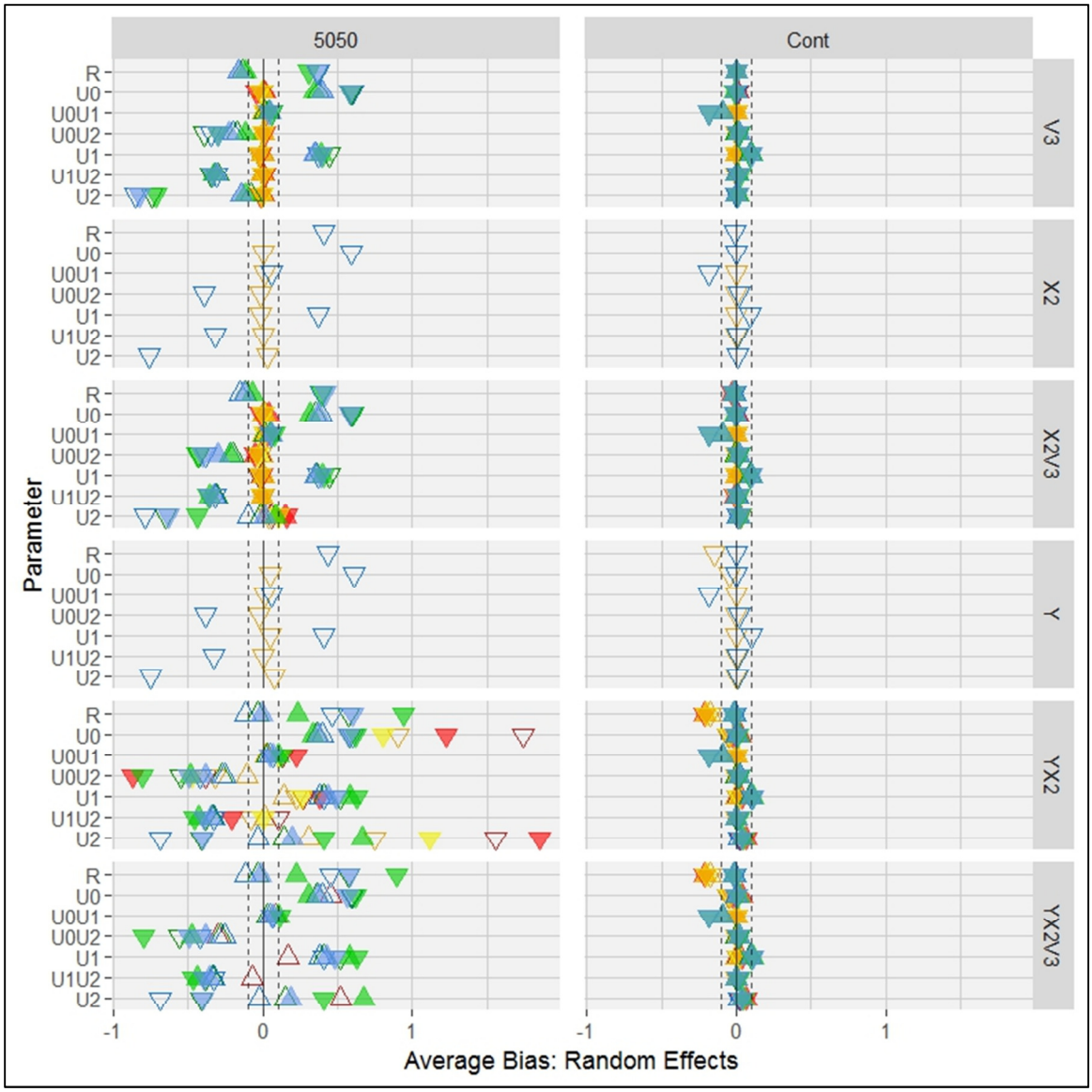


Figure 4.4. Average bias of the random effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ✸ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across ICC, J, or MAR rate.

Turning to the BLImP/Mplus CI coverage rates, patterns of results are noticeably different when comparing the dichotomous and continuous data results[23]. For γ_{01}, γ_{02}, and γ_{03}, BLImP/Mplus returns coverage rates that are nearer to the cutoff of .95 for the .50/.50 data than for the continuous data. For γ_{03}, these coverage rates are at, above, or slightly below .95 for all patterns of missingness (shown by the turquoise star shapes in the left column of plot panels). Coverage for the random slope fixed effects shows that BLImP/Mplus that coverage rates are higher with dichotomous data for γ_{20} but are much lower for γ_{10}. With the exception of the L2 random slope values, the amount of overlap (or disparity) between upward-facing and downward-facing triangles is similar between data types. A greater difference between green and blue markers shows that a difference exists between coverage rates for conditions with $J = 25$ and conditions with $J = 50$ with lower rates for the conditions with more groups. Again, the patterns for the rates of missingness observed with continuous data are repeated for the .50/.50 split data with worse coverage for some lower missingness rate conditions.

Random Effects CI Coverage

Figure 4.6 shows CI coverage for the random effects. For this set of parameters, coverage rate differences between data types are apparent for both analysis procedures. Although coverage rates for R/JAGS for all V3, X2V3, X2, and Y conditions and for the single YX2V3 condition are above .95, coverage rates below .95 are observed for the YX2 condition. For this missingness pattern, coverage rates below .95 are observed for

[23] Note that Mplus does not provide intercept (γ_{00}) estimates for outcomes that are categorical.

conditions with *ICC* = .3 and *J* = 50 (shown by the red and yellow markers that are below the guideline at .95).

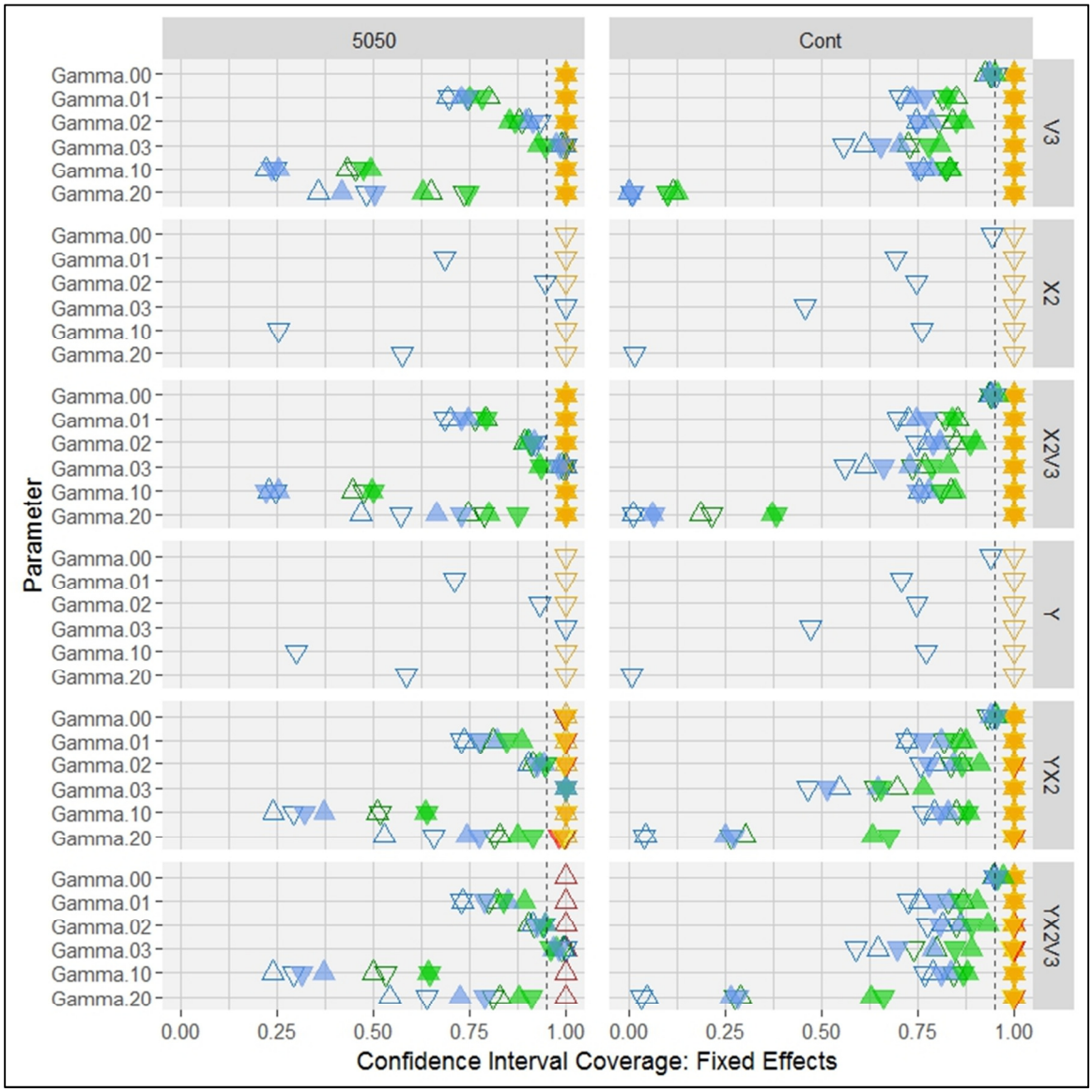


Figure 4.5. Confidence interval coverage of the fixed effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ✸ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across ICC, J, or MAR rate.

Looking at coverage rates for BLImP/Mplus, we see that with minimal exception, coverage is worse for the dichotomous data conditions than those observed for the continuous data. Coverage for the covariances between L2 random effects and for $\sigma^2_{u_1}$ appears to be slightly larger for the dichotomous data across all missingness patterns. Across both data types, conditions with more L2 groups return lower coverage rates. The pattern for ICC value does differ across data type with more conditions with $ICC = .1$ showing lower coverage rates with dichotomous data than those observed for continuous data (shown by downward-facing green and blue triangles located closer to the guideline than upward-facing triangles). This pattern for ICC was observed in coverage rates for the fixed effects as well, but the difference is more extreme for coverage of random effects parameters.

Fixed Effects RMSE

RMSE estimates for the fixed effects parameters are shown in Figure 4.7. Across all datatypes, analysis procedures, and missingness patterns RMSE values are larger for the dichotomous data than for the continuous data[24]. For the continuous data conditions, all RMSE values are below .5 with most below values of .25. In the dichotomous data conditions, RMSE values range as high as 1.5 for BLImP/Mplus (as shown by the green markers that are beyond the guideline at 1.0).

[24] An additional guideline has been added to the RMSE plot at a value of 1.0. Due to the extreme RMSE values observed this additional guideline helps with interpretation of difference between data type and fixed vs. random effects.

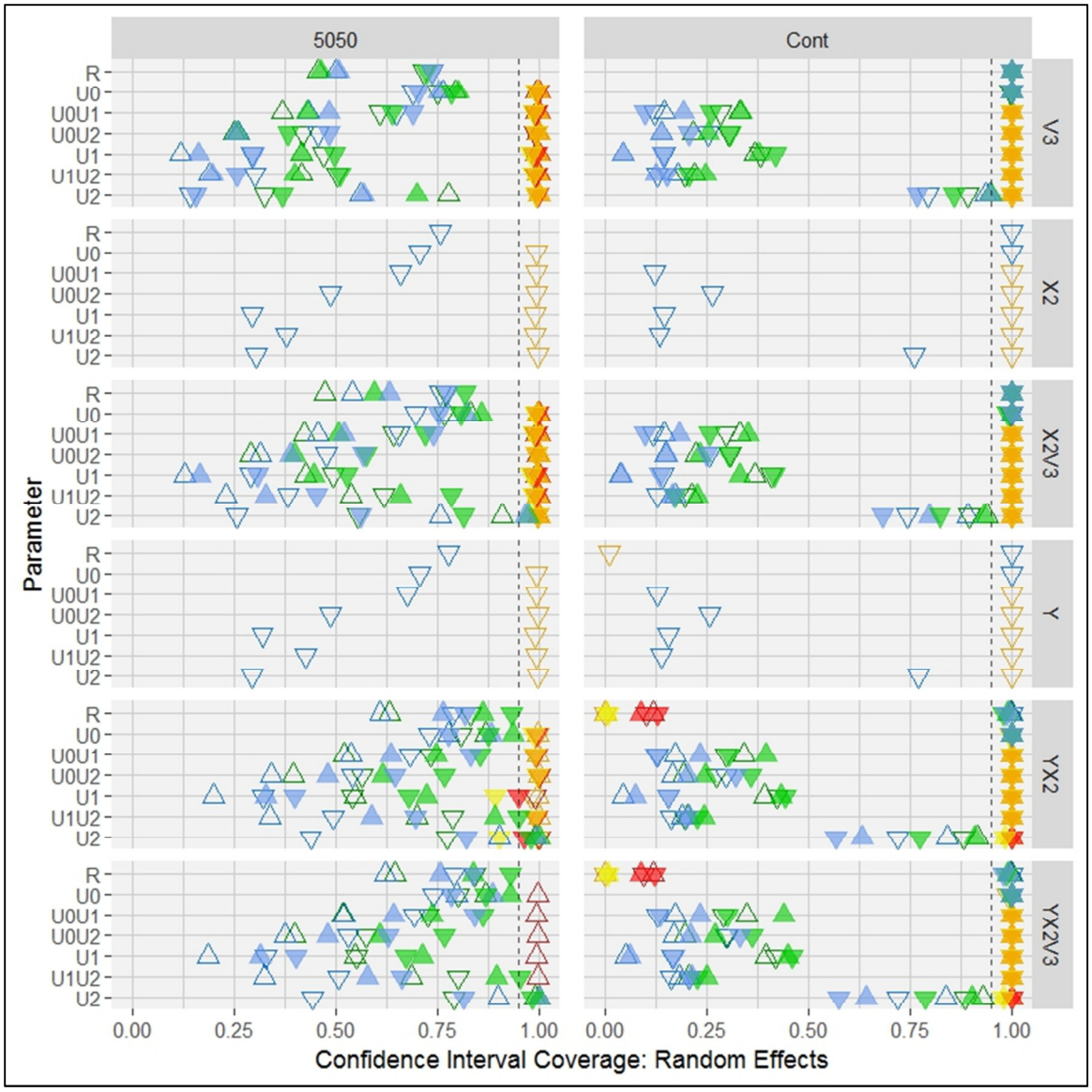


Figure 4.6. Confidence interval coverage of the random effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ★ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across *ICC*, *J*, or MAR rate.

Looking at the RMSE values for the R/JAGS conditions with dichotomous data, we see large RMSE values for the fixed effects associated with the L2 intercept. RMSEs of .25 or smaller were observed for the fixed effects for the L2 slopes. Differences in

RMSE across manipulated variables are more noticeable for the YX2 and YX2V3 missingness patterns for both analysis procedures.

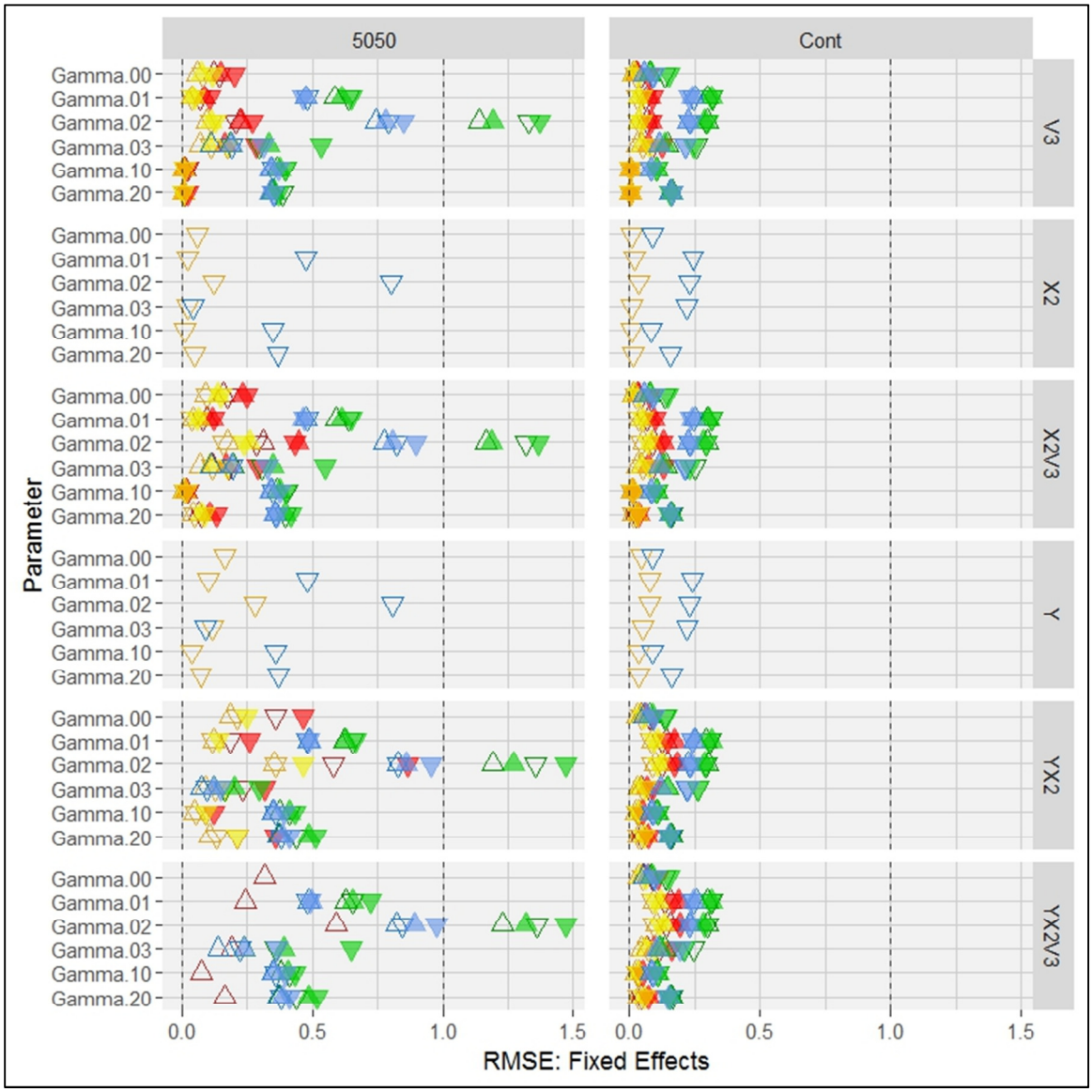


Figure 4.7. RMSE of the fixed effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ★ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across *ICC*, *J*, or MAR rate.

The values of RMSE for BLImP/Mplus show the most difference between dichotomous and continuous data for γ_{01} and γ_{02}. RMSEs for γ_{03} were similar across data types for most missingness patterns and smaller than RMSEs for continuous data when missingness was on X2 only or Y only. As more variables had missingness, larger differences were observed. Again, patterns of more extreme values were observed for conditions with $J = 50$ and conditions with $ICC = .3$ (shown by blue, downward-facing triangles).

Random Effects RMSE

In Figure 4.8, random effects variance and covariance estimates for R/JAGS with V3, X2V2, X2, and Y missingness patterns show RMSE values that were larger than those for continuous data as all markers for in the plots with .50/.50 split data for these missingness patterns are closer to the guideline at 1.0 than values for the continuous data in these missingness patterns. The RMSE values observed with .50/.50 data for V3, X2, X2V3, and Y are smaller than those observed with dichotomous data and missingness was present on YX2 or YX2V3. Due to the lack of convergence with R/JAGS for YX2 and YX2V3, statements on the differences between dichotomous and continuous data results cannot be made for these conditions.

Looking at RMSEs for BLImP/Mplus, values for the .50/.50 split were much higher than those observed for continuous data. Many RMSEs were at or near a value of 1 for dichotomous data across the missingness patterns. Values tended to be most extreme for the YX2 and YX2V3 missingness patterns than for the other patterns of missingness. RMSEs for dichotomous data, the RMSEs and YX2 and YX2V3

missingness were much smaller than those observed for R/JAGS for the same data type and missingness patterns.

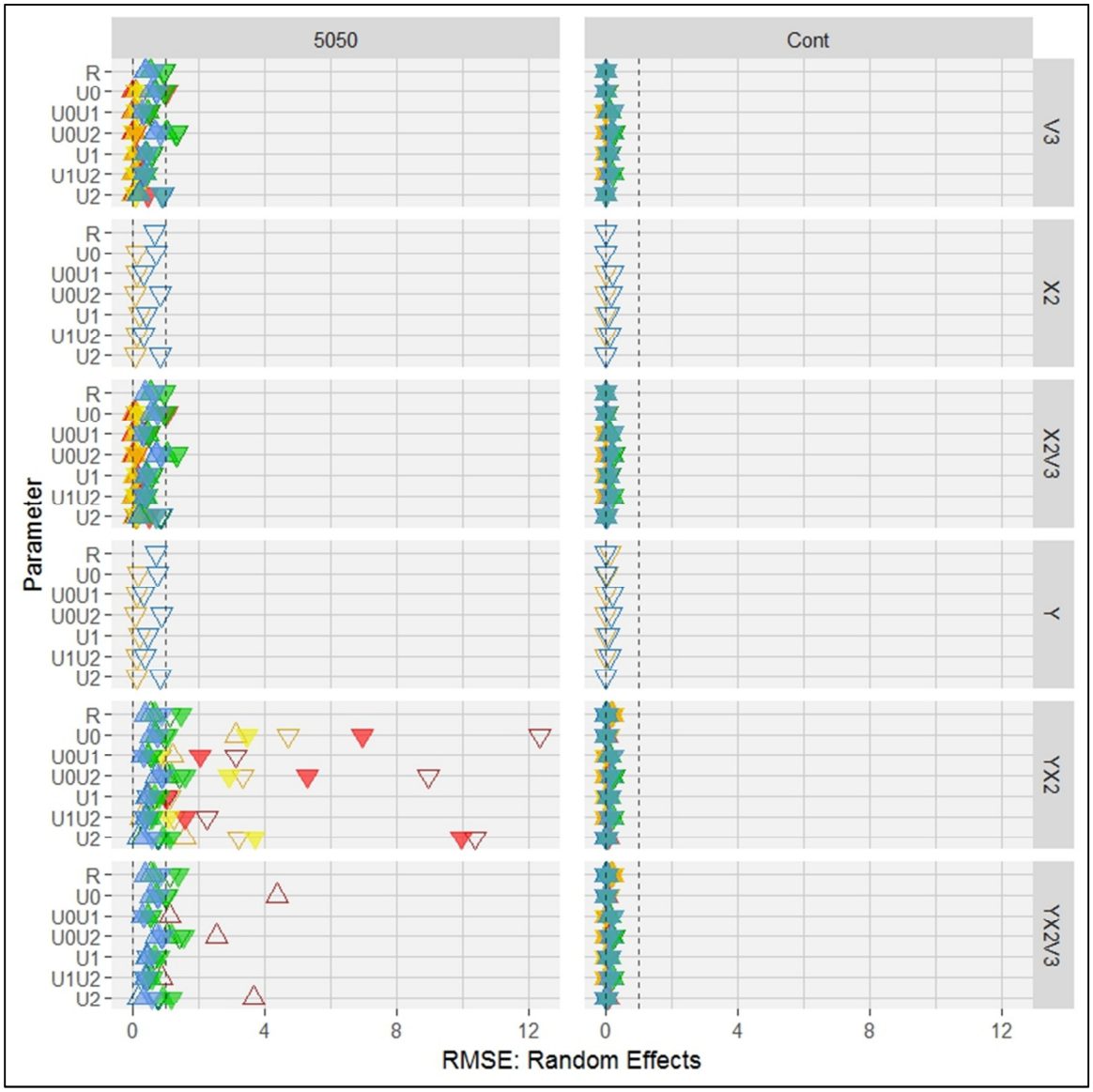


Figure 4.7. RMSE of the random effects for .50/.50 split and continuous data. Data types are shown across columns. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *ICC* = .3 is shown by▼, *J* = 25 is shown by red (R/JAGS) or green (BLImP/Mplus), *J* = 50 is shown by yellow (R/JAGS) or blue (BLImP/Mplus), *MAR* = 20% is shown by △ or▽, and *MAR* = 40% is shown by ▲ or▼. A single ✸ shape of either orange or turquoise in a row indicates that no difference in the value of the statistic across *ICC*, *J*, or MAR rate.

Follow-up Analyses

During analysis of the results for both the continuous data and the .50/.50 split dichotomous data some patterns of results for the BLImP/Mplus conditions showed that the lower missingness rate of *MAR* = 20% produced more extreme $BIAS_{AVG}$ values, worse CI coverage, and larger RMSE values. This pattern of results runs counter to theory that as missingness *increases* the impact on recovery of parameter estimates would increase as well.[25] Because this unexpected pattern was observed, three sets of additional analyses were conducted in an attempt to better understand what was occurring in these instances. These additional analyses also provided a view of the results which allowed the performance of the chosen imputation software to be isolated from the performance of the chosen analysis software.

Influence of Missing Data Rates

In order to assess whether the method of imposing missingness was a potential influencing factor in the unexpected results patterns showing that lower missingness rates returned worse values of bias, coverage and RMSE, additional analyses were conducted using the V3 missingness pattern. The condition from the V3 missingness pattern with *ICC* = .3 and *J* = 25 was selected for this follow-up study because it was one of the most simple missingness patterns in which this pattern for MAR rate was observed (see $BIAS_{AVG}$ and RMSE values for γ_{03} in Figure 4.1 as an example). The values of $BIAS_{AVG}$, CI coverage, and RMSE from the original set of conditions with *ICC* = .3, *J* = 25, *MAR* = 20% and *MAR* = 40% were compared to results from estimation of versions of the

[25] This effect for rate of missingness was documented—among others—by Enders and Keller (in press).

datasets with *MAR* = 40% in which 20% and 10% missingness imposed manually. This manual method of imposing missingness created a new set of 500 replications with 20% missingness by taking the 40% missingness version of the data in each replication and randomly replacing half of the missing values with the values from the complete data resulting in a dataset with approximately 20% missingness on V3. Using this 20% missingness version of each dataset, half of the missing values were again repopulated with the values from the complete data to create a dataset with approximately 10% missingness on V3.

Imposing missingness in this way allowed for direct comparison of different rates of missingness imposed on the same dataset. Whereas the procedure used for the main set of analyses did not require that the missing values in a *MAR* = 20% missingness dataset make up half of the missingness in the *MAR* = 40% dataset in a given replication. This method for manually imposing missingness does make that requirement. A more direct comparison of versions of each dataset in which the exact missing values are the same across rates of missingness (with additional missing values at higher rates of missingness) eliminates the confounding that could be caused by comparing results from conditions in which the individual missing values are not constant across datasets in a given replication. Analysis results of these two new versions of the data along with results from the previously discussed 40% and 20% missingness versions of the condition with *ICC* = .3, *J* = 25, and missingness on V3 are shown in Figures 4.9 and 4.10.

Fixed effects. Looking at Figure 4.9, values from this condition with 40% and 20% missingness imposed during data generation are shown in the filled-in markers and outlined green markers respectively. Values for versions of this condition in which

missingness was imposed in this section are shown in purple for the 20% missingness rate and pink for the 10 percent missingness rate. This figure shows that for some parameters, there is little difference between missingness rates or in method for imposing missingness as all markers overlap. An example of this pattern is values for γ_{20} in which all four triangles ae almost perfectly overlapping for all three statistics. A parameter in which the unusual patterns for missingness are observed would be γ_{03} in which even the pattern with 10% missingness (pink outlined downward-facing triangle) still shows a lower coverage rate than was found with 40% missingness.

Random effects. Turning to the random effects results found in Figure 4.10, the patterns are of values are similar across missingness rate for many of these parameters for the three datasets that are variations of one another (i.e., 40% using the original method for imposing missingness, 20% using this test method of "removing" missingness, and 10% using this same method). Figure 4.10, does show that there is at times a difference between these three variations on a condition and the version of the condition with 20% missingness that was imposed using the original method. This pattern is not surprising as the 40% and 20% versions of a condition do not come from the same complete dataset. The comparison between conditions with 20% and 40% missingness compares two separately drawn datasets with either 40% or 20% missingness imposed on the complete data for each replication within a condition. It should be noted again that this pattern of conditions returning worse statistics with lower missingness rates than with higher missingness rates is more extreme for certain parameters than for others.

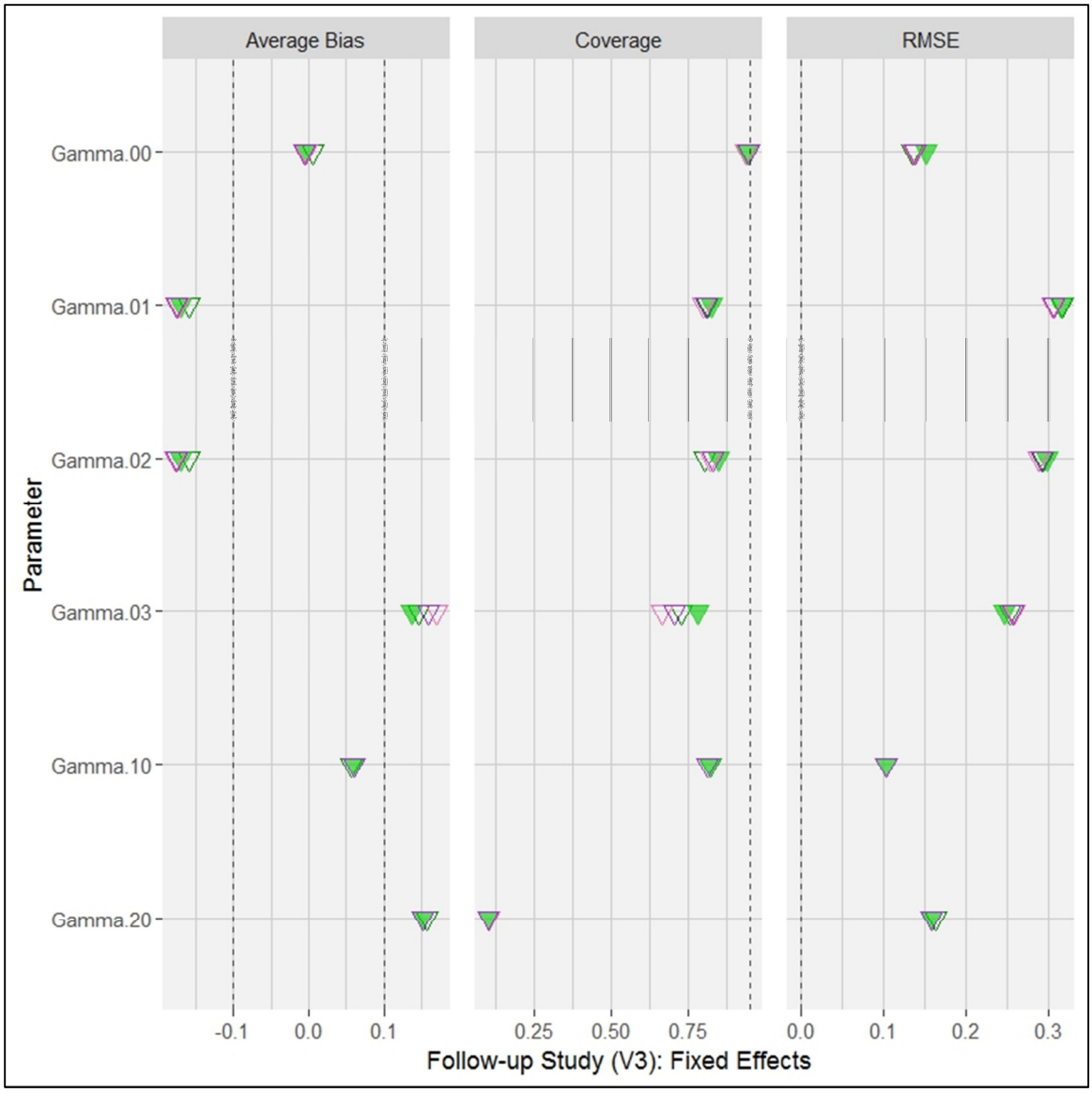


Figure 4.9. Comparison of BLImP/Mplus fixed effects results with V3 missingness for different methods of imposing missingness. The pattern for 40% and 20% missingness imposed during data generation are shown in green and the 20% (shown in purple) and 10% (shown in pink) missingness imposed manually from the 40% missingness version of this conditions are also shown. Each parameter is found on a line of each plot row. *ICC* = .3 is shown by▼, *J* = 25 in green, *MAR* = 20% is shown by▽, and *MAR* = 40% is shown by▼.

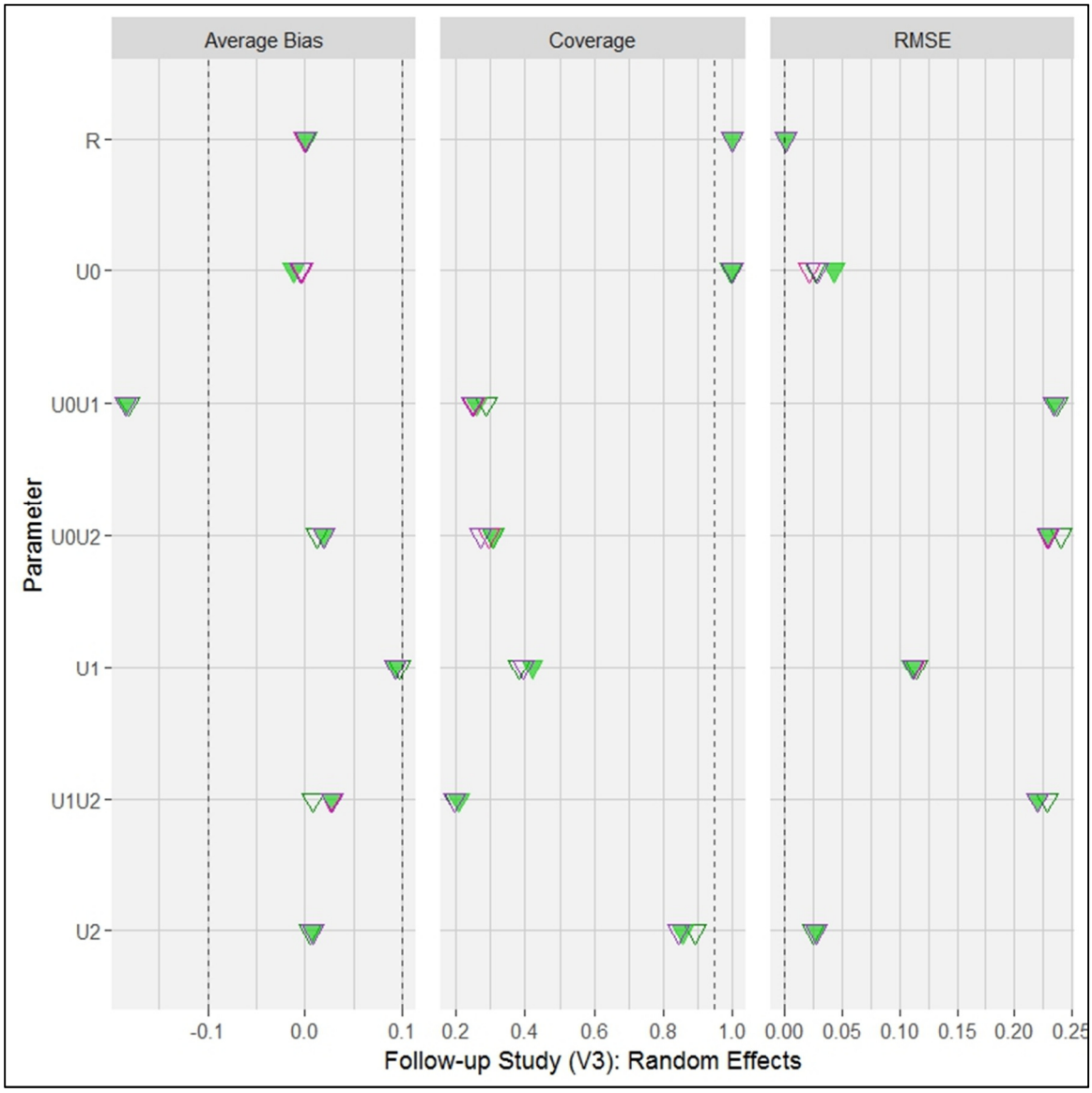


Figure 4.10. Comparison of BLImP/Mplus random effects results with V3 missingness pattern for different methods of imposing missingness. The pattern for 40% and 20% missingness imposed during data generation are shown in green and the 20% (shown in purple) and 10% (shown in pink) missingness imposed manually from the 40% missingness version of this conditions are also shown. Each parameter is found on a line of each plot row. *ICC* = .3 is shown by▼, *J* = 25 in green, *MAR* = 20% is shown by▽, and *MAR* = 40% is shown by▼.

Conclusions. Because the pattern of worse parameter recovery with lower rates of missingness was not as clearly observed in the versions of the data in which missingness had been manually imposed as it had been in the originally run of data, it seems that this

counterintuitive pattern of results may have been influenced by the choice of method for imposing missingness. However, the differences between the two methods appear to be minimal and the patterns observed were not universal across parameters. Because this investigation did not identify a clear cause for the unexpected patterns of results, the remaining follow-up studies focus on the comparability of the imputed datasets.

Imputation in R/JAGS with Analysis in Mplus

To further explore what might be causing R/JAGS and BLImP/Mplus to return such different values of $BIAS_{AVG}$, CI coverage, and RMSE, two additional follow-up studies were conducted to test whether there were differences in the imputed datasets coming from JAGS and BLImP. One way to explore similarity between the two imputed datasets was to impute data using one software routine and run the analysis and pooling phases of the routine in the other analysis framework. Similarity between parameters estimates for data imputed in BLImP and analyzed in Mplus and data imputed in JAGS and then analyzed similarly in Mplus would provide evidence that the inputs to the analysis are similar (i.e., if the analysis model is the same between the two sets of imputed data and the parameter estimates coming from the analysis are the same or similar, then the assumption can be made that the inputs to those analyses—the imputed datasets—are also similar to each other.

For this analysis, two additional continuous data conditions were analyzed using datasets in which missingness was imputed in R/JAGS and analyzed using Mplus. Conditions which were examined included (a) $ICC = .2$, $J = 50$, and 20% missingness on Y only and (b) $ICC = .1$, $J = 25$, and 40% missingness on YX2. These conditions were chosen as they capture opposite extremes of the manipulated variables with high ICC

values, a larger number of clusters, but low missingness or a low ICC value, with a small number of clusters, but with high rates of missingness. They Y missing condition was chosen as it was one of the most simple conditions with missingness on only one variable. The YX2 missingness pattern was chosen as an example of complex condition with missingness on multiple variables.

After imputing missingness in R/JAGS, imputed datasets for each replication were saved out by thinning the 5,000 samples at a rate of 250 resulting in 20 imputed datasets for each replication. The choice to save out 20 datasets was made to replicate what occurred in conditions run with BLImP in which 20 imputed datasets were saved out for each replication of data.

Then, each set of 20 datasets was analyzed using analysis models previously used for these missingness patterns with complete data in BLImP/Mplus. For Y only missingness, Figures 4.11 and 4.12 show the comparison between the typical R/JAGS results, the typical BLImP/Mplus results (both previously discussed and shown in previous figures in rows labeled "Y"), and the results from the follow-up condition imputed in R/JAGS and analyzed in Mplus.

Fixed effects Yjags. Figure 4.11 shows the fixed effects results for $BIAS_{AVG}$, CI coverage, and RMSE. As detailed in Chapter 3, the condition examined here had an *ICC* = .30 (shown by the downward-pointing triangle markers), 50 L2 groups (shown by the yellow and blue markers for R/JAGS and BLImP/Mplus respectively), and a missingness rate of 20% (shown by the use of outlined markers).

Paying closest attention to the location of the black markers denoting the Yjags results, it is evident that the values of $BIAS_{AVG}$, CI coverage, and RMSE very closely

aligned with those from BLImP/Mplus for this condition. For the fixed effects, this was most noticeable for values of γ_{01}, γ_{03}, γ_{10}, and γ_{20}. Some very small differences existed between these three sets of results for γ_{00} and γ_{02}, but Yjags and BLImP/Mplus values were still closer to each other than to R/JAGS values.

Random effects Yjags. The results for the random effect parameters in Figure 4.12 show more variation between the three sets of results than was observed in Figure 4.11, but the differences are still very small. Again, values of all three statistics for BLImP/Mplus and Yjags are either nearly identical or very close to each other. The disparity between values from the BLImP/Mplus and R/JAGS and between Yjags and R/JAGS was similar across statistics.

Fixed effects YX2jags. To confirm that this pattern was not influenced by choice of manipulated variables, the condition with YX2 missingness, *ICC* = .1, *J* = 25, and a missingness rate of 40% was also examined by imputing missing data using R/JAGS and conducting the MLM analysis in Mplus. Figures 4.13 and 4.14 show the comparison across these three sets of results for the selected condition with YX2 missingness. These figures add an additional series of black markers to indicate the version of this condition which was imputed in R/JAGS and analyzed in Mplus which will be referred to as Yjags for Y missingness and YX2jags for the condition with missingness on Y and X2[26].

[26] These conditions will be called "Yjags" and "YX2jags" in order to distinguish them from the R/JAGS or BLImP/Mplus analyses for the Y missingness pattern.

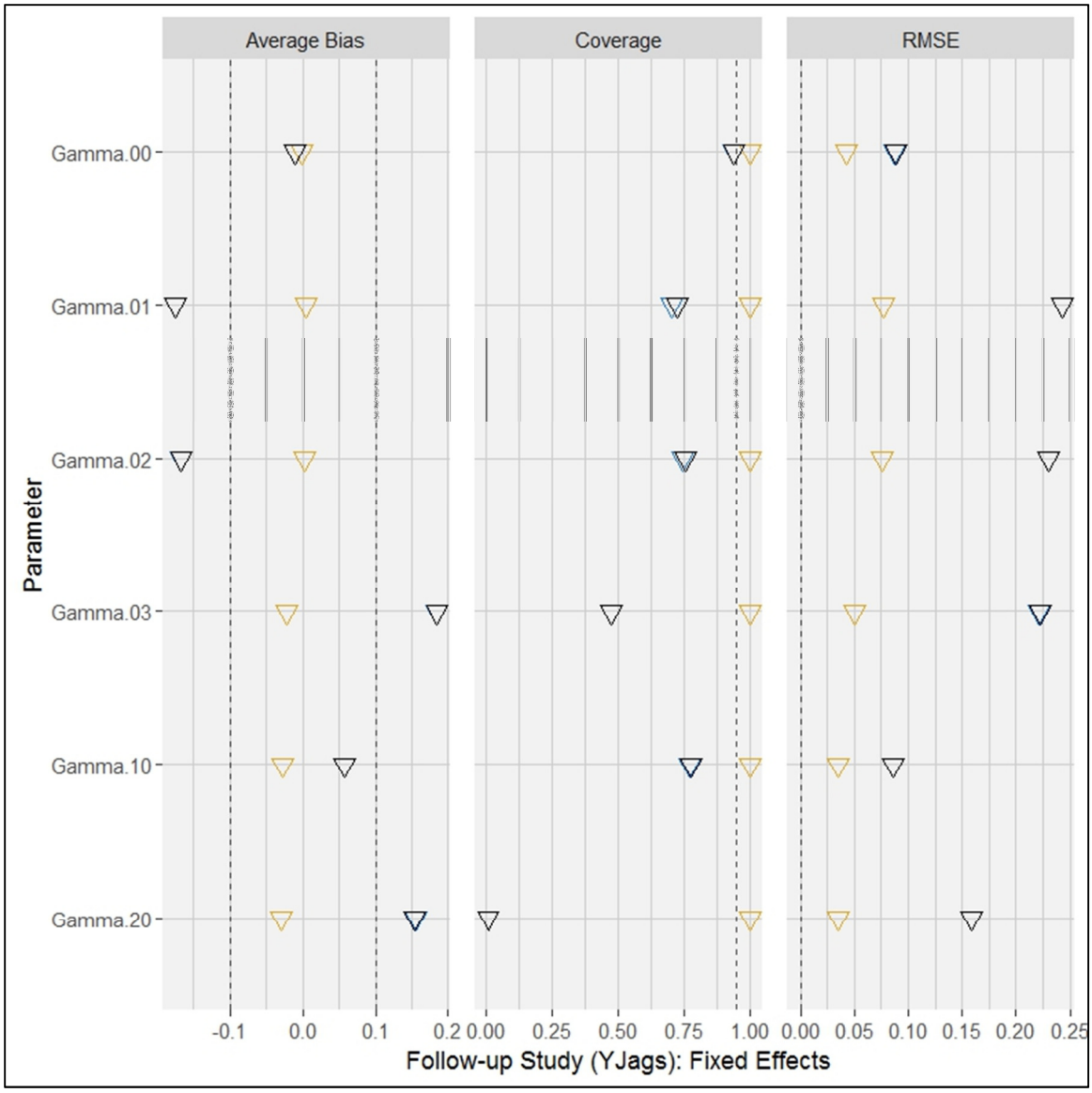


Figure 4.11. Comparison of fixed effects results for Y missingness with data imputed/analyzed in R/JAGS, imputed/analyzed in BLImP/Mplus, and imputed in R/JAGS then analyzed in Mplus. Data imputed/analyzed in R/JAGS are shown in yellow, imputed/analyzed in BLImP/Mplus are shown in blue, and data imputed in R/JAGS and then analyzed in Mplus are shown in black. All data for this comparison had *ICC* = .3, *J* = 50, and 20% missingness. Each parameter is found on a line of each plot row. *ICC* = .3 is shown by▼, *J* = 50, *MAR* = 20% is shown by▽.

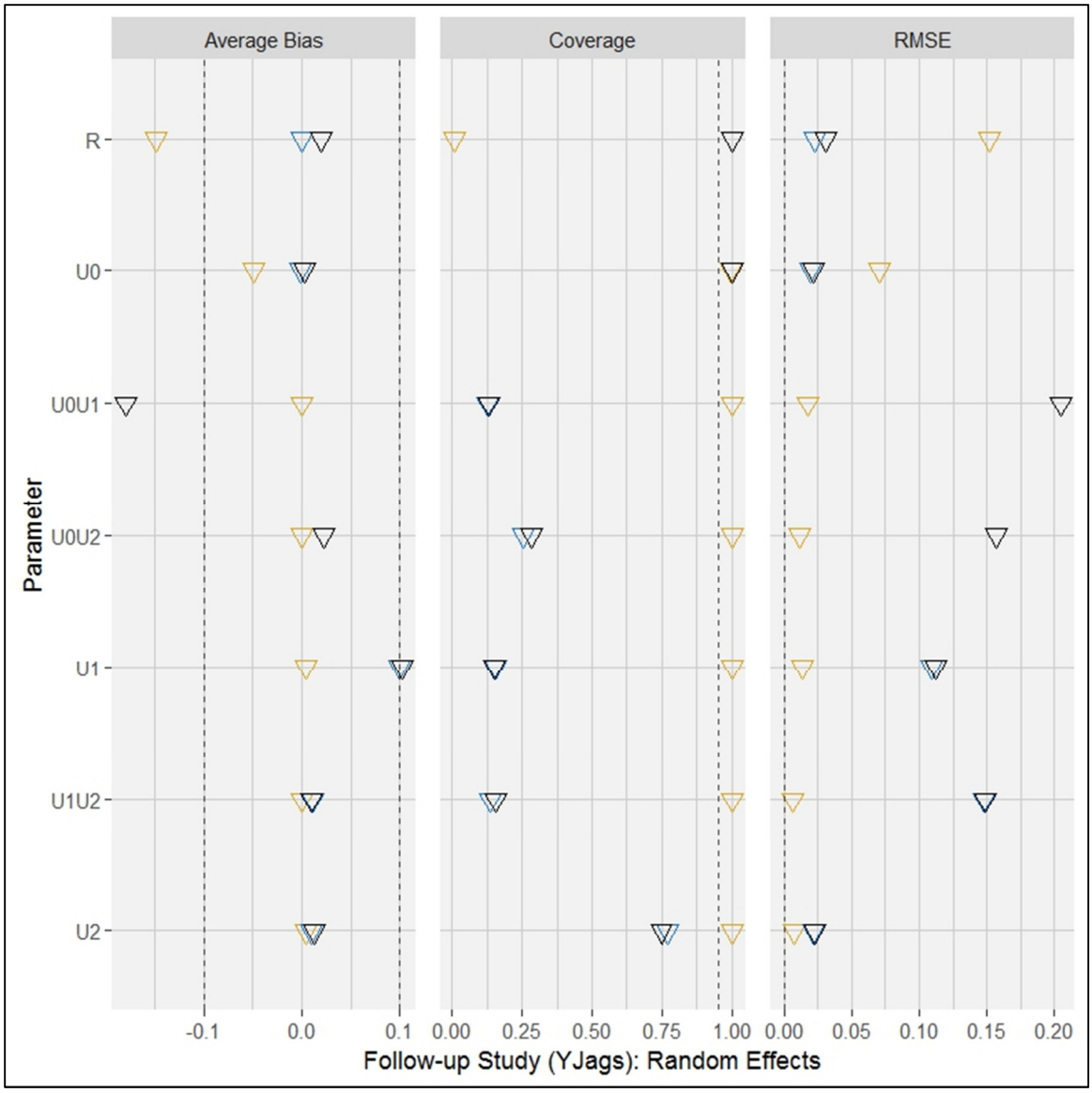


Figure 4.12. Comparison of fixed effects results for Y missingness with data imputed/analyzed in R/JAGS, imputed/analyzed in BLImP/Mplus, and imputed in R/JAGS then analyzed in Mplus. Data imputed/analyzed in R/JAGS are shown in yellow, imputed/analyzed in BLImP/Mplus are shown in blue, and data imputed in R/JAGS and then analyzed in Mplus are shown in black. All data for this comparison had *ICC* = .3, *J* = 50, and 20% missingness. Missingness pattern is shown by each row of plots. Each parameter is found on a line of each plot row. *ICC* = .3 is shown by▼, *J* = 50, *MAR* = 20% is shown by▽.

Figure 4.13 shows the results for the fixed effects parameters. Looking at the estimates for $BIAS_{AVG}$, CI coverage, and RMSE shows a similar pattern to that observed in the condition with Y missingness discussed in the previous section: values of these

statistics tend to be more similar between BLImP/Mplus and YX2jags (shown in green and black respectively) than between either of these conditions and the R/JAGS condition (shown in red). The differences between BLImP/Mplus and YX2jags is somewhat more disparate here than was observed in Figure 4.11 for the Yjags condition but still tend to be closer to each other than each set is to values from R/JAGS.

Random effects YX2jags. Figure 4.14 shows the results of this comparison across the three versions of the condition with $ICC = .1$, $J = 25$, and $MAR = 40\%$. For most parameters, the patterns of similarity for values of $BIAS_{AVG}$, CI coverage, and RMSE across versions of this condition were the same as was observed for the random effects in the Yjags condition. Values for u_0 are more similar across the three versions of here in the YX2jags condition than was observed for u_0 in the Yjags condition (Figure 4.12). Again, R/JAGS tended to return values that were closer to the guideline values than did BLImP/Mplus or YX2jags (again, with the exception of r_{ij}).

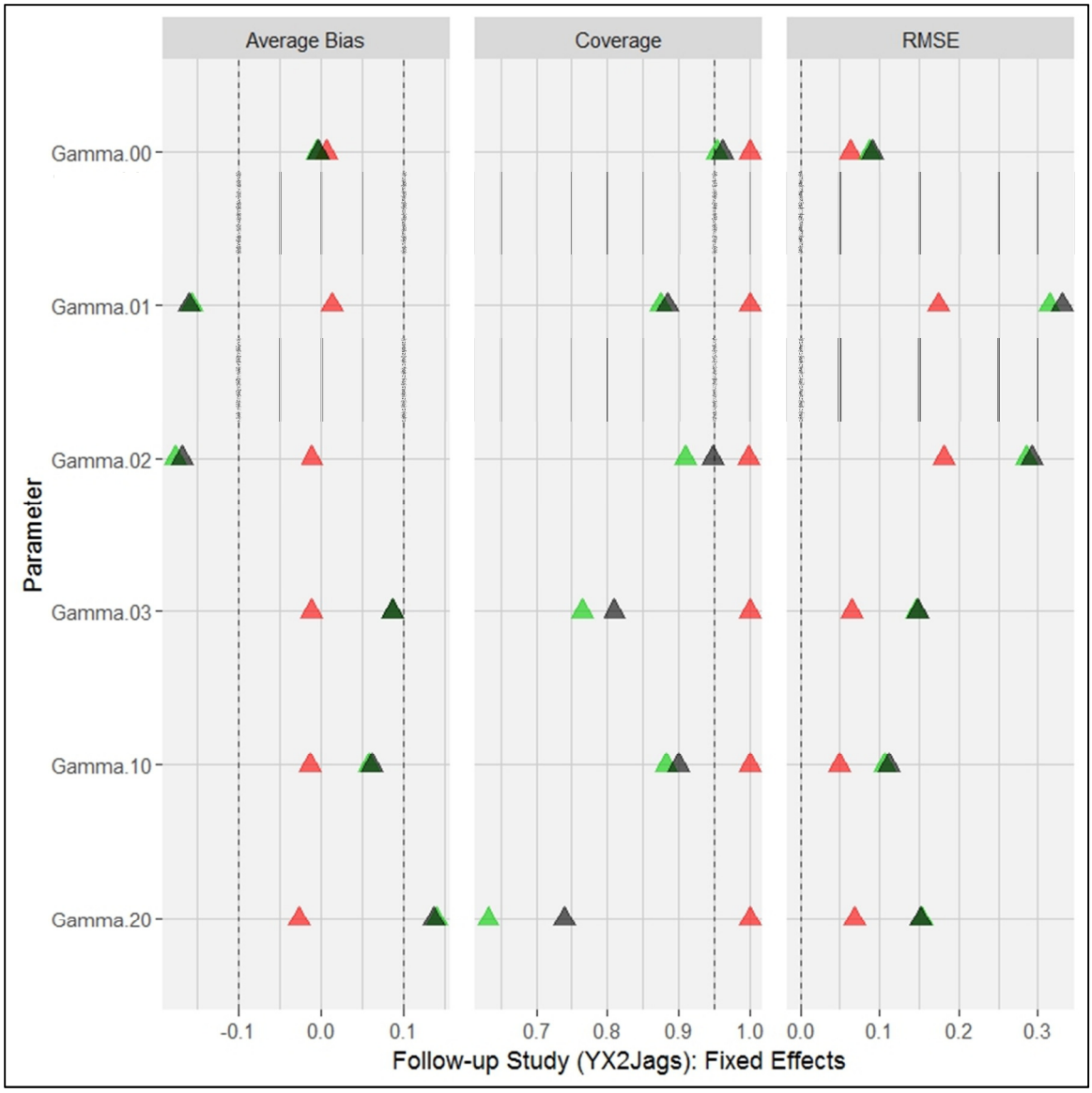


Figure 4.13. Comparison of fixed effects results for YX2 missingness with data imputed/analyzed in R/JAGS, imputed/analyzed in BLImP/Mplus, and imputed in R/JAGS then analyzed in Mplus. Data imputed/analyzed in R/JAGS are shown in red, imputed/analyzed in BLImP/Mplus are shown in green, and data imputed in R/JAGS and then analyzed in Mplus are shown in black. All data for this comparison had *ICC* = .1, *J* = 25, and 20% missingness. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *J* = 25, *MAR* = 40% is shown by▲.

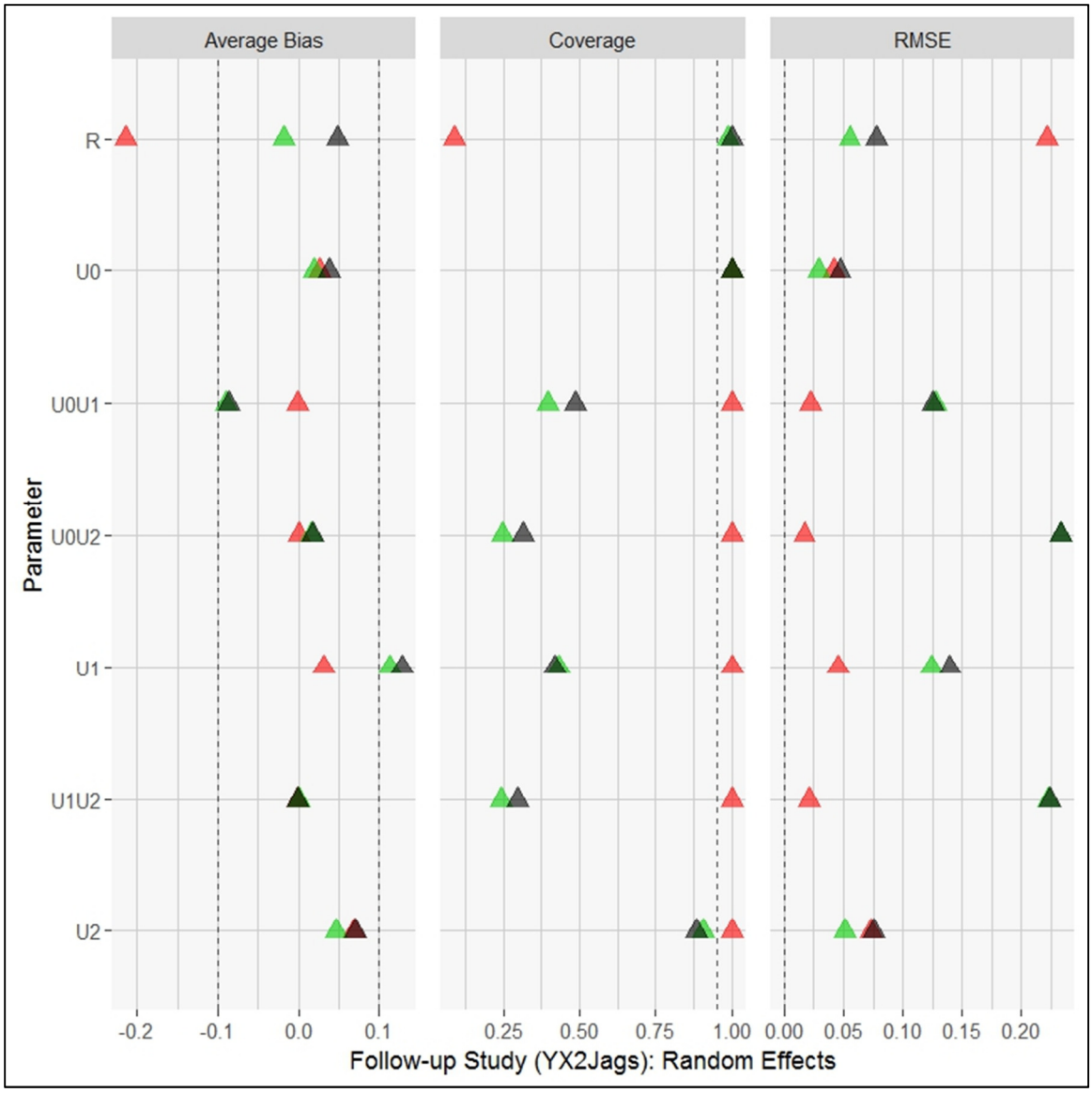


Figure 4.14. Comparison of random effects results for YX2 missingness with data imputed/analyzed in R/JAGS, imputed/analyzed in BLImP/Mplus, and imputed in R/JAGS then analyzed in Mplus. Data imputed/analyzed in R/JAGS are shown in red, imputed/analyzed in BLImP/Mplus are shown in green, and data imputed in R/JAGS and then analyzed in Mplus are shown in black. All data for this comparison had *ICC* = .1, *J* = 25, and 40% missingness. Each parameter is found on a line of each plot row. *ICC* = .1 is shown by▲, *J* = 25, *MAR* = 40% is shown by▲.

Chapter 5

DISCUSSION AND CONCLUSIONS

Three potential data characteristics which can influence the results and interpretation of an analysis are whether the data are clustered in nature, whether the data are categorical or continuous, and whether the data contain missing values. When analyzing multilevel data with the added complexity of missing data and/or categorical variables, the choice of modeling procedures for multiple imputation and analysis and pooling becomes increasingly important. Previous work explored multiple imputation procedures for two-level MLMs with missingness using joint modeling which groups data by the pattern of missingness and imputes missingness for each pattern (JM; Andridge, 2011; Black et al., 2011; Drechsler, 2015; Enders, 2011a; Resche-Rigon, White, Bartlett, Peters, & Thompson, 2013; Shin & Raudenbush, 2007, 2011; Yucel, 2008). This work with JM was followed by exploration of fully condition specification (FCS; Andridge, 2011; Liu, Taylor, & Berlin, 2000; van Buuren, 2011) methods of imputing missingness iteratively by imputing missingness on each variable individually.

Enders, et al. (in press) went on to compare results of imputation using JM and FCS (via custom SAS software) for continuous and categorical variables in a two-level MLM and found the results for FCS with multilevel categorical data with missingness to be promising. The work of Enders & Keller (in press) carried these efforts forward by further exploring the imputation of categorical variables modeled as normally distributed latent continuous variables using FCS and assessed the recovery of accurate imputations. Their custom software BLImP was developed to implement FCS under these conditions, and although some previous examination was conducted to investigate the accuracy of

imputation via BLImP, further work was needed to determine the influences of data characteristics and modeling choices in the recovery of accurate imputations.

The current work attempted to explore two frameworks for implementing MLMI with categorical and continuous data: a) a formally Bayesian approach with imputation, analysis, and pooling phases completed via Bayes estimation procedures and b) an approach which uses Bayesian estimation for the imputation phase only with the analysis and pooling phases completed from the frequentist perspective using common MLM software. The fully-Bayesian portion of this work was implemented in JAGS which is a popular software choice for researchers and data scientists working from a Bayesian perspective. The combined Bayesian estimation/frequentist analysis line of work was continued previous imputation efforts using BLImP for to impute the missingness and Mplus to estimate the MLM of interest.

Because the current work is couched in the field of educational measurement, some characteristics of the research were intended to be representative of applied educational research studies and differed from what had been most recently explored by Enders and Keller (in press). Primarily, these differences were in the rate of missingness with the possibility of 20% and 40% missing on each variable for which missingness was imposed, the number of clusters with small sets of 25 and 50 clusters being modeled, and with small ICC values of .1 and .3 Specifically, the choices for number of clusters and ICC values were intended to be representative of applied educational research studies which reported an average of between 20 and 28 students in each classroom (Black et al., 2011; Cosgrove & Cunningham, 2006; Du, 2009; Frempong, Reddy, & Kanjee, 2011; Goldstein et al., 2007; Guo & Zhao, 2000; Patrick, 2000; Muñoz & Chang, 2008) with

ICC values ranging from .00 to .7 (Andridge, 2011; Drechsler, 2015; Enders & Keller, in press; van Buuren, 2011; Zhao & Yucel, 2009).

Because the current endeavor is a simultaneous evaluation of these two analysis procedures, it the added benefit of providing a better understanding of the potential impact of the choice of analysis framework in recovery of accurate MLM parameter estimates. In other words, the choice to interpret results of the Bayesian imputation from BLImP within a frequentist analysis framework in Mplus may provide a different view of the results when compared to imputation, analysis, and interpretation within a fully Bayesian framework using JAGS. The results of this study show that this is indeed the case.

Continuous Data

First, considering the difference between analysis procedures for continuous data, results from the current work showed lower $BIAS_{AVG}$, higher CI coverage rates, and smaller RMSE values for R/JAGS than for BLImP/Mplus across all patterns of missingness, manipulated data characteristics, and for most parameters of interest. Regarding the fixed effects, R/JAGS returned more accurate parameter estimates than BLImP/Mplus. A few instances did occur in which BLImP/Mplus returned values of $BIAS_{AVG}$, CI coverage, or RMSE values that were similar to those of R/JAGS or were at least within the chosen guideline values. One example of this trend is the $BIAS_{AVG}$ values for γ_{00} with continuous data in which both R/JAGS and BLImP/Mplus analyses showed $BIAS_{AVG}$ which were at/near 0 (shown in Figure 4.2). For recovery of random effects, R/JAGS returned more accurate parameter estimates when the missingness

pattern was simple (e.g., missing on V3, X2, or X2V3), but as missingness extended to *Y*, R/JAGS parameter estimates for the variances σ_r^2, $\sigma_{u_0}^2$, and $\sigma_{u_2}^2$ were less accurate than estimates recovered for BLImP/Mplus (e.g., estimates of σ_r^2 for theYX2 missingness pattern were within the +/- .1 cutoff for BLImP/Mplus but below the -.1 cutoff for R/JAGS).

Looking at differences in BIAS_{AVG}, CI coverage, and RMSE for each of the manipulated data characteristics of ICC, *J*, and MAR rate with continuous data shows both similarities and differences in recovery of accurate estimates using R/JAGS and BLImP/Mplus. Examining BIAS_{AVG} for R/JAGS shows that for the fixed effects parameter, conditions with *ICC* = .3, *J* = 25, or *MAR* = 40% returned more extreme values than conditions with *ICC* = .1, *J* = 50, or *MAR* = 20%. For the random effects parameters (shown in Figure 4.2), patterns were similar across manipulated variables with the exception, again, of σ_r^2 which returned very extreme values of BIAS_{AVG} and some difference between *J* = 25 (shown in red) and *J* = 50 (shown in yellow).

Other than values observed for σ_r^2, all other BIAS_{AVG} were well below +/- .1 for R/JAGS. CI coverage rates for R/JAGS were similar (all at or near a value of 1) across both fixed and random effects parameters for all manipulated variables. RMSE values for R/JAGS showed high values for *ICC* = .3, *J* = 25, or *MAR* = 40% for fixed effects. For random effects, large RMSEs were most noticeable for conditions with *J* = 25. Some slight differences in RMSE were also observed for ICC with larger RMSE values found for γ_{00} with *ICC* = .3 and with complex missingness patterns of YX2 and YX2V3.

Results for BLImP/Mplus showed little differences between levels of *J* or MAR rates for $BIAS_{AVG}$ across most fixed and random effects. Differences by ICC were observed for γ_{03} and the covariance $u_{0j}u_{1j}$ with *ICC* = .3 showing more extreme values. CI coverage showed fixed effects results in which *ICC* = .3, *J* = 50, or *MAR* = 20% returned worse coverage rates than other conditions. For random effects, ICC values showed instances in which either ICC value resulted in worse coverage rates (i.e., no clear pattern for worse parameter recovery with one ICC value over the other), patterns of *J* = 50 again returned worse coverage, and MAR either showed no difference between levels or showed slightly worse coverage for *MAR* = 40%. RMSE values across ICCs for fixed effects were only consistently different for γ_{03} with *ICC* = .3 returning higher RMSEs than for *ICC* = .1. Differences in RMSE for levels of *J* showed that *J* = 25 returned high RMSE values. For random effects, these patterns of large RMSEs for *ICC* = .3 and *J* = 25 were again observed with a difference in pattern for MAR rate with slightly larger RMSEs for *MAR* = 20% than for *MAR* = 40%[27].

Categorical Data

The performance of both analysis procedures was noticeably worse with dichotomous data than with continuous data. With dichotomous data, convergence issues were observed in both R/JAGS and BLImP/Mplus which worsened as the ratio of categories became more extreme (i.e., more issues with convergence for .90/.10 split than for.50/.50 split). Although convergence issues with dichotomous data were observed for

[27] This issue of recovering less accurate parameter estimates will be discussed more fully in the later section on follow-up analyses.

both analysis procedures, these issues were much more widespread in R/JAGS and increased further with the complex missing data patterns. Estimation of models in R/JAGS which took only 500-1,500 cycles to adapt with continuous data were still unable to converge after 100,000 adaptation cycles.

Results for the dichotomous .50/.50 split conditions show additional differences between levels of the manipulated variables that were not observed with continuous data. This was true for both R/JAGS and BLImP/Mplus with more extreme values of $BIAS_{AVG}$, smaller CI coverage rates, and much larger RMSEs for dichotomous data. Looking at R/JAGS, values for $BIAS_{AVG}$ were still within +/- .1, and CI coverage rates remained near 1.0 for most parameters with dichotomous data (random effects for missingness on YX2 and YX2V3 were exceptions to this with extreme $BIAS_{AVG}$ and low CI coverage). RMSE values were much more extreme with dichotomous data and for random effects with the patterns of large values observed in simple missingness patterns becoming even more extreme for YX2 and YX2V3 missingness. Patterns of larger $BIAS_{AVG}$ values for $ICC = .3$, $J = 25$, or $MAR = 40\%$ were observed with dichotomous data as well.

Results for BLImP/Mplus show large $BIAS_{AVG}$ values for $ICC = .3$ and $J = 25$ for fixed effects γ_{01} and γ_{20} but large $BIAS_{AVG}$ for $ICC = .1$, $J = 25$, and $MAR = 40\%$ for γ_{02}. For random effects, $ICC = .3$, $J = 50$, and $MAR = 20\%$ showed large $BIAS_{AVG}$ values for $\sigma^2_{u_2}$ but large $BIAS_{AVG}$ were returned for $\text{cov}(u_{0j}u_{2j})$, $\text{cov}(u_{1j}u_{2j})$, and σ^2_r with $ICC = .3$, $J = 25$, and $MAR = 40\%$. Looking at CI coverage rates again shows similar patterns with lower CI coverage for $J = 50$ and $MAR = 20\%$ for most fixed and random effects parameters. Differences in ICC show low CI coverage rates for $ICC = .1$ which differs

from continuous data cases in which *ICC* = .3 returned many instances of low CI coverage. RMSE values for dichotomous data were again much more extreme than those observed for continuous data with larger RMSEs returned with *ICC* = .3, *J* = 25, or *MAR* = 40% for both fixed and random effects.

Follow-up Analyses

V3 Missingness Pattern

The observed results of extreme BIAS_{AVG}, low CI coverage, and large RMSEs for the missingness rate of *MAR* = 20% with BLImP/Mplus was unexpected given previous MLMI research. Past research has shown either a minimal effect of missingness rate (Zhao & Yucel, 2009) or showed that higher missingness rates produced less accurate parameter estimates (Enders & Keller, in press). Given the pattern observed in the current study, further investigation was conducted to assess whether the method of imposing missingness was influencing the recovery of accurate parameter estimates.

The follow-up study for the V3 missingness pattern compared results from BLImP/Mplus for additional lower missingness rates imposed on the same exact dataset to the original conditions with 40% and 20% missingness in which the two missingness rates were imposed on different datasets (with similar data characteristics) in a given replication. This comparison between missingness rates on the same exact dataset in Figures 4.9 and 4.10 showed little difference in values for most parameters in the MLM with 40% missingness (shown in solid green), 20% with values added back in to the 40% missing datasets (shown in purple), and 10% missingness rates (shown in pink) with additional values again added back in to the dataset to create a lower rate of missingness.

Some slight differences were observed for γ_{03} in which lower missingness rates returned larger BIAS_{AVG}, lower CI coverage, and larger RMSEs and cov($u_{0j}u_{2j}$) which returned larger RMSEs. In the remaining parameters, the patterns of lower missingness rates and similar or better results than were observed for the 40% missingness rate and were, therefore, more like the patterns expected from previous research. The results of this study showed that the patterns observed with the lower missingness rates were related to the way in which missingness was imposed and not indicative of an issue in the data. To explore another possible factor that might be contributing to this pattern of worse parameter recover with lower missingness rates in BLImP/Mplus, the final two follow-up studies were explored.

Yjags and YX2jags

To further investigate both this finding for missingness rates and to confirm that BLImP and JAGS were producing comparable imputed values, the Yjags and YX2jags conditions were examined. These conditions compared results from continuous R/Jags and BLImP/Mplus conditions to data which were imputed in R/JAGS and analyzed in Mplus. The choice of these two conditions provided a snapshot of conditions that were different in complexity with a condition with a simple missingness pattern, lower MAR rates, higher ICC values, and a larger number of clusters (Yjags) and a condition with a complex missingness pattern, higher MAR rates, lower ICC values, and fewer clusters (YX2jags).

As the results for these conditions show, parameter estimates for Yjags/YX2jags were more similar to results from BLImP/Mplus than they were to results from R/JAGS even though the imputed datasets were created in R/JAGS. The similar values from

Mplus highlight the influence of the analysis modeling choices in the recovery of accurate parameter estimates and the comparability between imputations from JAGS/BLImP. However, the differences observed between R/JAGS and Yjags/YX2jags estimates indicate that the analysis phase procedures in JAGS and Mplus do, in fact, influence the accuracy of the parameter estimates coming from the analysis of interest. This finding may be quite informative for MLM researchers working in MLMI within a frequentist framework.

Implications

The current study has many implications for MLMI research, and these implications differ in importance based on the goals of a particular research endeavor and the guiding framework to which a researcher subscribes. If a researcher is working from a fully-Bayesian perspective and her MLM analysis of interest contains only continuous data and does not specifically focus on values of the L1 variance σ_r^2, then R/JAGS is the preferred imputation and analysis method. This is due to the very minimal $BIAS_{AVG}$ values, very high CI coverage rates, and low RMSE values found with R/JAGS. In the current work, this pattern of accurate parameter recovery for R/JAGS was observed regardless of ICC value, number of clusters, missingness rate, or missingness pattern for continuous data.

If the researcher is working from a frequentist perspective with continuous data, then imputation in BLImP would be a recommended option, especially if the focus of the research is to assess differences in σ_r^2. However, the choice of a frequentist analysis software other than Mplus may be preferred due to the influence that Mplus showed on

the accuracy of parameter estimates recovered here with a two-level MLM with random intercepts and slopes for imputed datasets.

If a researcher is working from a Bayesian perspective and is working with dichotomous data with a simple missingness pattern such as missingness on predictors only at L1, L2, or L1 and L2, then R/JAGS would be still be the recommended imputation and analysis software as it continues to perform consistently and returns accurate parameter estimates for both fixed and random effects under these patterns of missingness with dichotomous .50/.50 split data. If data are dichotomous and missingness patterns are more complex containing missingness on both predictors and the outcome, then R/JAGS cannot be trusted to return accurate parameter estimates.

This is an area in which improvement must be made either to the current estimation power of JAGS and/or additional work must be done to allow for further means of modeling dichotomous data in JAGS. Other Bayesian estimation software has been recommended for MLM such as STAN (Carpenter, et al., 2016); however, this software also has limitations with categorical data and/or missing data. Further work is needed to create or implement reliable modeling techniques for MLMI with dichotomous data.

Similarly, the standard frequentist approach to MLMI with Bayesian imputation and frequentist analysis models also produced inaccurate parameter estimates in the presence of data with missingness across missingness patterns. Although previous work such as Enders and Keller (in press) have had much success with MLMI procedures via BLImP for dichotomous data, the current work emphasizes that the choice of analysis

software has a meaningful influence on the recovery of parameter estimates which is heightened with dichotomous data and complex missingness patterns as shown by the three follow-up studies. Even though R/JAGS and BLImP seem to produce similar imputations, the differences in recovery of accurate parameter estimates between R/JAGS and Yjags/YX2jags indicate that analysis of imputed data in Mplus may negatively influence the recovery of accurate parameter estimates. If the frequentist analysis model in Mplus was not influencing the recovery of parameter estimates, then we would expect the parameter estimates from the Yjags/YX2jags conditions to be more closely aligned with values observed for R/JAGS than was observed. Some other potential causes for these results include the choice of algorithm in the BLImP software which differed across continuous and categorical data and the loss of information between analysis of several thousand samples in the R/JAGS procedure and analysis of only 20 sampled JAGS datasets in Mplus in the Yjags and YX2jags conditions.

Researchers working from a frequentist framework with a focus on MLMI should explore whether other analysis routines for MLM influence the recovery of accurate parameter estimates in the same manner. By imputing missingness in BLImP, the analysis can be conducted via another frequentist MLM program. One such analysis tool to explore is the lme4 package in R (Bates, et al., 2016), although many other frequentist MLM software programs are available.

Conclusions

The results of this study show R/JAGS as a promising method for MLMI with continuous data. For the chosen manipulated conditions in this study, neither program performed well with extreme splits of .90/10 dichotomous data. For the .70/.30 split,

.90/.10 split, and more complex missingness patterns with .50/.50 split, R/JAGS was unable to converge after a reasonable number of adaptation/burn-in iterations. This difficulty with convergence makes R/JAGS estimation unreasonable for the average MLM researcher with typical computing resources and dichotomous data. BLImP showed better performance at these extreme splits and may be more reasonable option for MLMI with dichotomous data. However, researchers may want to explore other options for analyzing imputed data from BLImP given the results from Yjags and YX2jags showed the influence of the analysis routine on recovery of accurate parameter estimates.

The current work provided an important comparison between MLMI within a formally Bayesian framework for imputation and analysis and common approach in the MLM field with Bayesian imputation procedures and frequentist analysis routine. The results of this study show that the fully Bayesian techniques may return better parameter estimates in many research scenarios. Further work is needed to refine software and modeling capabilities to handle fully Bayesian models for complex MLMs with dichotomous data and missingness. Additionally, MLMI using BLImP was shown to be a promising method for imputing dichotomous data; however, further study of the influence of analysis procedure is needed to detangle the influence of analysis procedure from imputation process in the recovery of accurate parameter estimates.

REFERENCES

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal, 53*(1), 57-74.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics, 38*(9), 1845-1865.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Boyle M. H., & Willms, J. D. (2001). Multilevel modelling of hierarchical data in developmental studies. *Journal of Child Psychology and Psychiatry, 42*(1), 141-162.

Carpenter, J. R., & Goldstein, H., (2005). Multiple imputation in MLwiN. *Multilevel Modeling Newsletter, 16*, 9-18.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software* (in press).

Clarke, P. (2008). Theory and methods: When can group level clustering be ignored? Mutilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health, 62*(8), 752-758.

Cosgrove, J. & Cunningham, R. (2006). A multilevel model of science achievement of Irish students participating in PISA 2006. *The Irish Journal of Education, 2011*(xxxix), 57-73.

Drechsler, J. (2015). Multiple imputation of multilevel missing data—Rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40*(1), 69-95.

Du, P. (2009). Factors influencing students' adaptability in school. *Chinese Education and Society, 41*(5), 21-35.

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

Enders, C. K., (2011a). Analyzing longitudinal data with missing values. *Rehabilitation Psychology, 56* (4), 267-288.

Enders, C. K., (2011b). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16.

Enders, C. K., & Keller, B. T. (in press). A fully conditional specification approach to multilevel imputation of categorical and continuous variables.

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

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245.

Frempong, G., Reddy, V., & Kanjee, A. (2011). Exploring equity and quality education in South Africa using multilevel models. *Compare: A Journal of Comparative and International Education, 41*(6), 819-835. doi: 10.1080/03057925.2011.607488

Garson, G. D. (2013). Fundamentals of hierarchical linear and multilevel modeling. In Garson, G. D., *Hierarchical Linear Modeling: Guide and Applications*. Thousand Oaks, CA: SAGE Publications, Inc.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2013). *Bayesian Data Analysis: 3rd Edition.* Boca Raton, FL: Taylor & Francis Group, LLC.

Goldstein, H. (2011a). Missing data, partially observed data and multiple imputation. In Goldstein, H., *Multilevel Statistical Models:4th Edition* (chapter 16). Great Britain: John Wiley & Sons, Ltd.

Goldstein, H. (2011b). Multilevel models for discrete response data. In Goldstein, H., *Multilevel Statistical Models:4th Edition* (chapter 4). Great Britain: John Wiley & Sons, Ltd.

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.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*(1), 1-25. doi: 10.1080/00273171.2012.640589

Guo, G. & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology, 26*(2000), 441-462.

Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In de Leeuw , J. & Meijer, E., *Handbook of Multilevel Analysis* (Chapter 6). New York: Springer Science + Business Media, LLC.

Hedeker, D., & Mermelstein, R. J., (2011). Multilevel analysis of ordinal outcomes related to survival data. In Hox, J. J., & Roberts, J. K. *Handbook of Advanced Multilevel Analysis* (chapter 7). New York: Taylor & Francis Group, LLC.

Hung, L. (2010). The multigroup multilevel categorical latent growth curve models. *Multivariate Behavioral Research, 45*(2), 359-392.

Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2013). Direct likelihood analysis and multiple imputation for missing item scores in multilevel cross-classification educational data. *Applied Psychological Measurement, 38*(1), 61-80.

Lee, V.E., & Bryk, A.S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*(3), 172-192.

Little, R.J., & Rubin, D.B. (2002). *Statistical analysis of missing data* (Second edition). New York, NY: Wiley-Interscience.

Liu, M., Taylor, J. M. G., & Berlin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics, 56*, 1157-1163.

Maier, K. S., (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 27*(3), 271-289.

Meng, H. (2007). A comparison study of IRT calibration methods for mixed-format tests in vertical scaling. (Doctoral dissertation, University of Iowa). Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1523&context=etd.

Muñoz, M. A., & Chang, F. C. (2008). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education, 20*, 147-164.

National Assessment of Educational Progress. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics

Patrick, W. J. (2001). Estimating first-year student attrition rates: An application of multilevel modeling using categorical variables. *Research in Higher Education, 42*(2), 151-170.

Powers, D. A. (2012). Multilevel models for binary data. In Lott II, J. L., & Antony, J. S., *Multilevel Modeling Techniques and Applications in Institutional Research* (chapter 4). San Francisco: Wiley Periodicals, Inc.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Application and data analysis methods*. Thousand Oaks, CA: Sage Publications, Inc.

Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A. E., & Thompson, S. G. (2013). Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine, 32*, 4890-4905.
Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D.B., (1987). *Multiple imputation for nonresponse in surveys.* New York City: John Wiley & Sons.

Schafer, J. L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

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.

Serban, N., Staicu, A., & Carroll, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics, 69*, 902-913.

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

Shin, Y. & Raudenbush, S. W. (2011). Efficient analysis of Q-level nested hierarchical general linear models with missing data. *International Journal of Biostatics, 9*(1). doi:10.1515/ijb-2012-0048

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science, 46*(1), 218-237.

Teachman, J. (2011). Modeling repeatable events using discrete-time data: Predicting marital dissolution. *Journal of Marriage and Family, 73*(June 2011), 525-540.

van Buuren S (2011). Multiple imputation of multilevel data. In *The Handbook of Advanced Multilevel Analysis(10),* (Eds, Hox J, J. & Roberts J, K.) Routledge, Milton Park, UK, 173-196.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with inferable non-response. *Philosophical Transactions of the Royal Society, A*, 1-16.

Zhao, E., & Yucel, R. M. (2009). Performance of sequential imputation method in multilevel applications. In American Statistical Association Proceedings of the Survey Research Methods Section. American Statistical Association, Alexandria, VA, pp. 2800–2810.

APPENDIX A

YX2V3 BLIMP MODEL CODE WITH CONTINUOUS DATA

DATA: /u/home/k/klkunze/orpheus/dat/ic1j1nj1mi1rep430.dat;
VARNAMES: id a1 a2 a3 y x1 x2 v3 ym xm2 vm3 y_5050 x2_5050 v3_5050 y_5050m x2_5050m v3_5050m y_7030 x2_7030 v3_7030 y_7030m x2_7030m v3_7030m y_9010 x2_9010 v3_9010 y_9010m x2_9010m v3_9010m;

MODEL: id ~ ym:x1:xm2 a1 a2 a3 vm3;
ORDINAL: ;
NOMINAL: ;
BURN: 1000;
THIN: 500;
NIMPS: 20;
MISSING: 999;
SEED: 90291;
OUTFILE: /u/home/k/klkunze/orpheus/temp;
OPTIONS: separate hov;

APPENDIX B

YX2V3 BLIMP MODEL CODE WITH DICHOTOMOUS DATA

DATA: /u/home/k/klkunze/orpheus/ic1j1nj1mi1rep430.dat;
VARNAMES: id a1 a2 a3 y x1 x2 v3 ym xm2 vm3 y_5050 x2_5050 v3_5050 y_5050m x2_5050m v3_5050m y_7030 x2_7030 v3_7030 y_7030m x2_7030m v3_7030m y_9010 x2_9010 v3_9010 y_9010m x2_9010m v3_9010m;

MODEL: id ~ y_5050m:x1:x2_5050m a1 a2 a3 v3_5050m;
ORDINAL: y_5050m x2_5050m v3_5050m;
NOMINAL: ;
BURN: 1000;
THIN: 500;
NIMPS: 20;
MISSING: 999;
SEED: 90291;
OUTFILE: /u/home/k/klkunze/orpheus/temp;
OPTIONS: separate hov;

APPENDIX C

YX2V3 MPLUS MODEL CODE WITH CONTINUOUS DATA

```
DATA: file = ${2};
type = imputation;

VARIABLE:
names = id a1 a2 a3 y x1 x2 v3 ym xm2 vm3 y_5050 x2_5050 v3_5050 y_5050m
x2_5050m v3_5050m y_7030 x2_7030 v3_7030 y_7030m x2_7030m v3_7030m y_9010
x2_9010 v3_9010 y_9010m x2_9010m v3_9010m;
usevariables = id ym x1 xm2 vm3 w1 w2;
cluster = id;
within = x1 xm2;
between = w1 w2 vm3;

DEFINE:
w1 = cluster_mean(x1);
w2 = cluster_mean(xm2);

ANALYSIS:
type = twolevel random;

MODEL:
%within%
b1x1 | ym on x1;
b2x2| ym on xm2;
%between%
ym on w1 w2 vm3;
ym; b1x1; b2x2; ym with b1x1; ym with b2x2; b1x1 with b2x2;"
[b1x1];
[b2x2];
savedata:
results = ${1};
tech3 = ${1}_covB.dat;
```

APPENDIX D

YX2V3 MPLUS MODEL CODE WITH DICHOTOMOUS DATA

```
DATA:
file = ${2};
type = imputation;

VARIABLE:
names = id a1 a2 a3 y x1 x2 v3 ym xm2 vm3 y_5050 x2_5050 v3_5050 y_5050m
x2_5050m v3_5050m y_7030 x2_7030 v3_7030 y_7030m x2_7030m v3_7030m y_9010
x2_9010 v3_9010 y_9010m x2_9010m v3_9010m;
usevariables = id y_5050m x1 x2_5050m v3_5050m w1 w2;
categorical = y_5050m;
cluster = id;
within = x1 x2_5050m;
between = w1 w2 v3_5050m;

DEFINE:
w1 = cluster_mean(x1);
w2 = cluster_mean(x2_5050m);

ANALYSIS:
type = twolevel random;
algorithm = integration;
integration = 10;

MODEL:
%within%
b1x1 | y_5050m on x1;
b2x2| y_5050m on x2_5050m;
%between%
y_5050m on w1 w2 v3_5050m;
y_5050m; b1x1; b2x2; y_5050m with b1x1; y_5050m with b2x2; b1x1 with b2x2;
[b1x1];
[b2x2];
savedata:
results = ${1};
tech3 =  ${1}_covB.dat;
```

APPENDIX E

YX2V3 JAGS MODEL CODE WITH CONTINUOUS DATA

```
#############################################################################
#JAGS model for continuous data YX2V3 missingness
#############################################################################

model {
 for(i in 1:N)
  {
    #use when X.2 contains missingness
    X.2[i] ~ dnorm(mu.X.2[i], tau.X.2)
    mu.X.2[i] <- eta0[s[i]] + eta1[s[i]]*X.1[i] + eta2[s[i]]*a2[i] + eta3[s[i]]*a1[i]

#############################################################################
# Distribution of the data at Level 1
# Regression model
# s represents the group/school level
# b represents betas: e.g., b[s[i],1] is beta.00 for student i in school s as it
#is the first beta element in the matrix of betas
#############################################################################
    Y[i] ~ dnorm(mu.Y[i], tau.r)
    mu.Y[i] <-b0[s[i]] + b1[s[i]]*X.1[i] + b2[s[i]]*X.2[i] + b3[s[i]]*a1[i] + b4[s[i]]*a2[i]

}  #closes loop over N

#############################################################################
# Prior specifications for Level 1 parameters
#############################################################################
#L1 prior for Y
tau.r ~ dgamma(alpha.r, beta.r)
#prior for X.2 when missingness is present on X.2
tau.X.2 ~ dgamma(alpha.X.2,beta.X.2)

#############################################################################
# Distribution of the data at Level 2
# Regression model
#############################################################################
for(j in 1:J) {
      b0[j] <- B[j,1]
      b1[j] <- B[j,2]
      b2[j] <- B[j,3]
      b3[j] <- B[j,4] #when Y is missing
      b4[j] <- B[j,5] #when Y and X2 are missing

      B[j,1:5] ~ dmnorm (B.hat[j,], Tau.B[,]) #when Y and X2 are missing
         B.hat[j,1] <- Gamma.00 + Gamma.01 * W.1[j] + Gamma.02 * W.2[j] +
         Gamma.03 * V.3[j] + Gamma.04 * A.3[j]
```

```
        B.hat[j,2] <- mu.b1
        B.hat[j,3] <- mu.b2
        B.hat[j,4] <- mu.a1 #include when Y is missing
        B.hat[j,5] <- mu.a2 #include when Y and X2 are missing

        #format Ws for JAGS
        startplace[j] <-((j-1)*nperclust+1)
        endplace[j] <- ((j-1)*nperclust+nperclust)
        W.1[j] <- mean(X.1[(startplace[j]:endplace[j])])
        W.2[j] <- mean(X.2[(startplace[j]:endplace[j])])

        #for missingness on X.2
        eta0[j] <- E[j,1]
        eta1[j] <- E[j,2]
        eta2[j] <- E[j,3] #with missingness on X2only
        eta3[j] <- E[j,4] #with missingness on X2 and Y

        E[j,1:4] ~ dmnorm (E.hat[j,], Tau.E[,]) #for misssingness on both Y and X2
        E.hat[j,1] <- Omega.00 + Omega.04 * A.3[j] #when missingness is on X2 and V
        E.hat[j,2] <- Omega.01 * W.1[j]
        E.hat[j,3] <- Omega.02
        E.hat[j,4] <- Omega.03  #with missingess on Y and X2

        #for missing on V3
        V.3[j] ~ dnorm(mu.V.3[j], tau.V.3)
        mu.V.3[j] <- alpha.0 + alpha.1 * A.3[j]

} #close loop over J

############################################################################
# Prior specifications for Level 2 parameters
############################################################################
#priors for means of randomly varying slopes
    mu.b1 ~ dnorm(0, .0001)
    mu.b2 ~ dnorm(0, .0001)
    mu.a1 ~ dnorm(0, .0001) #for missingness on Y
    mu.a2 ~ dnorm(0, .0001) #for misssingness on X2 and Y

#prior for Tau matrix on B when Y and X2 are missing (dims:5x5)
    Tau.B[1:5,1:5] ~dwish(R.B[,],5)

#prior for Tau matrix on E for missingness on X2 and Y
    Tau.E[1:4,1:4] ~dwish(R.E[,],4)
```

```
#priors for gammas
      Gamma.00 ~ dnorm(0, .0001)
      Gamma.01 ~ dnorm(0, .0001)
      Gamma.02 ~ dnorm(0, .0001)
      Gamma.03 ~ dnorm(0, .0001)
      Gamma.04 ~ dnorm(0, .0001)

#priors for E.hat matrix for X2 missingness
    Omega.00 ~ dnorm(0, .0001)
    Omega.01 ~ dnorm(0, .0001)
    Omega.02 ~ dnorm(0, .0001)
    Omega.03 ~ dnorm(0, .0001)
    Omega.04 ~ dnorm(0, .0001) #include when V3 is also missing

#priors for V3 with missingness on V.3
    alpha.0 ~ dnorm(0, .0001)
    alpha.1 ~ dnorm(0, .0001)
    tau.V.3 ~ dgamma(alpha.V.3, beta.V.3)

#############################################################################
#Compute variances and standard deviations from precisons
#############################################################################
#Level1 variances
    sigma.squared.tau.r <- 1/tau.r
    sigma.r <- sqrt(sigma.squared.tau.r)
#Level2 variance when Y and X2 are missing
    sigma.squared.B[1:5,1:5] <- inverse(Tau.B[,])

#when missingness is present on X.2only
    sigma.squared.tau.X.2 <- 1/tau.X.2
    sigma.X.2 <- sqrt(sigma.squared.tau.X.2)

#when missingness is present on X.2 and Y
    sigma.squared.E[1:4,1:4] <- inverse(Tau.E[,])



#when V.3 is missing
    sigma.squared.tau.V.3 <- 1/tau.V.3
    sigma.V.3 <- sqrt(sigma.squared.tau.V.3)

}  #closes loop over model
```

APPENDIX F

YX2V3 JAGS MODEL CODE WITH DICHOTMOUS DATA

```
#############################################################################
#JAGS model for continuous data YX2V3 missingness
#############################################################################
model {
 for(i in 1:N)
  {
      #for missing on Y and X2 at L1
      X.2[i] ~ dbern(P.X[i])
      P.X[i] <-phi(eta0[s[i]] + eta1[s[i]]*X.1[i] + eta2[s[i]]*a2[i] + eta3[s[i]]*a1[i])

#############################################################################
# Distribution of the data at Level 1
# Regression model
# s represents the group/school level
# b represents betas: e.g., b[s[i],1] is beta.00 for student i in school s as it
#is the first beta element in the matrix of betas
#############################################################################
      #for missing on Y
      Y[i] ~ dbern(P.Y[i])
      #adds a1 and a2 when Y is missing
      P.Y[i] <-phi(b0[s[i]] + b1[s[i]]*X.1[i] + b2[s[i]]*X.2[i] + b3[s[i]]*a1[i] +
      b4[s[i]]*a2[i])

 }  #closes loop over N

#############################################################################
# Prior specifications for Level 1 parameters
#############################################################################
      #L1 prior for Y
      tau.r ~ dgamma(alpha.r, beta.r)
      #prior for X.2 when missingness is present on X.2
      tau.X.2 ~ dgamma(alpha.X.2,beta.X.2)









#############################################################################
# Distribution of the data at Level 2
# Regression model
# g represents gamma elements in the regression model
```

```
# K is the number of beta coefficients (K=3)
###############################################################################
for (j in 1:J) {

    b0[j] <- B[j,1]
    b1[j] <- B[j,2]
    b2[j] <- B[j,3]
    b3[j] <- B[j,4] #when Y is missing
    b4[j] <- B[j,5] #when Y and X2 are missing

    B[j,1:5] ~ dmnorm (B.hat[j,], Tau.B[,]) #when Y and X2 are missing, dims:5x5
    B.hat[j,1] <- Gamma.00 + Gamma.01 * W.1[j] + Gamma.02 * W.2[j] + Gamma.03 *
    V.3[j] + Gamma.04 * A.3[j]
    B.hat[j,2] <- mu.b1
    B.hat[j,3] <- mu.b2
    B.hat[j,4] <- mu.a1 #include when Y is missing
    B.hat[j,5] <- mu.a2 #include when Y and X2 are missing


    #format Ws for JAGS
    startplace[j] <-((j-1)*nperclust+1)
    endplace[j] <- ((j-1)*nperclust+nperclust)
    W.1[j] <- mean(X.1[(startplace[j]:endplace[j])])
    W.2[j] <- mean(X.2[(startplace[j]:endplace[j])])


    #for missingness on X.2
    eta0[j] <- E[j,1]
    eta1[j] <- E[j,2]
    eta2[j] <- E[j,3] #with missingness on X2only
    eta3[j] <- E[j,4] #with missingness on X2 and Y


    E[j,1:4] ~ dmnorm (E.hat[j,], Tau.E[,]) #for misssingness on both Y and X2
    E.hat[j,1] <- Omega.00 + Omega.04 * A.3[j] #when missingness is on X2 and V
    E.hat[j,2] <- Omega.01 * W.1[j]
    E.hat[j,3] <- Omega.02
    E.hat[j,4] <- Omega.03  #with missingess on Y and X2



    #for missing on V3 only
    V.3[j] ~ dbern(P.V3[j])
    P.V3[j] <-phi(alpha.0 + alpha.1 * A.3[j])
```

```
} #close loop over J

#############################################################################
# Prior specifications for Level 2 parameters
#############################################################################
      #priors for means of randomly varying slopes
      mu.b1 ~ dnorm(0, .0001)
      mu.b2 ~ dnorm(0, .0001)
      mu.a1 ~ dnorm(0, .0001) #for missingness on Y
      mu.a2 ~ dnorm(0, .0001) #for misssingness on X2 and Y

      #prior for Tau matrix on B when Y and X2 are missing (dims:5x5)
      Tau.B[1:5,1:5] ~dwish(R.B[,],5)
      #prior for Tau matrix on E for missingness on X2 and Y
      Tau.E[1:4,1:4] ~dwish(R.E[,],4)

      #priors for gammas
      Gamma.00 ~ dnorm(0, .0001)
      Gamma.01 ~ dnorm(0, .0001)
      Gamma.02 ~ dnorm(0, .0001)
      Gamma.03 ~ dnorm(0, .0001)
      Gamma.04 ~ dnorm(0, .0001)

      #priors for E.hat matrix for X2 missingness
      Omega.00 ~ dnorm(0, .0001)
      Omega.01 ~ dnorm(0, .0001)
      Omega.02 ~ dnorm(0, .0001)
      Omega.03 ~ dnorm(0, .0001)
      Omega.04 ~ dnorm(0, .0001) #include when V3 is also missing

      #priors for V3 with missingness on V.3
      alpha.0 ~ dnorm(0, .0001)
      alpha.1 ~ dnorm(0, .0001)

#############################################################################
# Compute variances and standard deviations from precisions
#############################################################################

      #Level1 variances
      #sigma.squared.tau.r <- 1/tau.r
      #sigma.r <- sqrt(sigma.squared.tau.r)
      #Level2 variance when Y and X2 are missing
      sigma.squared.B[1:5,1:5] <- inverse(Tau.B[,])
```

```
    ##when missingness is present on X.2only
    #sigma.squared.tau.X.2 <- 1/tau.X.2
    #sigma.X.2 <- sqrt(sigma.squared.tau.X.2)

    #when missingness is present on X.2 and Y
    sigma.squared.E[1:4,1:4] <- inverse(Tau.E[,])

    ##when V.3 is missing
    #sigma.squared.tau.V.3 <- 1/tau.V.3
    #sigma.V.3 <- sqrt(sigma.squared.tau.V.3)

}  #closes loop over model
```