12-14-2015

Exploring Listwise Deletion and Multilevel Multiple Imputation in Linear Two-Level Organizational Models

Whitney Flemming Smiley
University of South Carolina - Columbia

EXPLORING LISTWISE DELETION AND MULTILEVEL MULTIPLE IMPUTATION IN LINEAR TWO-LEVEL ORGANIZATIONAL MODELS

by

Whitney Flemming Smiley

Bachelor of Science
James Madison University, 2009

Master of Arts
James Madison University, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2015

Accepted by:

Bethany Bell, Major Professor

Robert Johnson, Major Professor

Christine DiStefano, Committee Member

Michael Seaman, Committee Member

Jason Schoeneberger, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

DEDICATION

This dissertation is dedicated to my father and mother, Robert Flemming Smiley and Janice Overton Smiley for their support and encouragement throughout this process.

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely thank my parents who have supported and encouraged me through numerous years of college. While you may not know it, this dissertation would have never happened without you. Additionally I would like to thank my aunt Edna Jane McDaniel, who listened to me whenever I needed it and supported me much more than she probably knows.

I would also like to acknowledge the Department of Educational Research and Measurement at the University of South Carolina, especially those members of my doctoral committee for their input, valuable discussion, and accessibility. In particular I would like to thank Dr. Bethany Bell for taking me on as a student and helping me accomplish my goals and Dr. Jason Schoeneberger, one of the best SAS programmers I know, for answering e-mails very quickly (typically same day) when SAS and I couldn't agree.

I am also very grateful to my colleagues at College Board, who put up with me using their computers for several months to simulate, impute, and pool several terabytes of data. I especially want to mention Haifa Matos-Elefonte and Pamela Kaliski Pfluger as they allowed me to interrupt them at a moment's notice to check on my conditions. I'm lucky to have colleagues like you.

ABSTRACT

Problems of missing data are pervasive in social science research. Because of this, researchers have begun to use techniques after data collection to deal with missing data, including traditional methods (i.e. listwise deletion, pairwise deletion, and single imputation procedures) and modern procedures (i.e. multiple imputation and full information maximum likelihood). In the past, several organizations and researchers have warned that traditional missing data techniques (MDTs) can introduce bias into parameter estimates, and can result in a loss of statistical power (e.g., Becker & Powers, 2001; Wilkinson & the APA Task Force on Statistical Inference, 1999). However, previous research has shown that using a traditional method does not necessarily reduce statistical power or bias parameter estimates (Roth & Swizer, 1995). Research using traditional regression techniques has shown that sample size, percent of missing data, and missing data mechanism are key characteristics in determining under what conditions each MDT should be used. To further complicate matters, the multilevel modeling (MLM) literature has largely ignored the impact of missing data. Thus, it is not known if the results from single-level missing data research apply to hierarchical data.

The present simulation study compare the performance of multilevel multiple imputation (MLMI) and listwise deletion in the context of linear two-level organizational models with continuous predictors. Design factors of interest included missing data technique, missing data mechanism, level-1 sample size, level-2 sample size, level-1 percent of missing data, and level-2 percent of missing data totaling

2,000 conditions. Design factors were evaluated on four outcomes, including bias, Type I error, statistical power, and confidence interval (C.I.) coverage. Results from this study showed that listwise deletion performed well for bias, level-2 Type I error rates, level-1 power, and C.I. coverage. Listwise deletion did have minor problems with Type I error rates at level-1, and power at level-2. MLMI performed well for level-2 Type I error rates, level-1 power, and level-2 C.I. coverage, but had minor issues with level-1 Type I error rates and level-2 power, and major issues with bias at both levels, and C.I. coverage at level-1. Recommendations for applied researchers based upon these results are discussed.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

Often in education, public health, sociology, and public policy, researchers encounter situations in which data are hierarchical or nested. For example, students can be nested within classrooms or schools and patients can be nested within hospitals. Because students (level-1) within a school (level-2) are more alike than students in different schools, these nested data structures need to be analyzed differently due to the lack of independence between the level-1 units. Specifically, when systematic differences exist between higher level units in data organized hierarchically, a correlation among lower-level units within each higher level unit can arise, thus violating the assumption of independence (Snijders & Bosker, 1999). Ignoring the nesting of data can impact estimated variances and the available power to detect treatment or covariate effects (Donner & Klar, 2000; Julian, 2001; Moerbeek, 2004; Murray, 1998; Shadish, Cook & Campbell, 2002), can inflate Type I error rates (Wampold & Serlin, 2000), and can lead to substantive errors in interpreting the results of statistical significance tests (Goldstein, 2003; Nich & Caroll, 1997).

Given the importance of accounting for dependencies in nested data, a class of models known broadly as multilevel models (MLM) has been developed. Conceptually, MLMs are an extension of regression where each level-2 unit has its own unique regression equation. Thus, instead of representing data with only one regression equation,

researchers can examine regression lines for each level-2 unit and how slopes and intercepts vary across the different units. This class of models requires additional assumptions, estimation of a greater number of parameters, and more complex specifications and model fit assessments than in traditional single-level methods (Goldstein, 1987; 2003; Longford, 1993; Raudenbush & Bryk, 1986; 2002; Snijders & Bosker, 1999).

Similar to traditional linear regression techniques, linear multilevel models can specify dependent variables as a function of a linear combination of both categorical and continuous variables. Estimates of the relationships between predictors and the dependent variable are termed fixed effects and are interpreted similarly to single level regression coefficients (i.e. for every one unit change in the predictor the predicted value of the dependent variable changes equal to the slope value associated with the predictor). However, unlike traditional regression techniques, multilevel models allow for specification of units at different levels. For example if children are nested within schools, multilevel modeling specifies children at level-1 and schools at level-2. This can also be extended to have more than two levels (e.g. children nested within schools nested within districts). Specification of these levels allows for decomposition of variance in outcomes and predictors to be within-unit (within each distinct higher level unit) and between-unit (across the distinct higher-level units). When predictors exist at the lowest and highest levels of the hierarchy, researchers use multilevel models to specify cross-level interactions to investigate how higher level predictors influence the estimates of lower-level predictors. Additionally, researchers can understand how the performances of lower level variables fluctuate among higher level units by modeling random effects.

Specifying a variable as having a random effect informs researchers how slopes and intercepts vary across the different units.

Figure 1.1 shows a conceptual illustration of random intercepts and slopes. Ignoring color, all of the data points are represented by the black regression line, which is an example of how these data would be represented using single-level regression. A MLM with random slopes and intercepts allows researchers to specify a unique group effect through random intercepts and random slopes, thus allowing each school to have its own regression line with its own intercept and own slope. Figure 1.1 shows three schools where school A is represented by the blue data points, school B the red data points, and school C the green data points. By allowing each school to have its own slope and intercept, we can see that School A, on average, has lower overall performance on the outcome, but the slope associated with the predictor is stronger compared to the other schools. School C, on the other hand, has the highest overall performance but the weakest slope. Lastly, School B has average achievement slightly lower than School C, but has a stronger slope, but not as strong as school A.

Estimates of variability among random intercepts and slopes are known as variance components, and provide a description of the distributions of the random intercept and slope effects, and their potential covariance (Agresti, Booth, Hobert & Caffo, 2000). The within-unit variance captures the error in prediction among units within a particular group (at level-1), where each level-1 unit deviates from the overall level-2 mean for the model. The between-unit variance captures the amount of variability between the higher-level units. In the case of a two-level model examining children nested within schools, the within-unit variance captures how much a child deviates from

the overall mean of all children in the same school. The between-unit variance captures the extent to which schools vary from each other. The ratio of the between-group variance to total variance (i.e. the combination of between-group and within-group variance) is known as the intra-class correlation coefficient (ICC). The ICC provides a metric for the degree of correlation or dependence among lower-level units (Bloom, 2005; Kreft & De Leeuw, 1998; Snijders & Bosker, 1999). The ICC is scaled 0 to 1 with larger values providing more evidence that multilevel modeling should be used.

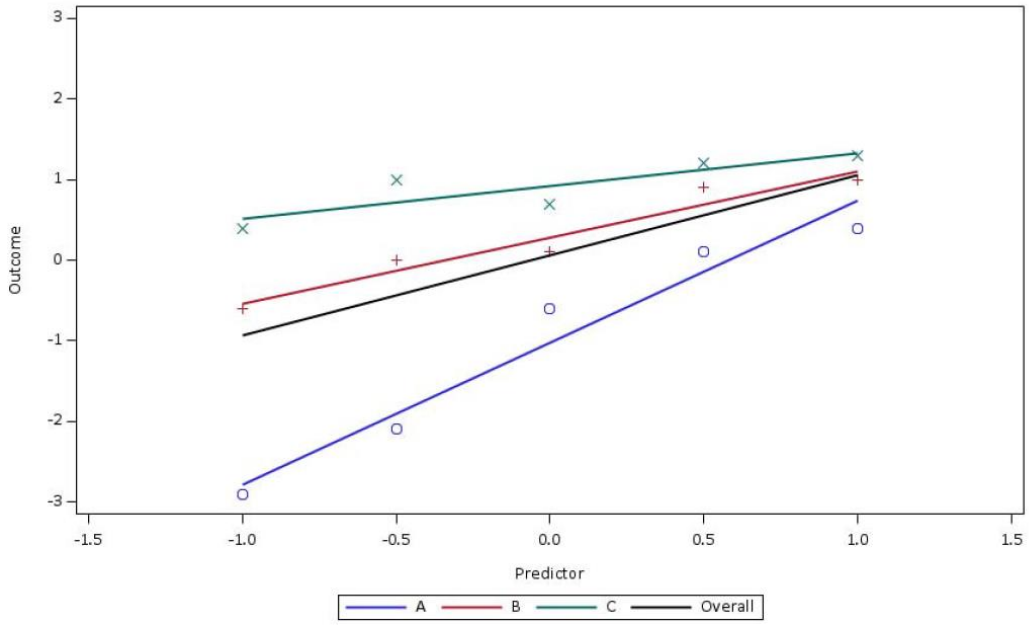

Figure 1.1 Conceptual Illustration of Random Intercepts and Random Slopes.

Due to the fact that MLM involves an increased number of parameters that needs to be estimated, the model-building process can become complicated. Although there are many ways to build MLMs, here I provide a relatively simple and straightforward method to find the most parsimonious process to identify the best fitting model for a two-level model:

1. The null model is estimated in order to calculate the ICC. The null model is estimated by specifying the outcome of interest as the dependent variable and a random intercept with no predictors. The within-and between-group variance is then used to calculate the ICC (i.e. the ratio of the between-group variance to total variance).
2. The next step is to expand the null model by adding the level-1 fixed effects. In this step, the results indicate the relationship between the level-1 fixed effects and the outcome. Examination of the fit indices, including Akaike's Information Criteria and Bayesian Information Criterion in addition to the Likelihood Ratio Test, can inform the researcher if the expanded model fits better than the null model.
3. Once the expanded model has been estimated and deemed to be a better fitting model than the null model, the researcher can add random slopes for every level-1 predictor. This will expand the results from Step 2 and reveal if the relationships between level-1 predictors and the outcome vary between the level-2 units. From this point fit indices will then be explored to verify that the random slopes need to be included in the model.
4. Once the model in Step 3 is estimated, the level-2 fixed effects are added to the model. This will indicate the average relationship between level-2 predictors and the outcome, while controlling for level-1 predictors.

It is important to point out that this model building procedure is general in nature. Depending on the fit indices, intermediate models may need to be estimated (e.g. some level-1 slopes may vary across level-2 and some may not and intermediate models may be estimated to determine out which ones should be allowed to vary). Also note that the model building procedure becomes more complex when more levels are added and when interactions are present (i.e. more steps are needed).

Because multilevel models consist of both fixed and random effects, there are assumptions associated with both of these effects. The assumptions related to bias in the fixed effects are similar to assumptions associated with traditional regression models; however the assumptions occur at multiple levels. These assumptions are based on the relationship of fixed effects to residual terms in the model and include:

1. Level-1 predictors are independent of level-1 residuals.
2. Level-2 predictors are independent of every level-2 residual.
3. Predictors at each level are not correlated with residuals at other levels.

The assumptions related to the random effects determine the accuracy of hypothesis tests and confidence intervals and include:

4. Each level-1 residual is independent and normally distributed with a mean of 0 and variance σ^2 for every level-1 unit within each level-2 unit.
5. Level-2 errors are multivariate normal with a mean of 0 and a variance to be estimated and potential covariance among the level-2 errors. The level-2 random error vectors are also independent from each other.
6. Level-1 and level-2 errors are independent.

Assumption 4 is similar to traditional regression, however in multilevel modeling the residuals are independent within a higher-level unit instead of across the entire data set. Assumptions 5 and 6 are unique to multilevel modeling and are due to the nested data and specification of multiple levels and residuals associated with multiple levels.

Two-Level Linear Multilevel Models

Raudenbush and Bryk (2002) presented an example of a two-level linear model with random intercepts and slopes where students (level-1) are nested within schools (level-2) that models data collected from the 1982 High School and Beyond Survey (HSB). At level-1 a student outcome variable is regressed onto a single student-level predictor. In Raduenbush and Bryk's (2002) example, the outcome was score on a math assessment and the student-level predictor was student SES, represented by the following equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(SES)_{ij} + r_{ij} \qquad r_{ij} \sim N(0, \sigma^2)$$

where Y_{ij} represents the score on the math assessment for student i in school j, β_{0j} represents the average math achievement across all students, β_{1j} represents the relationship between student SES and math achievement, and r_{ij} represents an individual student's error term. Each school is allowed to have its own intercept and the error (r_{ij}) is approximately normal with a mean of 0 and a covariance R.

At level-2, parameters at level-1 may serve as an outcome variable for level-2 predictors. Raudenbush and Bryk (2002) show an example where two level-2 predictors are used to explain intercepts and slopes, by the following equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(MEAN\ SES)_j + \gamma_{02}(SECTOR)_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(MEAN\ SES)_j + \gamma_{12}(SECTOR)_j + u_{1j}$$

In this equation, SECTOR is a dummy coded variable where 1 represents private schools and 0 represents public schools and MEAN SES represents the average SES of students within a particular school. The equation for β_{0j} represents several main effects. Thus, γ_{00} represents the overall mean score for public schools, γ_{01} is the regression coefficient associated with school-level SES, and γ_{02} is the regression coefficient associated with private schools. The equation for β_{1j} represents cross-level interactions between level-1 and level-2 predictors, where γ_{10} represents the regression coefficient for SES, γ_{11} represents the regression coefficient for the interaction between MEAN SES and student - level SES, γ_{12} represents the regression coefficient for the interaction of SES and school SECTOR. How these variables interact can be easily seen when the level-1 and level-2 models are combined into the following equation:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01}(MEAN\ SES)_j + \gamma_{02}(SECTOR)_j + \gamma_{10}(SES)_{ij} \\ & + \gamma_{11}(MEAN\ SES)_j(SES)_{ij} + \gamma_{12}(SECTOR)_j(SES)_{ij} + u_{0j} \\ & + u_{1j}(SES)_{ij} + r_{ij} \end{aligned}$$

This equation is arranged such that the fixed effects (in blue) of the model are toward the front with the random part of the equation (in green) listed in the middle and the error term (in orange) toward the back. Y_{ij} represents the score on the math assessment for student *i* in school *j*, γ_{00} represents the overall mean score for public schools, γ_{01} is the regression coefficient associated with school-level SES, and γ_{02} is the regression coefficient associated with private schools. γ_{10} represents the regression coefficient for SES, γ_{11} represents the cross-level interaction between school-level SES and student SES,

γ_{12} represents the school-level cross-level interaction between sector and student SES. Thus the fixed effects portion of the model shows that this equation consists of three main effects and two cross-level interactions.

For the random part of the model, the estimates are called variance components. For example, multilevel models can have a random intercept (u_{0j}) to indicate that intercepts in the lower levels are allowed to vary among higher levels. Additionally, researchers can specify random slopes (u_{1j}) through theory or the model building process to indicate that slope coefficients for lower level variables fluctuate among higher level units. Note in these models level-1 effects are assumed to be unrelated to errors at level-2. The variance components (u_{0j} and u_{1j}) are assumed to be multivariate normally distributed with means of 0 and covariance matrix:

$$Var\begin{bmatrix}u_{0j}\\u_{1j}\end{bmatrix} = \begin{bmatrix}\tau_{00} & \\ \tau_{10} & \tau_{11}\end{bmatrix} = \boldsymbol{T}$$

where τ_{00} represents the population variance among the school means, τ_{11} is the population variance among slopes, and τ_{10} is the population covariance between intercepts and slopes.

As evidenced by Raudenbush and Bryk's two-level linear model with random intercepts and slopes, MLM is an extension of traditional regression techniques with a more complicated model building process. As such, MLMs are susceptible to many of the same complications as those introduced in traditional regression techniques, including missing data. Missing data can be even more detrimental in MLMs given the complexity of the model. That is, since data are nested, missingness can occur at both levels

influencing the results of the model. For example, if children are nested within a school, complete information has to be gathered from the child (e.g. student achievement) and from the school (e.g. percent of students receiving free or reduced lunch). Given that data are collected from both the students and the school it is more likely missing data will emerge, influencing the results of the model.

Statement of the problem

Problems of missing data are pervasive in empirical social science research. Researchers spend much time and effort devising sampling efforts that represent their population of interest. For example, suppose an applied researcher selects 1,000 people from their target population and is confident that the sampling technique used is representative of the population. If all data are present, researchers can feel fairly confident that his or her results are representative. However, if after sampling missing data are present on some observations and subsequently get deleted from the analyses, the researcher becomes less confident that his or her results actually represent the population of interest. Moreover, if observations with missing data are deleted from the sample, analyses are often conducted on smaller samples than the initial number of selected cases (Allison, 2002), thus decreasing statistical power.

There are many causes of missing data. Sometimes people refuse to answer questions in surveys, people overlook survey items, interviewers may neglect to ask some questions, respondents may indicate that they do not know the answer to some questions, people may drop out of research studies completed longitudinally, etc. Whatever the reason for missing data, such data are a notable problem since nearly all standard

statistical methods presume that information is available on every case for all variables included in the analyses. Because of this, investigators spend a lot of time, money, and effort in order to minimize the occurrence of incomplete data or non-response among respondents (Mason, 1999).

Even after rigorous methods for minimizing missing data are used, it is unlikely that researchers have complete data (Allison, 2002). Because of this, researchers have begun to use techniques after data collection to deal with missing data, including the more traditional methods (i.e. listwise deletion, pairwise deletion, and single imputation procedures) and two modern procedures that are typically recommended by researchers (i.e. multiple imputation and full information maximum likelihood [FIML]; Schafer & Graham, 2002). While there are pros and cons to both traditional and modern methods, this section provides a brief overview of the most common missing data techniques (MDTs) in the literature, whereas a more thorough explanation of each MDT is discussed in Chapter 2.

Listwise Deletion. With this method, any case that has missing data for any variable of interest in a given study is discarded from the analysis. The primary reason that this method is frequently used is due to convenience. This method is commonly available in software packages such as SAS and SPSS and has no subjective decision making required to implement the procedure. Ultimately, if data are incomplete, that case (or in most social science research, that person) is removed from the analysis.

Pairwise Deletion. This method is similar to listwise in that incomplete cases are removed from analyses, however with this method, researchers attempt to use as much

data as possible by eliminating cases on an analysis-by-analysis basis. Because of this, each analysis has a different sample size.

Single Imputation Methods. This is an umbrella term for procedures where one replacement value is filled in for each missing data point prior to analysis. Examples include mean imputation, regression imputation, stochastic regression imputation, and hot-deck imputation. This method results in all observations having complete data, thus, no one is removed from the analyses.

Multiple Imputation (MI). Multiple imputation (MI) fills in missing values prior to analysis. Conceptually this procedure involves three stages to generate a complete data set—the imputation phase, the analysis phase, and the pooling phase. The imputation phase creates multiple datasets (the number of which is set by the researcher) of the data which contains different estimates of the missing values using Bayesian estimation principles. The analysis phase analyzes the new, filled in data sets from the imputation phase, and the pooling phase combines everything into a single set of results (Enders, 2010; Rubin 1987).

Full-information Maximum Likelihood (FIML). Maximum likelihood estimation is an iterative process that repeatedly uses different values for population parameter values until it finds estimates that most likely produced the observed data, by repeating the log-likelihood computations for each of the different values (Enders, 2010). The goal with this procedure is to find the combination of estimates that best fits the data, yet does not actually fill in the missing values.

In 1999 The American Psychological Association Task Force on Statistical Inference explicitly warned against the use of traditional MDTs, such as listwise and pairwise deletion. They stated,

> Special issues arise in modeling when we have missing data. The two popular methods for dealing with missing data that are found in basic statistics packages—listwise and pairwise deletion of missing values—are among the worst methods available for practical applications. (Wilkinson & the APA Task Force on Statistical Inference, 1999, p.598)

This statement was published after several research studies showed that the use of traditional MDTs can introduce bias into parameter estimates derived from a statistical model (Becker & Powers, 2001; Becker & Walstaf, 1990; Rubin, 1987) and can result in a loss of information and statistical power (Anderson, Basilevsky, & Hum, 1983; Kim & Curry, 1977) when strict assumptions are not met.

Although The American Psychological Association Task Force on Statistical Inference explicitly warned against the use of traditional MDTs, previous research has shown that using a traditional method does not necessarily reduce statistical power or bias parameter estimates. Research using traditional regression techniques has shown that sample size and percent of missing data are key characteristics in determining under what conditions each MDT should be used. For example, researchers conducting simulation studies using traditional regression have come to the general consensus that pairwise and listwise deletion methods work well when sample sizes are large, and the percent of missingness is small (Basilevsky et al. 1985; Kim & Curry, 1977; Roth & Swizer, 1995; Witta, 1992). Thus, in certain situations traditional methods can be used despite the underlying missing data mechanism (i.e., the way that missing data is related to complete data).

Unlike what we know about missing data in ordinary least squares (OLS) regression, research on missing data in MLM is lacking and largely ignored in applied articles (van Buuren, 2011) and, thus, there is great need for research and proper dissemination in this area. Since MLM is an extension of traditional regression techniques, it is plausible that missing data treatments perform similarly with these models. That is, under most conditions researchers using MLM should use a modern technique to deal with data, but in certain situations traditional techniques could be appropriate. As noted earlier, MLM is more complicated than traditional regression because of the nesting of data and the presence of both fixed and random effects. It is especially important to examine the impact of MDTs with MLMs, as research studies with nested data are susceptible to missing data due to data being collected from different sources. Few researchers, however, have examined the impact of MDTs on MLMs.

Despite the fact that little is known about the conditions under which each MDT should or should not be used, listwise deletion has been the common method used in research that estimates MLMs (e.g. Archibald, 2005; Bosker, Kremers, & Lugthart, 1990; Desimore, 2005; Finn, Gerber, Achilles, & Boyd-Zaharias, 2001; Goddard, Goddard, & Tschannen-Moran, 2007; Hill & Rowe, 1996; Kyriakides, Campbell, & Gatsis, 2000; Lamb, 2002; Marks, 2000; Opdenakker, 2001; Xue, 2002), while few MLM studies have used data imputation methods (e.g. Correnti & Rowan, 2007; Hill, Rowan, & Ball, 2005). This is probably because single-level MI has been developed to work with single-level analyses (e.g. PROC MI in SAS) but is inappropriate to use with MLMs. Using a single-level MI procedure in a MLM results in variance estimates that are biased toward zero and may yield other biased parameters (Mistler, 2013a). Instead, a multilevel

multiple imputation (MLMI) procedure would need to be used where separate imputation regression equations are needed for each level-2 unit, which is currently not available in most software packages. Further, the exploration of MDTs in MLMs is more complex because sample size and percent of missing data can happen at multiple levels. Thus, examining the impact of sample size and percent of missing data on MDTs in MLM requires the manipulation of these characteristics at both level-1 and level-2.

Since MLMs are applicable to various applied research settings, it is important for researchers to be aware of the impact of MDTs on the inferences drawn from their analyses in order to appropriately handle missing data. To date, most of the research on the impact of MDTs is conducted using traditional regression techniques. Whereas a few studies have been conducted on the impact of MDTs in MLMs (Gibson & Olejnik, 2003; Kwon, 2011; Zhang, 2005) these studies typically examine missingness at only one level or are methodologically flawed (as outlined in Chapter 2). Since most software packages that handle nested data require that missing data be treated prior to data analysis, either by listwise deletion or imputation (Cai, 2008), the purpose of this study is to compare the impact of listwise deletion and multilevel multiple imputation (MLMI) on a two-level linear model.

Study Design

Currently, researchers commonly use software such as HLM 7 (Raudenbush, Bryk, & Congdon, 2005) or SAS PROC MIXED (SAS Institute, 2008) to estimate MLMs. These programs require that missing data be handled prior to analysis by listwise deletion or imputation. Listwise deletion is the most dominate method for dealing with

missing data (Peugh & Enders, 2004). This is problematic for MLM analyses because deletion can result in a substantial loss in sample size, especially when missing data occurs at level-2. However, if multiple imputation procedures can be used in the case of MLM, researchers could retain their original sample size and the full statistical power of the study, provided that the method does not bias parameter estimates. The goal of this study is to compare listwise deletion and multilevel multiple imputation (MLMI) in the context of a linear random intercept two-level MLM.

Specifically, listwise deletion and MLMI were compared on four outcomes including a) bias in point estimates of the fixed effects, b) Type I error rates estimated for null fixed effects, c) statistical power for non-null fixed effects, and d) average confidence interval coverage for each fixed effect. In addition, several design factors were manipulated including a) missing data assumption, b) level-1 sample size, c) level-2 sample size, d) percent of missing data at level-1, and e) percent of missing data at level-2.

In order to evaluate both MDTs, a simulation study was conducted. Data were generated under a linear two-level model. Simulation of the parameter values was designed to generalize to current applied research articles. Once datasets were generated using SAS, certain percentages were deleted from two level-1 variables and two level-2 variables (one from a null variable to examine Type I error rate and one from a non-null variable to examine power) using two methods. The first method randomly deleted certain percentages from the data set to meet the assumption of missing completely at random (MCAR). The second method deleted a certain percentage of missing data based

upon the value of another variable in order to meet the assumption of missing at random (MAR). More information about this is provided in Chapter 2 and Chapter 3.

Once data were deleted to meet the assumption of MAR or MCAR, data were either imputed using Mplus (Asparouhov & Muthén, 2010) prior to being estimated with SAS PROC MIXED (a MLMI procedure), or the data were simply estimated with PROC MIXED (which listwise deletes by default). Comparisons of the parameter estimates produced by each MDT were compared to the parameter estimates from the complete data and the population parameter values (where appropriate) according to the four outcomes listed above.

Significance of the Study

While other modeling techniques commonly discuss the problem of missing data (e.g. SEM), researchers using traditional MLM have largely ignored the topic. However, MLM studies have gained popularity among researchers over the past several decades. This is a problem as nested data typically come from many sources (e.g. students and teachers) which increase the chances of having missing data. To date only a few research studies have examined MDTs in the context of hierarchical data (e.g. Cai, 2008; Gibson & Olejnik, 2003; Zhang, 2005). These studies, however, have methodological flaws or limitations such as using a single level imputation procedure, examining a small number of imputations during the imputation phase of the imputation procedure, only examining data that are MCAR, and only examining missing data at one level of the MLM. As yet no empirical study has been published comparing listwise deletion to multiple imputation

using a MLMI procedure. Further, most research only examines one level of the MLM and neglects to examine both levels simultaneously.

This research offers a thorough examination of the impact of MLMI and listwise deletion in MLMs, and has substantial implications for applied research. Since MLMs involve a further level of complexity (e.g. sample size at both level-1 and level-2), research on single-level analysis cannot easily be generalized to MLMs. Therefore, applied researchers who use MLM methods can benefit from this study. Further, methodologists who study MLM can also benefit and expand upon this study in order to inform best practices for studies with missing data.

Limitations of the Study

As with any research study, the generalizability of the results is limited to the design factors and facets manipulated in the study. For example, this study examined a random intercept model. Because of this, results may not generalize to more complex models such as models that have both random intercepts and random slopes. Also, univariate normality is assumed for all predictor variables and multivariate normality is assumed with the MLMI procedure. Applied researchers wishing to use the guidelines set forth by this study need to be cautious and mindful of how close their research scenarios mirror the design and data of this simulation study.

Chapter Summary

The next chapter presents a general review of the literature on missing data mechanisms, missing data treatments, and previous research studies on missing data in multilevel models. Chapter 3 discusses the methodology of the current study, which

provides an in-depth description of the methods used to address the main goals of the study.

CHAPTER II

Literature Review

Chapter 2 presents seven sections. In the first and second sections, an overview of missing data mechanisms and missing data treatments is presented. In the third section, a literature review of the consequences associated with missing data treatment selection is discussed. In the fourth and fifth sections an overview of missing data in multilevel modeling and an examination of how researchers have attempted to use multiple imputation in multilevel models is presented. The sixth section goes into detail on important factors to examine with missing data in multilevel models (MLMs). Finally, the seventh section discusses the goal of the current study.

Missing Data Mechanisms

Rubin (1976) introduced a classification system for the issues associated with missing data that are prevalent in current literature. The classification system developed describes the probability of missing data and is known as the missing data mechanism. Three missing data mechanisms were described by Rubin (1976) and are explained in this section including missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR).

Missing Not at Random. Missing not at random (MNAR) describes when the probability of missing data on a variable is related to the values on that variable itself, even after controlling for other variables (Rubin, 1976). Stated mathematically,

$$p(R|Y_{obs}, Y_{mis}, \Phi)$$

where *p* indicates a probability distribution, R is the missing data indicator, Y_{obs} is the observed part of the data, Y_{mis} indicates the missing part of the data, and Φ is a parameter or set of parameters that describes the relationship between R and the data.

One example of a missing data scenario that is MNAR could be a situation where English as a second language students have missing test scores on an English comprehension exam because they experienced reading comprehension problems during the exam. In this example, scores are missing for the students who are ESL and have the lowest reading comprehension ability.

Missing at Random. Rubin (1976) used the term missing at random (MAR) to classify data where the probability of missing data on a variable is related to some other measured variable in the analysis, but not the values of that variable, itself. Stated mathematically,

$$p(R|Y_{obs}, \Phi)$$

Where *p* indicates a probability distribution, R is the missing data indicator, Y_{obs} is the observed parts of the data, and Φ is a parameter or set of parameters that describes the relationship between R and the data. Stated in words, this equation states that the

probability of missingness depends on the observed data via some parameter (Φ) that relates the observed data to the missing data indicator.

There are many instances in which the missing data mechanism is MAR. For example, if a researcher was interested in job satisfaction from a nationally representative group of workers and also collected data on worker's income, it could be that workers with low income are less likely to report their satisfaction. In this case, MAR missingness may occur. That is, the missingness in the job satisfaction variable can be explained by worker income.

Missing Completely At Random. Rubin (1976) used the term missing completely at random (MCAR) to classify data where the probability of missing data on a variable is unrelated to other measured variables and is unrelated to the values of that variable, itself. Stated differently, data that meet the assumptions of MCAR are purely randomly missing and unrelated to other measured variables. Stated mathematically,

$$p(\mathrm{R}|\,\Phi)$$

Where *p* indicates a probability distribution, R is the missing data indicator, and Φ is a parameter or set of parameters that describes the relationship between R and the data. Conceptually, this states that some parameter (Φ) governs the probability of R, but missingness is no longer related to the data.

An example of how this could happen in the behavioral and social sciences is where children in a district are missing achievement scores due to personal reasons (e.g. some could move to different states, some could be ill the day before the test, etc.). In this situation, purely haphazard missingness occurs that is not related to any other variables.

Missing Data Techniques

Two main consequences of missing data are (a) a decrease in statistical power due to a loss of information and (b) the possibility of biased estimates for parameters and standard errors (Roth, 1994). Thus, conclusions drawn from data with missingness can be inaccurate if not treated accordingly. In the next two sections traditional and modern methods for dealing with missingness are presented along with pros and cons of each method.

Traditional Missing Data Treatments. Traditional techniques for dealing with missing data require the assumption that data are missing completely at random (MCAR). Of all the categorizations of missing data assumptions, this assumption is the most stringent. However, due to ease of use traditional techniques are appealing to researchers. Below are brief summaries of several traditional missing data techniques that are commonly seen in the literature or are most available in statistical packages.

Listwise Deletion. With this method, any observation with missing data on any of the variables included in the statistical analyses is discarded. Previous research has found that a majority of research studies (Peugh & Enders, 2004; Peng, Harwell, Liou, & Ehman, 2006) use listwise deletion. This method is primarily used for convenience. It is commonly available in software packages such as SAS and SPSS and has no subjective decision-making required to implement the procedure. Ultimately, if the data are incomplete, that observation, in most social science research, is thrown out of the analyses.

Although the implementation of listwise deletion is simple, research has shown that the disadvantages are vast. When the assumption of MCAR is not met, this technique can result in parameter bias, especially when the percent of missing data is high (Enders, 2010; Kromery & Hines, 1994). Additionally, the probability of detecting a difference if one exists (i.e., power) could be impacted due to the decrease in the overall sample size. As Kim and Curry (1977) show, 59% of data can be lost using listwise deletion if only 10% of the data were eliminated randomly from each variable in a data set with five variables. Elimination of 59% of the data would result in a dramatic decrease in power, while also limiting the generalizability of the sample to the population.

As an example, picture a data set where a researcher was interested in job satisfaction from a nationally representative group of workers and also collected data on worker's income. However, workers with low income might be less likely to report their satisfaction. If the researcher uses listwise deletion, those participants with low income scores are systematically deleted from the dataset. Provided there is a high correlation between the income measure and the job satisfaction measure, the remaining cases are unrepresentative of the complete data set, resulting in parameter bias. In addition to bias, eliminating participant data results in a lower sample size and thus lower power. This reduction in power is always a problem, even when data actually meet the assumption of MCAR (Enders, 2010).

Pairwise Deletion. This method is similar to listwise deletion in that incomplete cases are missing; however with this method researchers attempt to use as much data as possible by eliminating cases on an analysis-by-analysis basis. Because of this, each analysis has a different sample size. The use of deleting cases on an analysis-by-analysis

method can be more powerful than listwise because all of the available data are used for a given analysis. According to Kim and Curry (1977) this makes pairwise deletion attractive to researchers especially when there are a small number of missing cases on each variable relative to the total sample size and a large number of variables are involved. Although this is the default setting in SPSS, SYSTAT, and SAS for descriptive, correlation, and regression analyses this method is not as common as listwise deletion, with previous research supporting that only about 7.6% of researchers are using this method (Peng et al., 2006).

Major disadvantages of this method occur when the assumption of MCAR is not met. Consistent with listwise deletion, when this assumption is not met this procedure can result in parameter bias. Unlike listwise deletion, since the subsets of cases are used differentially across analysis issues with association can result. Specifically, due to the method by which the covariance and correlation matrices are calculated, values exceeding |1| can occur causing issues with nonpositive definite matrices (Enders, 2010; Little, 1992; Marsh, 1998; Wothke, 1993).

Single Imputation Methods. Single imputation is an umbrella term for procedures to treat missing data where, prior to analysis, one replacement value is calculated and assigned to each missing data point. Examples of such techniques include mean imputation, regression imputation, stochastic regression imputation, and hot-deck imputation. Single imputation methods can be attractive to researchers because they do not require deletion of cases, resulting in a complete set of data. Since the full data set is retained, power is not reduced.

However, techniques that utilize this procedure have serious disadvantages and most of the methods are generally not recommended for use (Enders, 2010). Even under the MCAR assumption the techniques can produce biased parameter estimates. In fact, mean imputation has been shown to be the missing data treatment with the most disadvantages (Brown, 1994; Enders & Bandalos, 2001; Gleason & Staelin, 1975; Kim & Curry, 1977; Kromery & Hines, 1994). Though single imputation procedures can cause bias, two of the methods were instrumental to the development of modern missing data techniques, such as multiple imputation. As such, they warrant some discussion here. The following section provides a more complete explanation of regression imputation and stochastic regression imputation and offers an argument as to why this research focuses solely on multiple imputation.

Prior to the development of stochastic regression imputation, researchers were using regression models to predict missing data from variables with no missing data, thus using the association among variables in a dataset to generate missing values (i.e. regression imputation). Below is an example of a regression model where two variables (parent income and parent highest education) predict a third missing variable (prior achievement):

$$PriorAch_i = \beta_0 + \beta_1(ParentIncome_i) + \beta_2(ParentalEd_i)$$

With this process, once data are imputed the values fall directly on a regression line with a nonzero slope resulting in a perfect correlation between the variables that are imputed, even when data are MCAR (Note that each missing data pattern involves its own regression equation). Thus, the imputed values lack variability that would have been

present had the data been complete, resulting in attenuated variances and covariances (Beale & Little, 1975; Gleason & Staelin, 1975; Kromery & Hines, 1994; Olinsky et al., 2003; Raymond & Roberts, 1987; Timm, 1970), thus biasing parameters.

In an attempt to adjust the attenuation of variance in regression imputation, stochastic imputation was developed (Enders, 2010). Stochastic imputation is also a regression based imputation procedure, yet goes a step further to replace the lost variability by adding a normally distributed residual term to each predicted score. Using the same equation above where I use two variables with complete data (parent income and parent highest education) to predict a third missing variable (prior achievement), the model would change to the following model:

$$PriorAch_i = \beta_0 + \beta_1(ParentIncome_i) + \beta_2(ParentalEd_i) + z_i$$

Adding the residual term (z_i) creates an imputed value that is a random variable from a normal distribution with a mean of zero and a variance equal to the residual variance from the regression model. Stochastic regression has been found to be useful in imputing data even when data are MAR (Little & Rubin, 2002) and also produce similar estimates to modern missing data techniques (discussed below; Gold & Bentler, 2000). While stochastic regression appears to be a useful tool because it does not produce biased parameter estimates, standard analysis techniques treat the imputed values as real data thereby ignoring the same additional sampling error from the missing data. Because of this, stochastic imputation procedures produce biased standard errors that increase the likelihood of making a Type I error (Enders, 2010).

Modern Missing Data Techniques. Due to the shortcomings in traditional techniques, modern techniques for treating missing data have been developed. Modern techniques require the assumption that data are missing at random (MAR). The assumption of MAR is less stringent then the MCAR assumption because the pattern of missingness can be explained by other measured variables. However, there is no way to test a MAR assumption with complete accuracy (Enders, 2010). Two MDTs have been developed that are considered to be "state of the art" (Schafer & Graham, 2002) including FIML and MI.

Full Information Maximum Likelihood. Full information maximum likelihood (FIML) is a missing data procedure that borrows information from observed data to estimate the parameters and the standard errors. FIML was developed from maximum likelihood estimation therefore a review of maximum likelihood estimation is provided. I also present a discussion about the slight differences between FIML and maximum likelihood for treating missing data.

In cases of complete-data, the first step of FIML is to specify a population distribution, which is typically the multivariate normal distribution in educational and psychological research (Enders, 2010). Once a population distribution is assumed, a probability density function is used based upon the assumed distribution. A probability density function is a continuous random variable for which the mathematical integral gives the probability that the value of the variable lies within the same interval. Using the probability density function, researchers can calculate the probability of obtaining a score value given a particular mean and variance, which is the essence of FIML. Using this probability density function, maximum likelihood estimation is an iterative process that

repeatedly auditions different values for population parameter values until it finds estimates that most likely produced the observed data (Enders, 2010). It does this by repeating the log-likelihood computations many times, each with different values of the population parameters. The algorithm evaluates the sample log-likelihood of the estimates until it chooses the most plausible parameter values. The ultimate goal of this procedure is to identify the unique combination of estimates that maximize the log-likelihood and thus produce the best fit to the data.

Using maximum likelihood estimation for missing data is conceptually the same, with a few more nuances, as the procedure with complete-data analysis. Similar to maximum likelihood estimation, a multivariate probability density function is specified that describes a multivariate normal distribution (in most cases). From here, the algorithm routinely substitutes a score vector and a population parameter value into the density function to assess the fit until it finds scores that are most probable given the assumed distribution. In situations with missing data, computations for individual log-likelihoods and standard errors must be altered to accommodate different amounts of data for each case (i.e. the size and contents of the matrices could change across observations and the log-likelihood computations for a particular observation depend on complete data; Enders, 2010). In addition, the case of missing data requires an iterative optimization algorithm (typically the EM algorithm) that is not needed in the case of complete data (complete data are more straightforward because familiar equations define the maximum likelihood parameter estimates; thus the optimization algorithm is not needed [Enders, 2010]).

Although slightly different than maximum likelihood, using FIML to treat missing data has the same ultimate goal as maximum likelihood estimation-- to identify the unique combination of estimates that maximize the log-likelihood and thus produce the best fit to the data. Take note that while the algorithm used in this procedure borrows information from other variables, it does not actually impute or replace the missing values. Rather, it uses all of the available data to estimate the parameters and standard errors (Enders, 2010).

Schafer and Graham (2002) regard FIML as a state-of-the-art method as it produces parameter estimates free from bias under the less stringent MAR assumption, thus providing accurate parameter estimates in cases in which traditional methods would fail. Power is also preserved in this method because the procedure uses all available observed information. There are two major disadvantages with the method. First, FIML is not widely available in traditional software packages such as SPSS and SAS. Common specialty software packages that include the procedure are typically structural equation software packages such as LISREL and MPLUS.

Second, FIML is typically more useful when missingness occurs in the dependent variable (Enders, 2014; van Buuren, 2011). This is because predicted variables are treated as fixed in statistical analyses. For the dependent variable we make distributional assumptions such as that the residuals are normally distributed. However, the predicted variables are treated as fixed and thus do not assume that the values of the predictors are sampled from a distribution, which is required for the incomplete variables when using FIML. As an example, the software package Mplus (which uses FIML by default) makes a sharp distinction between independent and dependent variables such that missingness in

independent variables are listwise deleted while dependent variables with missing data are treated with FIML. While there are workarounds in Mplus, these workarounds require alterations to the original model that require the predictors to no longer be fixed (Hox, 2014; C. Enders, personal communication, October 12, 2014).

Multiple Imputation. As an alternative to FIML, multiple imputation (MI) is another state-of-the-art technique (Schafer & Graham, 2002) that makes the same assumptions as FIML (i.e. data are MAR and multivariate normally distributed) yet actually fills in missing values prior to analysis. Note that this process reflects single imputation methods, which have been known to introduce issues. For example, above I mentioned how stochastic regression results in unbiased parameters (Little & Rubin, 2002), however increases the likelihood of making a Type I error because in standard analyses the imputed data are treated as "real" data. Research has also shown that the estimates yielded by stochastic imputation are similar to modern methods (Gold & Bentler, 2000; Newman, 2003). This is primarily because MI is an *iterative* version of stochastic regression imputation.

In addition to understanding stochastic imputation, understanding Bayesian estimation is important to fully understand MI since this is where the iterative portion of the procedure is rooted. Traditionally, psychological and educational disciplines define a parameter as an estimate of the true value. In the Bayesian paradigm, a parameter is a random variable that has a distribution. Instead of trying to obtain the true value, Bayesian analyses seek to describe the shape of the distribution. In order to accomplish this, Bayesian analysis typically involves specifying a prior distribution, using a likelihood function to summarize the different parameter values, and combining the prior

distribution and the likelihood to generate a posterior distribution. The posterior distribution describes the relative probability of different parameter values, and the goal of the Bayesian analysis is to describe the shape of this distribution.

Conceptually, MI involves three stages to generate a complete data set—the imputation phase, the analysis phase, and the pooling phase, outlined below:

1. **The imputation phase.** This step is conceptually an iterative stochastic regression. First, the procedure creates *m* multiple sets (*m* is determined by the researcher and should be greater than 1) of the data, which each contain different estimates of the missing values using Bayesian estimation principles. This is accomplished by using an iterative algorithm that repeatedly cycles between an imputation step (I-step) and a posterior step (P-step; Enders, 2010) and is used to determine plausible values for the missing data. Specifically, the I-step uses a stochastic regression procedure to impute the missing values, and the P-step uses the filled-in data to generate new estimates of the parameter. At each P-step, the iterative algorithm uses the filled-in data from the preceding I-step to define the posterior distribution of the parameter. Once the posterior distribution of the parameter is defined, Monte Carlo simulation is used to update the estimates from the posterior distribution. Once this is selected the process then iterates back to the I-step and uses the updated parameters from the P-step to derive slightly different regression equations (again, using stochastic regression from the updated distribution) from the previous I-step and then cycles back to the P-step to update the

parameters. This cycle repeats a number of times and generates several copies of the data, each of which contains a unique estimate of the missing values.

2. **The analysis phase.** Now that there are multiple sets of data, each of the datasets needs to be analyzed. Thus, this phase takes each of the *m* data sets created in the imputation phase and analyzes it with the specified technique set by the researcher (e.g. if the researcher was using regression each of the data sets created would be analyzed using the same regression model of substantive interest to the researcher). Stated differently, this step applies the same model the researcher would use if the data were complete, just to each of the *m* data sets created in the imputation phase, resulting in *m* number of results.
3. **The pooling phase**. Once each of the data sets has been analyzed, there are multiple sets of results from the last phase that that need to be combined into a single set of results. Rubin (1987) outlined formulas for pooling parameter estimates obtained during the analysis phase to combine results from each of *m* data sets into a final set of results. For example, in order to pool the means the arithmetic mean is taken across all *m* data sets, while pooling the standard errors is slightly more complex but the overall goal is the same.

MI is an umbrella categorization for many procedures that use this three step approach. However, depending on the type of data used, a different algorithm may be needed in the imputation phase. The most common algorithm, termed the data

augmentation algorithm (Schafer, 1997; Tanner & Wong, 1987), is the most widely used and handles multivariate normal data. Whereas conceptually it sounds like this procedure is simply creating fake data, research has shown that this procedure produces unbiased parameter estimates because the estimates are averaged over a number of plausible estimates (Rubin, 1987; Schafer & Graham, 2002). Through this process the procedure is not placing a significant weight on a single imputation like single imputation methods (Enders, 2010).

MI can be an effective approach that produces estimates that are consistent, asymptotically efficient, and asymptotically normal when data are MAR (Allison, 2002). The major disadvantage of MI is that there are several ambiguities, especially in the imputation phase. For example, a researcher must decide which variables to include in the imputation model, how many copies of the data should be made, and what algorithm to use. Thus, it is easy to misuse the procedure. However unlike FIML, this procedure is commonly available in most software packages and does not make distinctions between independent and dependent variables, and is thus more flexible than FIML.

Consequences associated with MDT selection

Two main consequences of missing data are (a) a decrease in statistical power due to a loss of information and (b) the possibility of biased estimates for parameters and standard errors (Roth, 1994). Recent literature has shown that traditional missing data techniques exhibit both of these problems depending on what type of missing data mechanism underlies the data. For example, listwise deletion leads to inflated standard errors for the parameter estimates in the case of MCAR and bias in the parameter estimates in the case of MAR (Allison, 2002). However, the general consensus is that

pairwise and listwise deletion methods work well when sample sizes are large, and the percent of missingness is small (Basilevsky et al., 1985; Kim & Curry, 1977; Roth & Swizer, 1995; Witta, 1992). Thus, using a traditional method may not reduce statistical power or bias estimates in some situations. Understanding in what situations traditional methods can be used is useful to applied researchers since employing traditional methods are much easier than employing modern methods.

In comparison, modern methods have been shown to retain power while not biasing parameter estimates when models are correctly specified (Rubin, 1996). These methods can also be used when data are MAR or MCAR, and thus seem like they are always be the safe choice (Enders, 2010). MI is an option in many software packages, however several decisions need to be made about how to conduct the MI. Researchers need to determine what imputation model needs to be used, how many copies of the data set should be used, and what auxiliary variables can be entered. Thus, MI is much more complex and misspecification may be an issue. Nevertheless, this technique is also more versatile in that it does not make distinctions between dependent and independent variables. In addition, FIML is not readily available in statistical packages and is more suited to treat missing data in dependent variables. Due to the complexity of modern MDTs, it is useful to understand under what conditions it is appropriate to use a traditional method and under what conditions researchers have to use a modern method when conducting multilevel research.

Missing Data in MLMs

The occurrence of missing data is not unique to single-level analyses. Missing data are also a serious concern for researchers using MLMs. Because data come from different sources (e.g. students and teachers) the probability of having missing data may be higher. In addition, if variables are missing for a level-2 predictor and deletion methods are used, all subsequent level-1 cases are deleted which could drastically reduce the overall sample size. Thus, some MDTs could affect how parameters are estimated at every level causing substantial bias to be present and thus wrongful conclusions could be drawn by applied researchers. Due to the need for MLMs for nested data and the propensity for MLMs to have missing data, more research needs to be conducted on MDTs in the context of MLMs.

The exploration of MDTs in MLMs has been lacking because MLMs are more complex than other models. Instead of looking at sample size and percent of missing data at just one level, we have to understand how both level-1 and level-2 sample size impact each technique as well as how percent of missing data at both level-1 and level-2 are impactful. Further, the method by which data can be treated becomes more complicated in MLMs. Enders (2014) stated that FIML is not a great option for MLMs due to software limitations and mathematical restrictions mentioned above. Thus, the best option for MLMs is to use MI. However, since MLMs are more complex imputation procedures can involve variables from different levels, creating more subjective decisions to make than when using a single level MI procedure. To date a variety of MI procedures have been used with MLM, outlined next.

Examining MI Usage in MLMs

MI has been implemented a few ways in MLMs. This section explores each of the methods used in the literature and the conclusions drawn from using restrictive MI, inclusive MI, and a single-level imputation procedure at multiple levels.

Restrictive Imputation for Level-2 Imputation. With this method, single-level MI is used to impute level-2 data using only the level-2 predictors in the imputation model and thus level-1 variables are ignored (and subsequently listwise deleted). Previous research has determined that including more variables rather than less is recommended (Collins, Schafer, Kam, 2001; Enders, 2010). Thus, including level-1 variables could offer valuable information, yet with restrictive imputation these data are left out.

An example of research using this technique was conducted by Gibson and Olejnik (2003). In this study, the impact of 5 MDTs (listwise deletion, overall mean substitution, group mean substitution, the EM algorithm, and MI at the second level of a two-level MLM) was examined under the assumption of MCAR. When implementing the MI procedure, they performed MI only at level-2 and included only level-2 predictors in the implementation model (which they acknowledge as a limitation). Results from Gibson and Olejnik (2003) show that in the case of estimating fixed effects, listwise deletion and the EM algorithm performed satisfactorily while mean substitution and MI was not as effective. For random effects, only listwise deletion performed satisfactorily except when the level-2 sample size was small (N=30) and the proportion of missingness was high (i.e., 40%).

Results from Gibson and Olejnik (2003) showed that using MI in this way lead to biased estimates while using listwise deletion introduced no bias. However, there are three methodological flaws that could be influencing the results. First, the researchers used a single level procedure that has no information from level-1 variables. Second, the imputation phase only contained three imputed data sets. The recommended minimum number of imputations for a single-level imputation is 20 (Graham, Olchowski, & Gilreath, 2007). While the number of data sets that should be imputed in multilevel models has not been systematically examined, in general, increasing the number of imputation leads to increased accuracy (Enders, 2010), thus using three imputed data sets is less likely to produce accurate estimates. Lastly, it is not surprising that listwise deletion performed well given that data were simulated in order to be MCAR, which is the correct assumption for the technique. Given this research, it is unclear how listwise would perform under the less strict assumption of MAR, which is more useful to applied researchers.

Inclusive MI for Level-2 Imputation. With inclusive MI, imputation at level-2 is performed similar to restrictive MI, however instead of using only level-2 variables, level-1 variables are aggregated to level-2 and used in the imputation model. Researchers have shown that the use of more variables is greatly preferred to using fewer variables (Rubin, 1996; Enders, 2010). With an inclusive strategy, more explanatory variables are being used so there is a reduced chance of omitting an important cause of missingness, while also increasing the probability of noticeable gains in efficiency and reduced bias (Collins, Schafer, & Kam, 2001).

A few studies have been performed to understand how restrictive and inclusive MI compare. For example, Kwon (2011) examined the impact of listwise deletion, mean substitution, restrictive and inclusive EM algorithm (a FIML approach using the EM algorithm), and restrictive and inclusive MI on the second level of a two-level MLM where the probability of missingness was MAR. Results showed that the number of level-2 predictors and sample size did not impact bias of the MDTs, while the proportion of missing data significantly impacted bias. Specifically, when the proportion of missing data increased, the relative bias among the MDTs tended to increase in most fixed effects and some random effects. Further, results showed the inclusive MI and listwise deletion generally outperformed the other MDTs that produced "practically acceptable" bias in most fixed effects that were highly related to missingness, however listwise deletion produced the largest RMSE and confidence intervals. Restrictive EM and inclusive EM performed well except in the cases with large proportion of missing data (30%). Lastly, restrictive MI and mean substitution produce bias with even a small proportion of missing data (less than 15%).

Additionally, Cai (2008) also used a simulation study in order to examine how listwise deletion, mean substitution, restrictive and inclusive EM, and restrictive and inclusive MI impact bias at level-2 in a 3-level MLM model under MAR missingness. For this study, restrictive MI only used level-2 variables during imputation and inclusive strategies used both level-1 and level-2 variables while disregarding level-3 variables (level-3 was simulated to have complete data). Results showed that the two MI methods did not produce satisfactory estimates for level-2 fixed effects, however inclusive MI outperformed the restrictive MI on estimates of fixed and random effects across all

conditions. Additionally, listwise deletion performed well when the level-2 sample size was small; however the precision was the worst. Lastly, it was determined that the restrictive EM method was effective in producing accurate and precise estimates for fixed effects and the inclusive EM performed well for estimating random effects.

Using MI at Multiple Levels. To date, MI usage has been examined using higher levels and very few studies have imputed at multiple levels. While the higher levels should be of more concern since missing a variable at the highest level and using deletion methods can lead to deleting every corresponding level-1 unit, imputation at both levels is more appropriate in order to retain as much power as possible. Currently the only research done on using MI usage at multiple levels was done by Swoboda and Kim (2010). In this study they compared single-level MI methods to a two-level MI method under MCAR and MAR in a three level random intercept model in their simulation study. Overall, their results showed that using a single-level MI procedure does not take into account the nested structure of the data and does not work correctly for missing data at levels beyond level 1. Thus, a more complicated MI procedure that takes into account clustering has not been investigated in the context of imputing at multiple levels in a MLM.

Table 2.1 provides a summary of the missing data techniques described in this section and Table 2.2 provides a summary of the research for each technique. Looking across this literature, MDT usage in linear MLM is not consistent resulting in confusion about which MDTs are acceptable to use with nested data. Some literature suggests that listwise deletion is an effective technique (Cai, 2008; Gibson & Olejnik, 2003; Kwon 2011); however listwise deletion results in large RMSE and confidence intervals (Cai,

2008; Kwon, 2011) and only performed well when the level-2 sample size was small (Cai, 2008). Additionally, the way people use the single-level MI procedure varies, resulting in conflicting statements about the utility of MI for MLMs. Most, but not all, researchers have determined that single-level MI is inappropriate to use with MLMs (Cai, 2008; Gibson & Oljnik, 2003; Kwon, 2001; Swoboda & Kim, 2010).

Important Factors for Examining Missing Data in MLMs

Researchers who have investigated missing data in traditional regression have concluded that percent of missing data and sample size are important factors when investigating the utility of missing data techniques. However, examinations of these design factors are more complicated with MLMs, as each factor needs to be examined at multiple levels. Because of this, each factor needs to be examined in the context of MLMs in order to determine if the general guidelines established by previous literature hold for each level of the MLM. More information about relevant design factor is presented in this section.

Sample Size and Percent of Missing Data. Most comparison studies examining MDTs in a linear regression context include sample size and percentage of missing data as study variables (see Kromery & Hines, 1994; Raymond & Roberts, 1987; Roth & Swizer, 1995; Witta, 1992). This is because sample size is one of the main factors affected using traditional deletion methods and the final sample size after deletion is based upon the percent of missing data in the sample. In addition, very general guidelines exist from previous literature that demonstrate modern MDTs produce better estimates

Table 2.1 *Names, Descriptors, and Description of Missing Data Techniques Used in Multilevel Models*

Full name of Technique	Acronym/Descriptor	Description
Listwise Deletion	LD	The practice of deleting cases that have missing data on analytic variables
Single-Level Multiple Imputation	MI	A data imputation procedure where a model is used to replace missing values. With single-level imputation, the same regression equations are used for each level-2 unit. This can be performed many different ways (see single-level multiple imputation at multiple levels, restrictive multiple imputation for highest level imputation, and inclusive multiple imputation or highest level imputation)
Single-level Multiple Imputation at Multiple Levels	Single-level MI at multiple levels	The use of a single-level multiple imputation model (e.g., PROC MI) applied separately to multiple levels of the hierarchy
Single- Level Restrictive Multiple Imputation for Highest Level Imputation	Restrictive MI	A single-level imputation procedure where level-2 data are imputed using only the level-2 predictors in the imputation model and thus level-1 variables are ignored (and subsequently listwise deleted).
Single-Level Inclusive Multiple Imputation for Highest Level Imputation	Inclusive MI	A single-level imputation procedure where level-2 data are imputed using the level-2 predictors in the imputation model as well as level-1 predictors aggregated up to level-2.
Restrictive Expectation Maximization	Restrictive EM	A Full Information Maximum Likelihood Approach using the Expectation Maximization (EM) algorithm on level-2 data with only level-2 variables included
Inclusive Expectation Maximization	Inclusive EM	A Full Information Maximum Likelihood Approach using the Expectation Maximization (EM) algorithm on level-2 data with level-2 variables included as well as level-1 variables aggregated to level-2
Multilevel Multiple Imputation	MLMI	A data imputation procedure where a model is used to replace missing values. With MLMI, separate imputation regression equations are needed for each level-2 unit, thus imputing at multiple levels and taking clustering into account. Thus, level-1 and level-2 data are used and level-1 data does not need to be aggregated.

Table 2.2 *Summary of Results from Previous Research Examining Missing Data Techniques in Multilevel Models*

	Listwise Deletion	Single Level MI at multiple levels	Restrictive MI	Inclusive MI	Restrictive EM	Inclusive EM
Fixed Effects	+ (MCAR) + (MAR)	- (MAR) - (MCAR)	- (MCAR) - (MAR)	+ (MAR) + (MAR)*	+ (MCAR) + (MAR)* + (MAR)	
Random Effects	+ (MAR) + (MCAR)*		- (MAR)	- (MAR)		+ (MAR)

*Note. + indicates satisfactory; - indicates unsatisfactory; * indicates with some exceptions*

than traditional MDTs when the sample size is "small" and the proportion of missing data is "high" (Basilevsky et al., 1985; Raymond & Roberts, 1987; Roth & Switzer, 1995). Similarly, previous research also generally recommends that deletion methods work well in estimation of regression coefficients when the sample size is "large" and the number of missing values is "small" (Basilevsky et al., 1985; Kim & Curry, 1977; Roth & Switzer; 1995; Witta, 1992) Thus, sample size and percent of missing data have a large influence on MDT selection and are interrelated.

In the context of MLM, much research has been conducted on the influence of sample size at multiple levels on accuracy of parameters, standard errors, confidence interval coverage, and statistical power (Austin, 2005; 2007; 2010; Bell, Morgan, Schoeneberger, Loudermilk, Kromery, & Ferron, 2014; Bell, Schoenenberger, Morgan, Ferron, & Kromery, 2010; Bell, Schoenenberger, Morgan, Zhu, Ferron, Kromery, 2011; French & Finch, 2011; Goldstein, 2003; Hox, 2010; Maas & Hox, 2004; 2005; Moineddin, Matheson & Glazier, 2007; Mok, 1995; Theall, Scribner, Broyles, Yu, Chotalia, Simonsen, Schonlau & Carlin, 2011; Van Der Leeden, Busing & Meijer, 1997). As a result of this literature, guidelines have been established to ensure adequate accuracy

and precision. However, all of these guidelines are based upon complete-data situations. With non-nested data, Kim and Curry (1977) showed that 59% of data can be lost using listwise deletion if only 10% of the data were eliminated randomly from each variable in a data set with five variables. Using a deletion method in MLM could result in an even more dramatic reduction of the sample size, especially if data are missing at level-2 as all the corresponding level-1 units will subsequently be deleted. Thus, when deletion methods are used researchers should keep in mind recommended sample sizes as well as the final analytic sample size after the deletion method has been used, which is determined by the percent of missing data in the sample.

Whereas very general rules for percent of missing data and sample size exist for selecting an MDT in linear regression, these rules do not easily generalize to MLM. This is because sample size and percent of missing data can occur at multiple levels. Thus, examining the impact of sample size and percent of missing data on MDTs in MLM requires the manipulation of these characteristics at both level-1 and level-2. To date, no research has been conducted to examine what sample sizes and percent of missing data at each level are optimal for listwise deletion and MLMI in MLMs.

Current Study

Unlike research on MDT usage in regression, research on MDT usage in linear MLM is not consistent, resulting in confusion on which MDTs are acceptable to use with nested data. Specifically, the MLM literature is consistent that inclusive strategies are better than restrictive strategies; however whether or not MI in general performs better than listwise deletion is inconsistent. Further, most studies that use MI use unorthodox methods of imputing, for example only including variables at one level of the MLM in

the analysis model, using single-level MI, and not using enough imputed data sets. In addition, only one known study examines how an imputation procedure could be used to impute at more than one level. This study was also flawed because a single-level procedure was used and shown to be ineffective.

The fact that the literature is inconsistent is a concern because missing data are a reality of educational and social science research, yet there is no guidance to applied researchers on which methods are the most effective. In practice, all of the major MLM software packages only use complete data thus requiring researchers to either delete or impute prior to analysis. If this is not done, most packages use deletion methods by default. Some imputation strategies have also been used that impute at only one level in inclusive or restrictive ways, but all of the imputation procedures in these packages to date have ignored the clustering of multilevel data.

The current study used multilevel multiple imputation procedure (MLMI) where separate imputation regression equations are needed for each level-2 unit, thus imputing at multiple levels and taking clustering into account. To date, using a MLMI technique has not been examined in the context of how it compares to listwise deletion on parameter bias. Understanding when listwise deletion can be used, or when a modern imputation technique (i.e. MLMI) should be used is of practical importance to applied researchers utilizing MLM. Thus, the ultimate goal of this study is to thoroughly examine the impact of MLMI and listwise deletion MDTs in MLMs. Specifically, the goal is to provide applied researchers with information regarding under what conditions MLMI and listwise deletion can be used to treat missing data. Thus, conditions examined in this

study were missing data techniques, percent of missing data at each level, sample size at each level, and missing data mechanisms.

CHAPTER III

Method

The current study was designed to evaluate the performance of multilevel multiple imputation (MLMI) and listwise deletion missing data treatments (MDTs) for handling missing data in a linear two-level random intercept models under both the missing at random (MAR) and missing completely at random (MCAR) mechanisms. Several design factors and criteria were used to evaluate the utility of both MDTs. Discussion of the methodology has three sections. First, explanations of the design factors are presented. The second section provides a description of the data generation process and the third section describes the procedure for determining important design factors.

Design Factors

In the current study, listwise deletion and MLMI were evaluated using the following design factors: missing data mechanism, level-1 sample size, level-2 sample size, percent of missing data at level-1, and percent of missing data at level-2. All design factors were completely crossed yielding 2,000 conditions. For each of the 2,000 conditions explored in the completely crossed design, 500 data sets were generated. Table 3.1 presents a summary of the levels of design factors and condition count for simulation.

Table 3.1 *Levels of Design Factors and Condition Count for Simulation*

Factor	Levels					Level Count
Level-1 Sample Size	20-35	35-50	50-65	65-100	100-150	5
Level-2 Sample Size	20	35	50	65	80	5
Level-1 Percent of Missing Data	0%	5%	20%	40%	70%	5
Level-2 Percent of Missing Data	0%	10%	20%	40%		4
Missing Data Mechanism	MAR	MCAR				2
Missing Data Technique	LD	MLMI				2
Total Conditions						2,000

In order to determine the range of each design factor applicable to applied researchers, a review of current multilevel modeling (MLM) articles was completed from 2013 by typing "Multilevel Modeling" or "Hierarchical Linear Modeling" into PsychInfo. To be included in the review, articles had to meet the following criteria:

- Include analysis from a two-level linear organizational model (e.g. students nested within schools);
- Be an applied research article (demonstration and methodology articles were not included); and
- Be peer reviewed.

Thirty-nine applied articles were retrieved from the search. Each article was reviewed, and information regarding sample size at each level, percent of missing data at each level (when available), and the number of variables at each level was recorded (see analysis

section for more information on the model). An explanation of each design factor and results from the review are explained below.

Missing Data Mechanism. As described in Chapter 2, Rubin (1976) introduced a classification system for missing data problems that is prevalent in current literature. Rubin (1976) used the term missing completely at random (MCAR) to classify data where the probability of missing data on a variable is unrelated to other measured variables and is unrelated to the values of that variable, itself. Traditional MDTs (such as listwise deletion) assume that data are missing MCAR. When the assumption of MCAR is not met, this technique can result in parameter bias, especially when the percent of missing data are high (Enders, 2010; Kromery & Hines, 1994). Some research suggests that when sample size is large and the percent of missingness is low this technique can be used even if the MCAR assumption is violated (Basilevsky et al., 1985; Kim & Curry, 1977; Roth & Swizer, 1995; Witta, 1992). However, the usefulness of listwise deletion under MCAR has not been verified with MLMs.

Rubin (1976) used the term missing at random (MAR) to classify data where the probability of missing data on a variable is related to some other measured variable in the analysis, but not the values of that variable, itself. This missing data mechanism is less stringent than the MCAR mechanism and is generally considered to be a more feasible assumption to make, especially in social science research (Enders, 2010). Using OLS regression methods, researchers recommend the use of modern techniques due to the less stringent assumptions needed and state that modern methods work better when the missing data mechanism is either MCAR or MAR. However, this has not been confirmed in MLM. Since the utility of listwise deletion and multiple imputation depend on

underlying missing data mechanisms, this study examined the effectiveness of listwise deletion and MLMI under both the MCAR and MAR assumption.

Percent of Missing Data. Previous research has shown that the percent of missing data is also related to performance of MDTs. Specifically, research with traditional regression has consistently found that if the percent of missing data is "small" then in some situations using listwise deletion may not bias parameter estimates (Basilevsky et al., 1985; Kim & Curry, 1977; Roth & Swizer, 1995; Witta, 1992). In order to examine percent of missing data in MLM an added level of complexity needs to be considered where the percent of missingness needs to be examined at each level.

In the 2013 applied articles that were reviewed to inform the design of the current study, only seven of the thirty-nine articles mentioned missing data at level-1. The median missingness reported was 32% with a first quartile of 10%, a third quartile of 59%, a minimum of 6% and a maximum of 67%. In order to capture these values, I simulated 4 different magnitudes of missing data including 5% (small amount), 20% (moderate), 40% (medium to large) and 70% (large amount of missing data). At level-2, no articles indicated missing data so the same percentages were simulated to match the level-1 missing data magnitudes, with the elimination of the 70% level as that magnitude of missingness at level-2 is unfeasible in a multilevel model with the level-2 sample sizes found in applied research journals. That is, it would not seem scientifically sound to analyze data using MLMs if they had 30 level 2 units with 70% of the units containing missing data. Instead, in this situation a researcher would be more likely to use single level regression with robust SEs – more of a contextual analysis model at which point the researcher could use single level imputation for the missing variable.

Sample Size. Unlike traditional regression, MLM has observations at multiple levels. Thus, sample size was examined at both level-1 and level-2. After reviewing applied, two-level organizational MLM articles, level-1 sample sizes had a median of 14.2 level-1 units per level-2 unit. Additionally the first quartile was 5.33 level-1 units and the third quartile was 53.7 units per level-2 unit (with a minimum of 2 units and a maximum of 2,508 units). Since the overall goal of this study is to provide guidelines as to when listwise deletion and MLMI should be used, and to ensure that design factors are relevant to applied researchers, the following Level-1 sample size ranges were used: 20-35, 35-50, 50-65, 65-100, and 100-150.

Notice that the sample size used did not actually capture the median number of level-1 units in applied MLM articles. This was done due to feasibility of the overall model that was used. Because the percent of missing data was distributed across two predictors, to allow the examination of missingness on both Type I error rates and statistical power (more details provided below in simulation section), a sample size smaller than 20 was not feasible. For example, if I were to include the median of 14 level-1 units per level-2 units, and took 5% of 14 (which is the smallest percent of missing data used in the study), this would yield a value less than 1 so no removal of cases would be made and thus missingness would be 0%. Thus, level-1 sample sizes were selected due to feasibility of the percent of missing data selected, while the last two ranges are used to examine how MDTs perform with larger values.

At level-2, the applied research articles had a median of 45 level-2 units. Additionally, the first quartile was 20.5 level-2 units and the third quartile was 129 level-2 units (with a minimum of 6 units and a maximum of 577). In order to capture level-2

sample sizes commonly used in the literature as well as have a feasible amount of design factor levels to examine at what sample size different MDTs can be utilized, level-2 sample sizes of 20, 35, 50, 65, and 80 were used.

Simulation

Data were generated using SAS IML (SAS Institute Inc., 2008) based on a linear two-level organizational model in which level-1 units are nested within level-2, with all level-1, level-2, and cross-level collinearity values of 0.25 (a small to moderate correlation based on Cohen, 1988). Simulated variables were generated from a normal distribution with a variance of 1.0 using RANNOR random number generation in SAS version 9.3 (SAS Institute Inc., 2008). All variables were generated to be normally distributed and continuous at both level-1 and level-2. The data were simulated such that the second predictor at each level had no effect ($\gamma = 0$; for estimating Type I error) while all remaining predictors had non-null effects ($\gamma \neq 0$; for estimating statistical power). The data simulation program was checked by examining the matrices produced at each stage of data generation. Descriptive statistics were generated and analytical models were conducted using simulated data sets to ensure the desired characteristics were achieved.

The 39 articles that were reviewed to inform the study design factors were also reviewed to determine non-null values worthy of detection and typical intra-class correlation coefficients (ICC). For each of the articles identified, the absolute values of gamma coefficients were tabulated separately for level-1 and level-2 predictors. Table 3.2 displays the basic descriptive statistics of the gamma values and ICC values obtained across articles.

Table 3.2 *Descriptive Statistics of Gamma Values and ICCs Obtained from 2013 Literature Search*

Statistic	Overall Gamma	Level-1 Gamma	Level-2 Gamma	ICC
Mean	0.706	0.688	0.721	0.152
Median	0.234	0.224	0.194	0.110
SD	1.159	1.479	1.410	0.142
Minimum	0.014	0.009	0.002	0.001
Maximum	4.769	7.007	6.804	0.530

The overall mean gamma value was 0.706 and the median was 0.234, indicating that there is positive skew in gamma values. Because of the positive skew, a gamma value of 0.47 was chosen indicating that a 1-unit change in simulated non-null predictors is associated with a 0.47 unit increase in the simulated dependent variable. The overall mean ICC value was 0.152 and the median was 0.110, indicating that there is a positive skew in ICC values. Because of this, an ICC value of .131 was used. Once data were generated, 500 replications were conducted for each of the design conditions in the study. Based on the work of Burton, Altman, Royston, and Holder (2006), the following formula calculates the level of accuracy obtained with a certain gamma value:

$$N_{sim} = \left(\frac{Z_{1-(\frac{\alpha}{2})}\sigma}{\delta} \right)^2$$

Where δ is the specified level of estimate accuracy desired, $Z_{1-(\frac{\alpha}{2})}$ is the 1-(α/2) quantile of the standard normal distribution and σ is the standard deviation of the parameter of interest. Solving for δ above using a gamma value of 0.47, a variance of 0.006 (based upon preliminary simulations), an alpha value of .05 and 500 replications, yielded an

estimate within 1.7% accuracy. Thus, five hundred replications were chosen as it provided adequate accuracy.

After each of the 2,000 datasets were generated and prior to data analysis, data were deleted from two level-1 predictors and two level-2 predictors (one null and one non-null variable at each level) to meet the assumption of MAR and MCAR. Thus, allowing me to examine the utility of both MDTs under both missing data mechanisms for both statistical power and type I error. In this study, X1 was the non-null variable at level-1, X2 was the null variable at level-1, W1 was the non-null variable at level-2, and W2 was the null variable at level-2 (the complete model is presented in the analysis section below). The process where observations were deleted from X1, X2, W1, and W2 depended on the missing data mechanism under study. Data deletion strategies for both MAR and MCAR mechanisms are presented below

MCAR Deletion. In order to ensure that missing data met the assumption of MCAR, two extra variables Z1 and Z2 (i.e. variables that were not used in the MLM analysis or the imputation process and not correlated to any other variable) were simulated from a uniform distribution ranging from 1 to 100 for each observation. Once the variables were created from the uniform distribution, values less than a pre-specified cutoff were deleted. X1 and W1 cutoffs were at the lower end of the distribution of Z1 and Z2, respectively and X2 and W2 cutoffs were at the upper end of the distribution from Z1 and Z2, respectively. Thus for X1, 2.5, 10, 20, and 35 were the cutoffs for 5%, 20%, 40%, and 70% missingness, respectively. The cutoffs for W1 were 2.5, 10, and 20 for 5%, and 40%, respectively. Cutoffs for X2 were 97.2, 90, 80, and 65 for 5%, 20%, 40% and 70%, respectively. Lastly, cutoffs for W2 were 97.2, 90, and 80 for 5%, 20%,

and 40%, respectively. This method (see code in Appendix A) allowed missingness on either the null missing variable or the non-null variable, but never on both variables. Due to the fact that two variables at each level had missingness, the total percent of missingness for a particular level was divided between the two variables that had missing data. Thus, if the condition called for 20% missing data at level-1 and 40% missing data at level-2, 10% were missing from X1 and X2 and 20% were missing from W1 and W2.

MAR Deletion. In order to ensure that variables were deleted to meet the assumption of MAR, the third variable at each level (X3 for level 1 and W3 at level 2) was simulated to follow a normal distribution, and the pattern of missingness was determined using this variable (this variable was used in the final MLM analysis and the imputation process). MAR was produced by selecting eligible observations for deletion in X1 for all the cases above the 50th percentile of X3 with a probability of 2.5%, 10%, 20%, and 35% of the total sample size to yield an expected 5%, 20%, 40%, and 70% percent of missing data at level-1. To create missingness in X2, eligible cases were selected below the 50th percentile of X3 with the same percentages previously stated.

Similarly, MAR was produced in level-2 by selecting eligible cases in W1 for all the cases above the 50th percentile of W3 with a probability of 2.5%, 10%, and 20% of the total sample size to yield an expected 5%, 20%, and 40% percent of missing data at level-2, while eligible cases in W2 were selected below the 50th percentile of W3. This method allowed missingness at either the null missing variable or the non-null variable, but never on both variables. Once eligible cases were identified, PROC SURVEYSELECT, a SAS procedure that allows stratified sampling (with and without replacement) by selecting samples within specified strata, was used to determine missing

data. For example, if a particular condition requires 5% missingness at level-1, PROC SURVEYSELECT was used to randomly select the eligible cases without replacement from the complete data with the probability of 2.5% for X1 and X2 (code provided in Appendix B).

After deleting data to meet the assumption of MCAR or MAR, data were analyzed by PROC MIXED (which listwise deletes) or imputed using an imputation method in MPLUS (Asparouhov & Muthén, 2010). The multilevel multiple imputation (MLMI) method in Mplus was used because single-level multiple imputation (i.e. PROC MI and PROC MIANALYZE in SAS) assume that data share a common mean vector and covariance matrix and fails to take clustering into account. Thus, if a single-level imputation algorithm was employed, a single regression equation that is common to every level-1 unit would be used regardless of the fact that level-1 units are nested within level-2 units. Using Mplus' multilevel imputation method allowed for each cluster to have a unique regression line, thereby eliminating bias that occurs due to model misspecification. As explained in Chapter 2, researchers using multiple imputation in MLMs predominantly use a single level imputation procedure at the highest level only. With the multilevel multiple imputation (MLMI) method, data were imputed at both levels simultaneously.

During this MLMI, all variables are treated as outcomes (regardless of missing data pattern) with all other variables as predictors and an unrestricted covariance matrix (Schafer, 1997). This method was used because this procedure has been shown to be effective with single-level analyses, preserves all available information during the imputation phase, has been extended to the multilevel case, is mathematically congenial

for intercepts as outcomes multilevel models as used in this study, and allows for imputation to happen for both level-1 and level-2 variables (Enders, Mistler, &Keller, 2015; Schafer, 1997). Rubin (1987) has shown that using a model that has all available information preserves the associations from the analysis of the model, and omitting important effects can bias parameter estimates toward zero (Rubin, 1987).

Additionally, there are various software packages that have extended this single-level routine to the multilevel context including MPlus (Muthén & Muthén, 1998–2012), the PAN and MLMMM packages in R (Schafer, 2001; Schafer & Yucel, 2002; Yucel, 2008), and SAS (Mistler, 2013). Slight differences arise in the way these software packages work. The MPlus version was chosen as it imputes at both level-1 and level-2 (The R packages only imputes level-1 variables), required less coding, and could be ran automatically using the MplusAutomation package in R (See an example of Mplus code in Appendix C). While the SAS macro did impute at both level-1 and level-2, it was a personal macro created to match the MPlus output, but various testing showed slight differences in the output from the MPlus and SAS macro (Enders, personal communication; Mistler, personal communication).To date, Mplus is the only software package where a multilevel imputation procedure is built into the software package that allows varaibles at both level-1 and level-2 to be imputed.

As stated in Chapter 2, there are several ambiguities that occur when performing multiple imputation consisting of which variables to include during the imputation model, how many copies of the data should be made, and what algorithm to use. During a simulation study these ambiguities are much clearer given that I knew exactly which variables were related to missingness. Thus, any inferences made from this research

assume that the imputation model is specified correctly. In addition, the number of copies made during the imputation phase was 20, as recommended by Graham, Olchowski, and Gilreath (2007). Once imputations were created they were stacked together into one file and then pooled in SAS using the MMI_ANALYZE (Mistler, 2013b) macro to create the final results. This macro accomplishes the analysis and pooling phase (see Chapter 2 for a description of these stages) by taking each of the 20 copies of the data and analyzing them separately and then pooling the results according to Rubin's (1987) formulas.

Model Analysis

After each data set was generated, the simulated sample was analyzed as a two-level organizational model via the PROC MIXED procedure in SAS (SAS Institute Inc., 2008) using the maximum likelihood estimation with Kenward-Roger degrees of freedom. Given the impact that model size can have on the imputation procedure, the number of predictors at both level-1 and level-2 was also examined in the 2013 articles reviewed to help develop study design factors. Specifically, as Rubin (1996, p. 479) states, "the advice has always been to include as many variables as possible when doing multiple imputation." Thus, with more variables in the imputation model, the more effective the imputation process (Rubin, 1996; Enders, 2010). This is primarily because MI (as well as FIML) uses associations among other variables in order to fill in missing values, and the more information received from variables, the more accurate and precise the estimates will be. This is especially true when variables have a high correlation ($r > .40$) with the missing analysis variable (Enders, 2010). Research also states that even when variables are provided with weak or zero correlations there is no harm in using more variables (Collins et al., 2001).

Because of this, great care was taken to make sure the model estimated in this study was as close to the average model in applied literature. After reviewing 2013 applied research articles, 34 out of 39 articles modeled fixed effects only, thus the model estimated in this study was a random intercept model. In addition, the article review yielded a median of four predictors at level-1 and a median of 2 predictors at level-2. However, in order to examine all criteria at level-2, three variables were needed. Specifically, at level-1 and level-2, there needed to be at least three variables -- one variable with a non-null effect with missingness (X1 and W1), one variable needed to have a null effect with missingness (X2 and W2), and one non-missing continuous variable needed to be present to determine the pattern of missingness (i.e., to delete data under the MAR assumption; X3 and W3). In addition, in order to be as close to the applied research articles as possible, a fourth variable was included at level-1. Thus, the following model was examined for both the complete data set as well as the data sets simulated to be MCAR and MAR:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(W1)_j + \gamma_{02}(W2)_j + \gamma_{03}(W3j)_j + \gamma_{10}(X1)_{ij} + \gamma_{20}(X2)_{ij} + \gamma_{30}(X3)_{ij} + \gamma_{40}(X4)_{ij} + u_{0j} + r_{ij}$$

where Y_{ij} represents an outcome variable for student *i* in school *j*, γ_{00} is the intercept, which represents the grand mean of the outcome variable across students and across schools, $W1_j - W3_j$ are school-level predictors for school *j* and $\gamma_{01} - \gamma_{03}$ represent slope coefficients associated with the corresponding W_j predictor, u_{0j} is an error term representing a unique effect for school *j*, and γ_{10}, γ_{20}, γ_{30}, and γ_{40} estimate the average effect of each of the four student-level predictors. The absence of an error term for the

individual level variables indicates that the effect of the student-level predictor is fixed, or held constant across schools.

Outcomes

Four outcomes were examined in this simulation study, including (a) bias in point estimates for the fixed effects, (b) type I error rates estimated for null fixed effects, (c) statistical power for non-null fixed effects, and (d) average confidence interval coverage for each fixed effect. This section will provide details on how these outcomes were calculated.

Bias. Because the goal of using a missing data treatment is to obtain estimates, standard errors, and *p*-values comparable to what we would expect if the data were complete, bias was captured by subtracting each fixed effect with missing data to its complete data counterpart (i.e., the same data set prior to deletion) and thus can be represented by the following equation:

$$Bias_{ik} = \hat{\lambda}_{ik(miss)} - \hat{\lambda}_{ik(comp)}$$

Where $\hat{\lambda}_{ik(miss)}$ represents the estimate of the *i*th fixed effect at level *k* under a condition that was subjected to either listwise deletion or MLMI (and thus had missing data), and $\hat{\lambda}_{ik(comp)}$ represents the estimate of the *i*th fixed effect at level *k* where data were complete. Thus, a value of zero would indicate that the estimate from the missing data is equal to the estimate for the complete data; a positive number indicates that the estimate for the missing data was larger than the estimate for the complete case; and a negative value indicates that the estimate

for the complete case. Once bias was calculated for all 500 replications, an average was taken. Thus, bias, in this study, is the average deviation of the treated data from the complete data across the 500 conditions. Bias has been calculated similarly in previous missing data research (Enders, Mistler, & Keller, in press).

Type I Error Rate. A Type I error occurs when a significance test results in rejection of a true null hypothesis. In the context of this study, a Type I error occurred when a variable was simulated to have no effect (i.e., $\lambda = 0$), but the hypothesis test associated with the estimate of that effect yielded a *p*-value for the estimate less than .05 (the a priori established α value). In this study, two variables with missingness were simulated to have a null effect, one at level-1 (X2) and one at level-2 (W2). In order to calculate Type I error rate, the proportion of the replications where X2 or W2 had a *p*-value less than .05 was recorded.

In addition to calculating Type I error rate the traditional way described above, a difference in Type I error rates was also calculated. Specifically, the Type I error rate from all 500 replications for each condition was calculated for the treated conditions (i.e., the data that contained missingness and was then subjected to either MLMI or listwise deletion). Likewise, the Type I error rate of the same 500 replications prior to deleting data was calculated for the complete observations (e.g., the simulated datasets before missing data were generated and treated). The difference of these two rates was calculated by subtracting the Type I error rate of the missing case to the Type I error rate of the complete case. Thus, a value of 0 indicates that Type I error of the missing condition is equal to the complete data; a positive value indicates that Type I error rate is higher for the missing condition; and a negative value indicates that the Type I error rate

of the missing condition is smaller than the Type I error rate of the complete data. Examining Type I error rates using both of these approaches allowed me to examine if the overall Type I error rate exceeds my a priori established alpha rate of .05 and how Type I error changes from the complete data.

Statistical Power. Power describes the probability of correctly rejecting a false null hypothesis. In the context of this study, power is the proportion of the total replications for each condition where a gamma value simulated to have an effect (i.e., $\gamma \neq 0$) yielded a *p*-value value below .05 (the established α). In this study, two variables with missingness were simulated to have a non-null effect, one at level-1 (X2) and one at level-2 (W2). Thus, to determine the power, I calculated the proportion of replications for each condition that had *p*-values less than .05.

Similar to Type I error rate, power for the treated data was also compared to the power from the complete data. Thus, for the results using this comparison method, a value of 0 indicates that the power of the missing condition is equal to the complete data; a positive value indicates that the power is higher for the missing condition; and a negative value indicates that the power of the missing condition is smaller than the power of the complete data. Using both of these methods allowed me to examine how power compares to the nominal level of .80 as well as how power of the treated data differed from the complete data.

Confidence Interval Coverage. Confidence interval coverage is the proportion of the replications where the confidence interval contained the value of interest. As stated previously, when dealing with missing data the goal is to have estimates that match the

complete data. The values of interest in this context are the gamma estimates produced by the complete data set, thus all coverage values represent the proportion of replications that contain the gamma estimate produced by the complete data set. In order to calculate this, a flag was created to indicate whether or not the confidence interval created by the model with the missing data contained the estimated gamma value from the model with complete data (the flag was a value of 1 when the confidence interval did contain the estimated gamma from the complete case and a value of 0 when the confidence interval did not contain the estimated gamma from the complete case) and averaging the flag variable across all 500 replications.

Basic descriptive statistics were conducted to examine the distribution, central tendency, and range information for each of the outcomes of interest (outlined in the next section). To identify the most influential design factors for each of the outcomes, a third-order analysis of variance (ANOVA) model was conducted for each outcome with every design factor entered as a crossed independent variable. ANOVA eta-squared (η^2) values were calculated to capture the proportion of outcome variance explained by each factor combination. All design factors or combinations of design factors with ANOVA eta-squared greater than .02 were plotted for further investigation. Once all plots were created, the results of the design factors with the largest eta-squared values and the design factors of upmost substantive interest were identified and are depicted and summarized in Chapter 4.

Chapter Summary

Chapter 3 explained in detail the methods for investigating the utility of MLMI and listwise deletion in two-level organizational models with continuous predictors. The next chapter, Chapter 4, contains the results of the study and has four sections, one for each of the four outcomes of interest. Within each section, a presentation of those conditions found to be associated with each outcome overall, by level-1, and by level-2 are included. Outcomes are presented in graphical form where appropriate, accompanied by explanatory text.

CHAPTER IV

Results

The present study was intended to compare the performance of multilevel multiple imputation (MLMI) and listwise deletion in the context of linear two-level organizational models with continuous predictors. In order to compare the two missing data techniques (MDTs), a Monte Carlo simulation study was conducted that mimics data used in current applied research articles (see the simulation section in Chapter 3). Design factors of interest included missing data technique, level-1 sample size, level-2 sample size, level-1 percent of missing data, and level-2 percent of missing data (see Table 2) totaling 2,000 conditions. Each condition was replicated 500 times. Four primary outcomes were examined in this Monte Carlo study: (a) bias in point estimates for the fixed effects, (b) Type I error rates estimated for null fixed effects and the difference in Type I error between the complete and treated conditions, (c) statistical power for non-null fixed effects and the difference in power between the complete case and treated conditions, and (d) average confidence interval coverage for each fixed effect.

To identify the most influential design factors for each of the outcomes, a third-order analysis of variance (ANOVA) model was conducted for each outcome with every design factor entered as a crossed independent variable. ANOVA eta-squared (η^2) values were calculated to capture the proportion of outcome variance explained by each factor combination. In this chapter, I present those conditions found to be associated with the

outcomes in graphical form where appropriate, accompanied by explanatory text. The results from the simulations are organized by outcome, with results for the overall outcomes presented first, then by level-1 and lastly by level-2. All effects and interactions mentioned in text or displayed graphically were deemed of some moderate practical significance based upon associated η^2 values (Cohen, 1988) or show the impact of my key design factors on the outcomes (i.e., MDT, missingness, and mechanism).

Bias

Overall Bias. The overall mean bias was very close to zero (M = -0.002, SD = 0.004, min = -0.014, max = 0.0084). However, noteworthy amounts of variance occurred for overall bias that was explained by the interaction between MDT and level-1 missingness η^2 = .280). Figure 4.1 depicts the distribution of overall bias as a function of the interaction between MDT and level-1 missingness. When percent of missing data at level-1 was 0% or 5%, bias was comparable between the two MDTs and close to zero. However, when missingness was 20% or above, small amounts of negative bias were present when MLMI was used (bias was -.003, -0.006, and -0.010 for 20%, 40% and 70%, respectively). Listwise deletion, however, remained close to 0 regardless of level-1 missingness.

Although the ANOVA did not show that missing data mechanism was an important predictor of bias, this design factor was of substantive interest due to previous research on listwise deletion for non-hierarchical data. Specifically, previous research has stated that listwise deletion produced biased estimates when the amount of missingness was high and data were missing at random (MAR), whereas imputation methods should not produce biased estimates (Enders, 2010), suggesting a three way interaction between

MDT, mechanism, and missingness. Figure 4.2 depicts overall bias as a function of the three way interaction of MDT, level-1 missingness, and mechanism ($\eta^2 < .001$) and Figure 4.3 depicts bias as a function of the three way interaction of MDT, level-2 missingness, and mechanism ($\eta^2 < .001$).

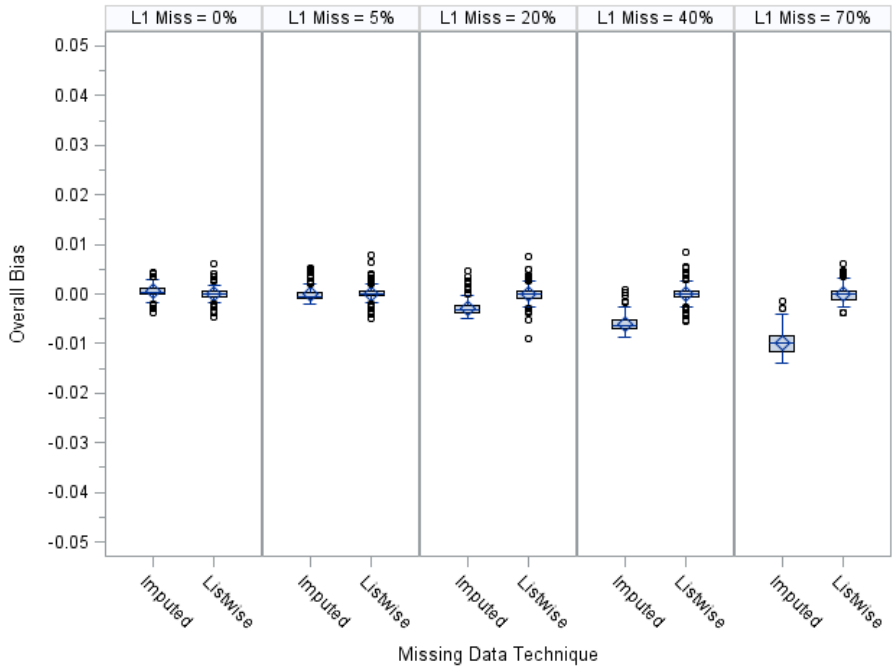

Figure 4.1 Distribution of Overall Bias by Missing Data Technique and Level-1 Missingness.

In Figure 4.2, the distribution of bias for listwise deletion was the same regardless of mechanism. Also, although MLMI produced negative bias as level-1 missingness increases (as shown previously); the mean was the same across the two missing data mechanisms, suggesting that the estimates of MLMI are equal regardless of mechanism. Similarly, Figure 4.3 shows the distribution of bias for listwise deletion and MLMI did not change as a function of data missing at random (MAR) or missing completely at random (MCAR) nor the percent of missing data at level-2.

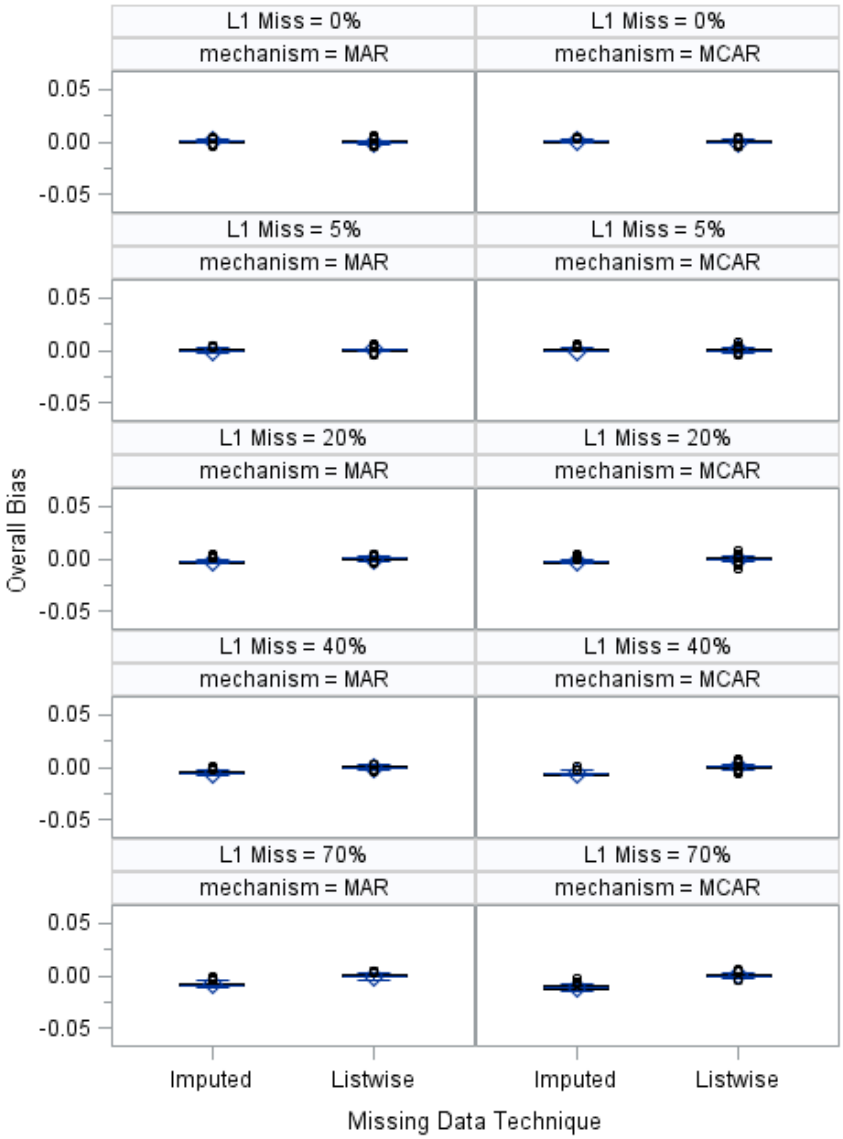

Figure 4.2 Distribution of Overall Bias by Missing Data Technique, Level-1 Missingness, and Missing Data Mechanism.

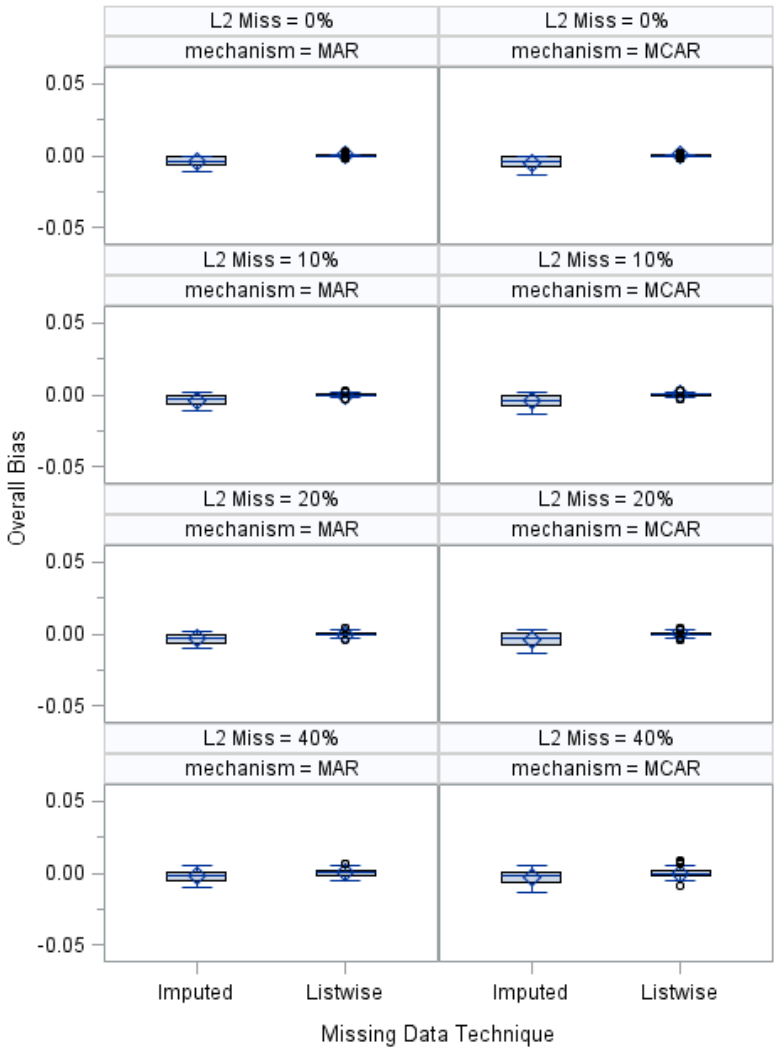

Figure 4.3 Distribution of Overall Bias by Missing Data Technique, Level-2 Missingness, and Missing Data Mechanism.

Level-1 Bias. Bias at level-1 was also close to zero (*M* = -0.008, *SD* = 0.013, min = -0.047, max = 0.0078). Noteworthy amount of variance in level-1 bias was explained by the interaction between MDT and level-1 missingness (η^2 = .307). Figure 4.4 shows that as level-1 missingness increased, MLMI resulted in bias, with values of -0.0032, -0.0129, -0.0248, and -0.0399 for 5%, 20%, 40% and 70% missingness, respectively. Listwise deletion, however, did not result in bias and was consistently around zero across

all levels of level-1 missingness. Even when level-1 missingness was at 70%, listwise deletion produced a mean level-1 bias of 0.00008.

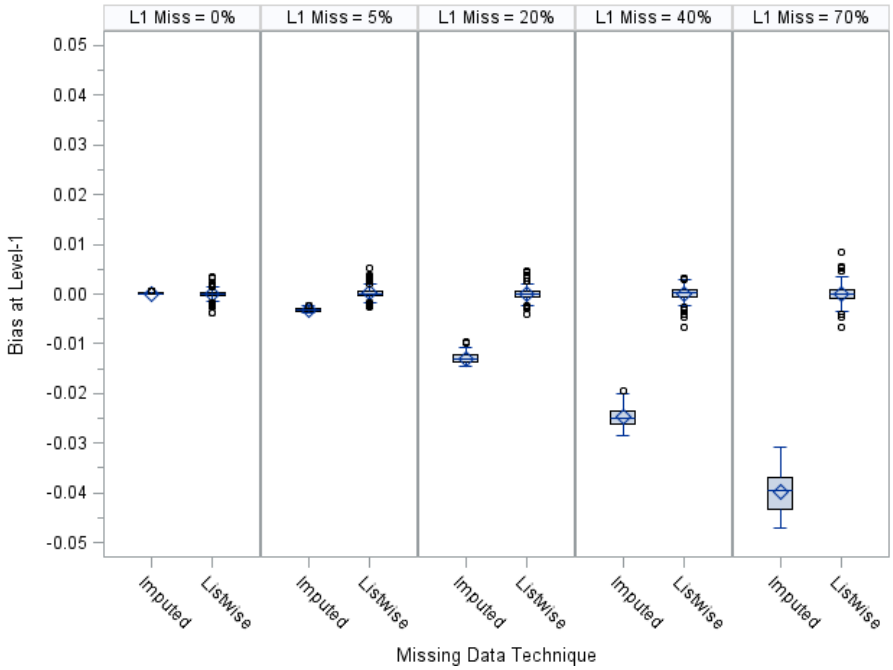

Figure 4.4 Distribution of Level-1 Bias by Missing Data Technique and Level-1 Missingness.

Level-2 Bias. Bias at level-2 was also close to zero (M = 0.004, SD = 0.007, min = -0.015, max = 0.036) with noteworthy variance explained by the interaction between MDT and level-1 missingness (η^2 = .239). Figure 4.5 depicts level-2 bias by this interaction and shows that MLMI estimates at level-2 were larger than the estimates for the complete data, with values of 0.0029, 0.0073, 0.0128, and 0.0203 for 5%, 20%, 40%, and 70%, respectively. Listwise deletion did not seem to produce biased estimates at level-2, regardless of the percent of missing data, with a bias value of -0.0001 with 70% level-1 missingness.

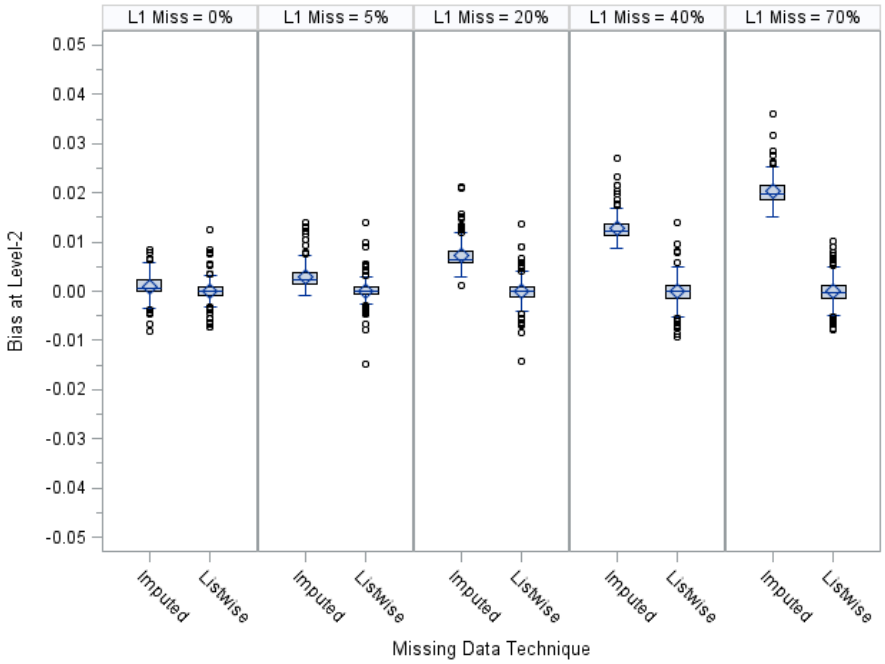

Figure 4.5 Distribution of Level-2 Bias by Missing Data Technique and Level-1 Missingness.

A very small amount of variance in level-2 bias was explained by the interaction between MDT and level-2 missingness (η^2 = .008). Figure 4.6 shows the distribution of level-2 bias by this interaction. Listwise deletion had very small amounts of variability when level-2 missingness was 0; however the variability slightly increased as level-2 missingness increased. The mean amount of bias across all levels of level-2 missingness was zero. For MLMI, bias was 0.0076, 0.0081, 0.0090, and 0.0108, when level-2 missingness was 0%, 10%, 20%, and 40%, respectively, demonstrating a slight increase in values as level-2 missingness increased.

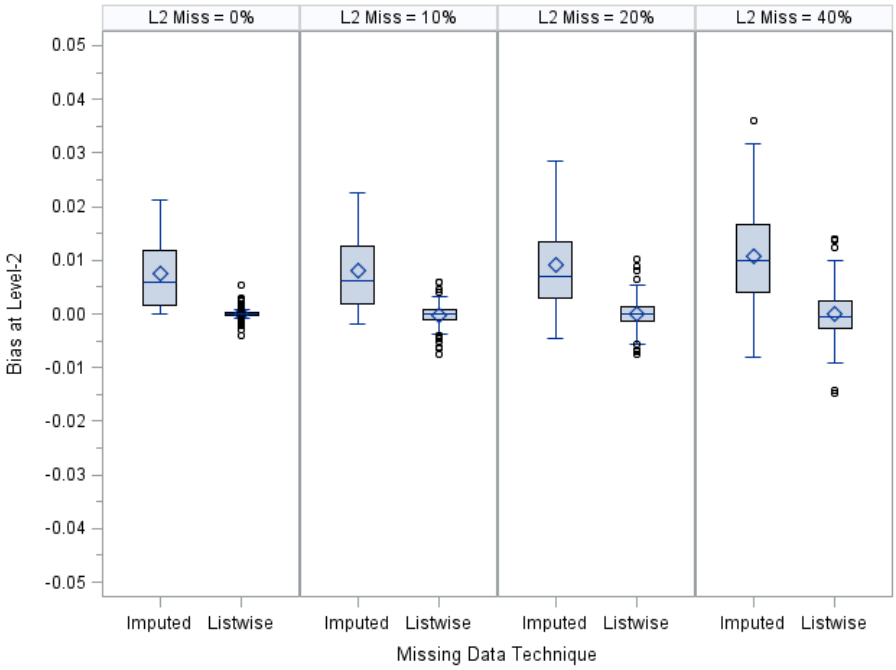

Figure 4.6 Distribution of Level-2 Bias by Missing Data Technique and Level-2 Missingness.

Type I Error Rate

Overall Type I Error Rate. Type I error rate was very close or below the .05 threshold, with an overall Type I error rate of .046 (min = .023, max = .073). Important design factors for Type I error rate included level-1 sample size (η^2 = .651), level-1 missingness (η^2 = .260), and the interaction between MDT and level-1 missingness (η^2 = .014). Figure 4.7 shows that as level-1 sample size increased, mean Type I error increased, with the mean Type I error of .040, .043, .0452, .0490, and .052 when level-1 sample size was 20-35, 35-50, 50-65,65-100, and 100-150, respectively. Thus, having a level-1 sample size of 100-150 produced an overall Type I error rate above the nominal .05 threshold.

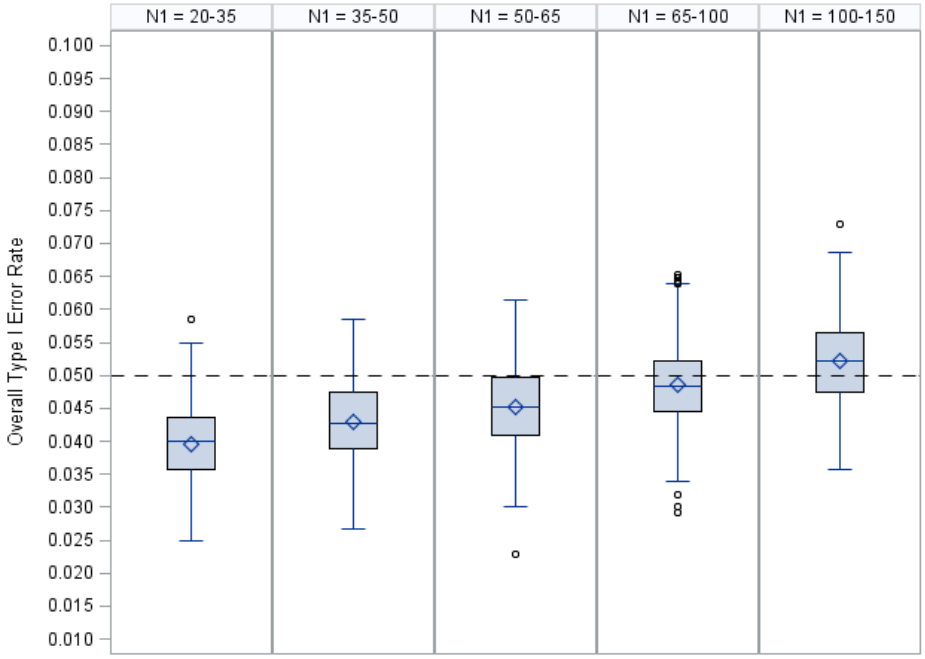

Figure 4.7 Distribution of Overall Type I Error Rate by Level-1 Sample Size.

As shown in Figure 4.8, as the percent of level-1 missingness increased, Type I error rates decreased. For MLMI, the Type I error rate was .050, .049, .047, .046, and .044 when level-1 missingness was 0%, 5%, 20%, 40% and 70%, respectively. When listwise was used, the mean Type I error rate was .047, .047, .046, .043, and .038 when level-1 missingness was 0%, 5%, 20%, 40%, and 70%, respectively. Thus, listwise deletion tended to produce Type I error rates more conservative than MLMI, and the difference between the two MDTs increased as missingness increased. Across all levels of missingness, on average, neither MDT produced Type I error rates above the .05 nominal threshold level.

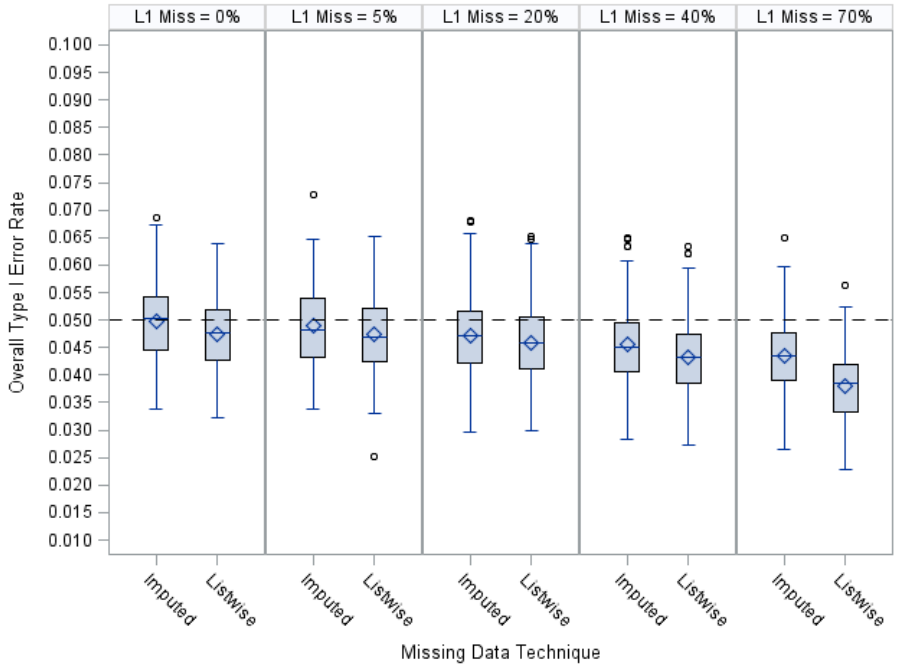

Figure 4.8 Distribution of Overall Type I Error Rate by Missing Data Technique and Level-1 Missingness.

Level-1 Type I Error Rate. At level-1, the mean Type I error rate was .037 (min = .012, max = .060). A noteworthy amount of variability in Type I error rate was explained by level-1 sample size (η^2 = .661), level-1 missingness (η^2 = .261) and the interaction between MDT and level-1 missingness (η^2 = .014). Figure 4.9 shows that as level-1 sample size increased, Type I error rate increased, with mean Type I error rates of .025, .032, .037, .043, and .051 when level-1 sample size was 20-35, 35-50, 50-65,65-100, and 100-150, respectively. Overall, when level-1 sample size was 20-35, Type I error rate was very conservative and got closer to the nominal .05 level as level-1 sample size increased.

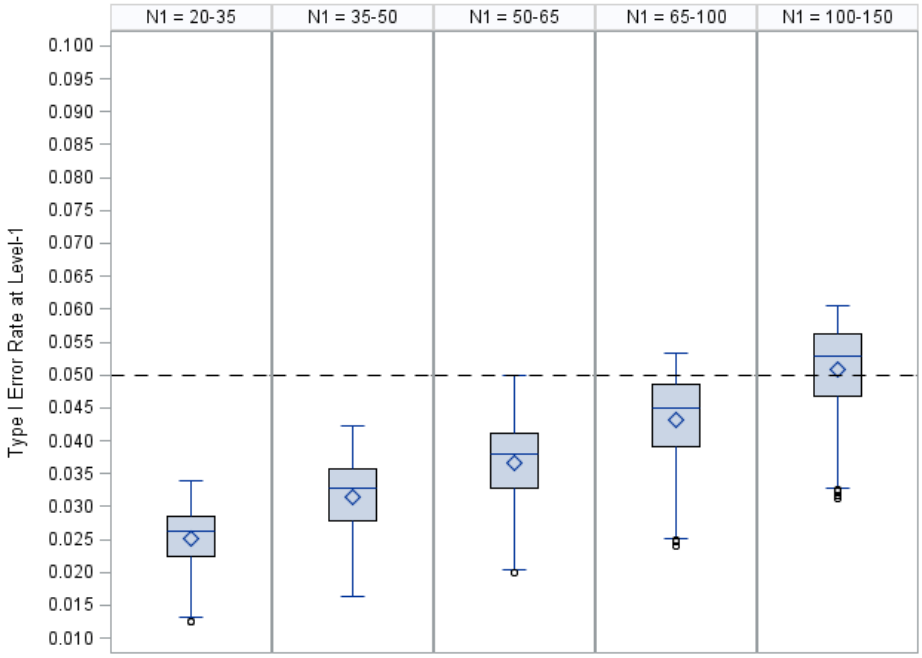

Figure 4.9 Distribution in Type I Error Rate by Level-1 Sample Size.

Figure 4.10 depicts the interaction of MDT and level-1 missingness since the interaction encompasses the main effect of level-1 missingness. This figure shows that as missingness increased at level-1, the Type I error rate for both listwise deletion and MLMI also increased. The mean Type I error across all levels of missingness was below the a priori established .05 alpha level across all levels of level-1 missingness. MLMI produced a mean of .043, .042, .039, .036, and .031 for level-1 missingness of 0%, 5%, 20%, 40%, and 70%, respectively. Listwise deletion produced a mean of .043, .042, .039, .034, and .024 for level-1 missingness of 0%, 5%, 20%, 40%, and 70%, respectively.

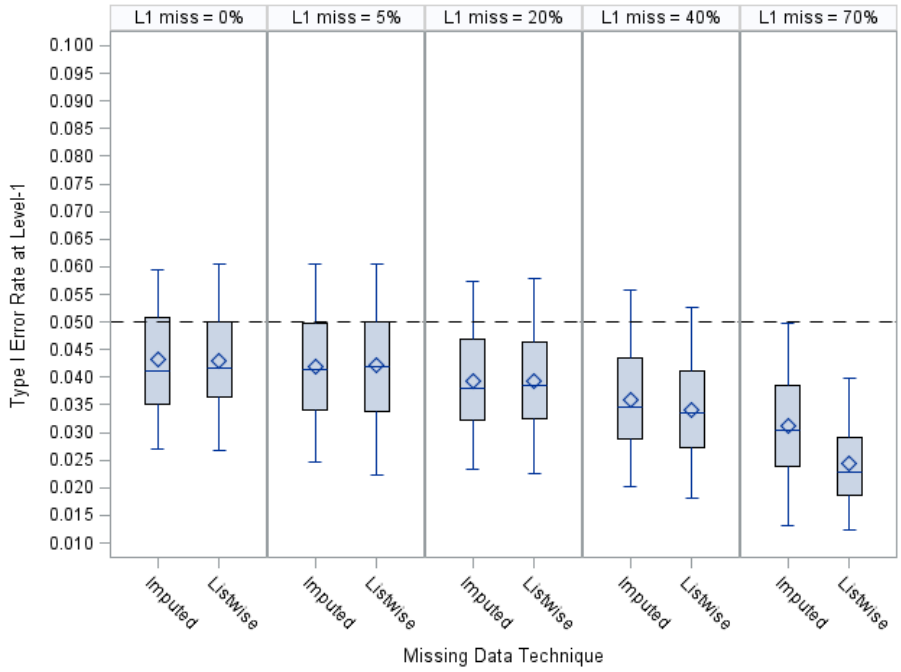

Figure 4.10 The Distribution of Level-1 Type I Error Rate by Missing Data Technique and Level-1 Missingness.

Level-2 Type I Error Rate. The mean level-2 Type I error rate was .054 (min = .020, max = .090) and was also close to the a priori established .05 alpha. The important design factors for Type I error rate at level-2 included the three-way interaction between level-1 sample size, level-2 sample size, and level-2 missingness (η^2 = .078), the three-way interaction between level-1 sample size, level-2 sample size, and level-1 missingness (η^2 = .061), the three-way interaction between level-1 sample size, level-1 missingness and level-2 missingness (η^2 = .054), the three-way interaction between level-2 sample size, level-1 missingness, and level-2 missingness (η^2 = .050), and MDT (η^2 = .028), however the overall R-squared was only .488, suggesting that only 48% of the variance in Type I error rate at level-2 was explained by the factors in this study, which is small for a simulation study. The small R-squared was a result of the lack of variability in Type I error rate at level-2.

Figure 4.11 depicts the level-1 sample size, level-2 sample size, and level-2 missingness interaction on Type I error rate at level-2 and Figure 4.12 shows the three-way interaction of level-1 sample size, level-2 sample size, and level-1 missingness on level-2 Type I error rate. Overall, Type I error rates at level-2 were very close to the a priori established alpha of .05.

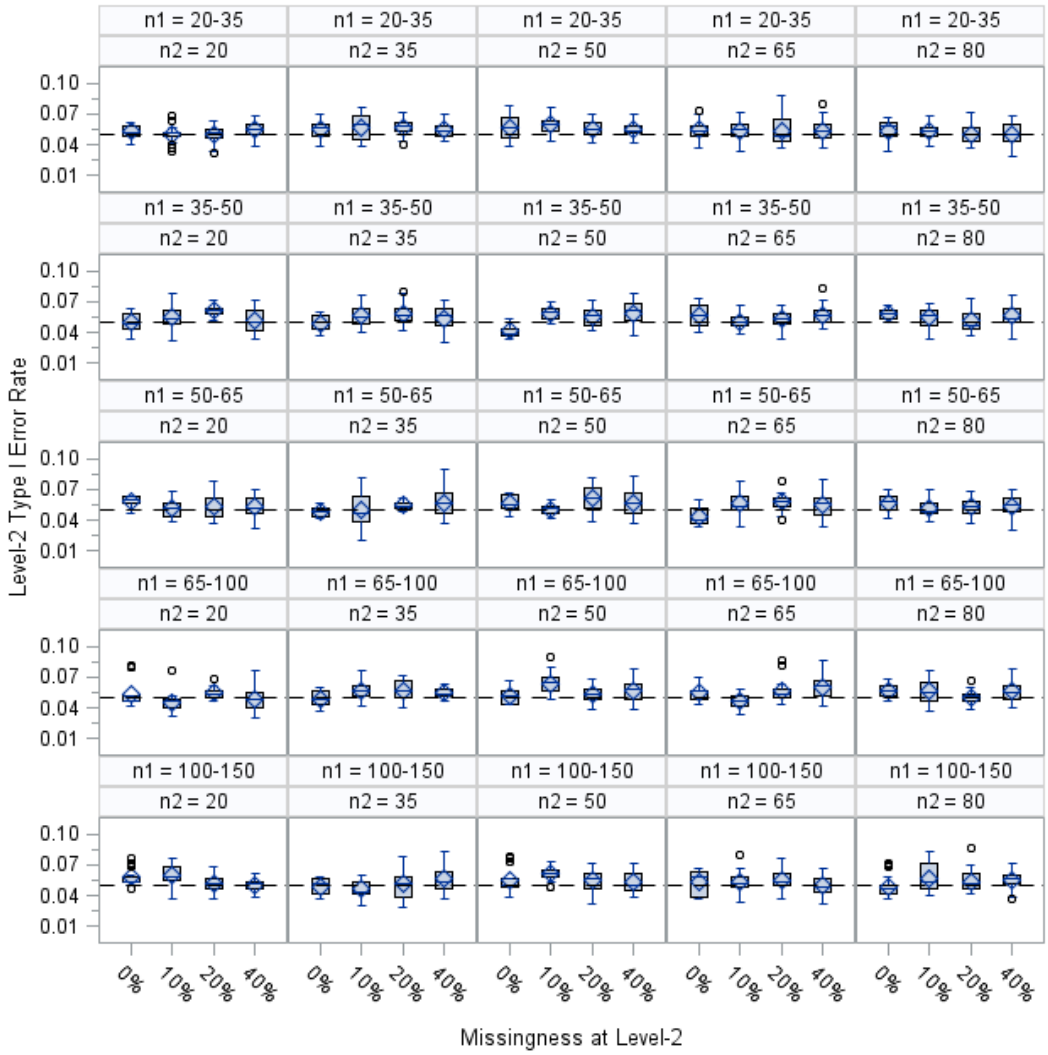

Figure 4.11 The Distribution of Type I Error Rates by the Interaction of Level-1 Sample Size, and Level-2 Sample Size, and Level-2 Missingness.

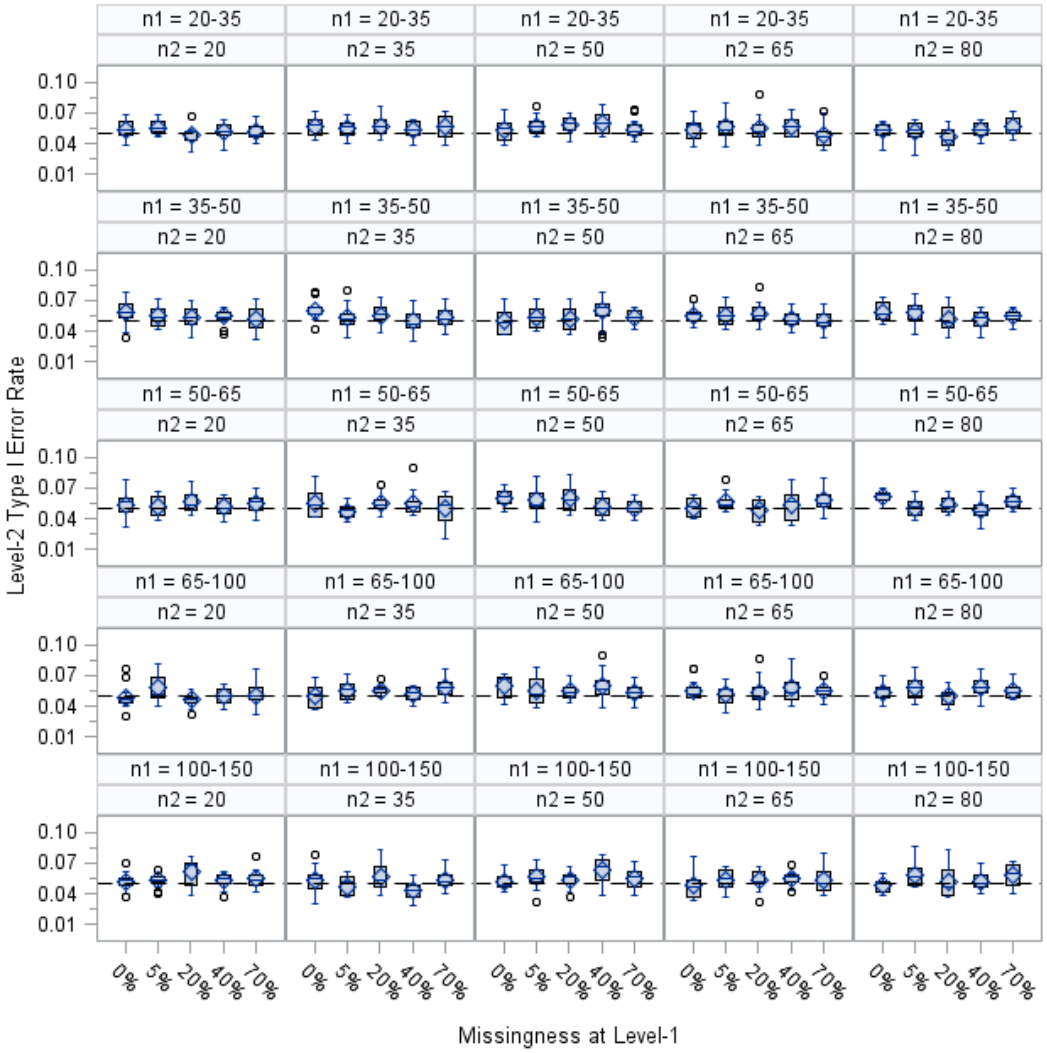

Figure 4.12 The Distribution of Level-2 Type I Error Rate by Level-1 Sample Size, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.13 shows the distribution of Level-2 Type I error rate by level-1 sample size, level-1 missingness, and level-2 missingness and Figure 4.14 shows the distribution of Level-2 Type I error rate by level-2 sample size, level-1 missingness, and level-2 missingness. Similar to the previous figures, values are close to the nominal value of .05

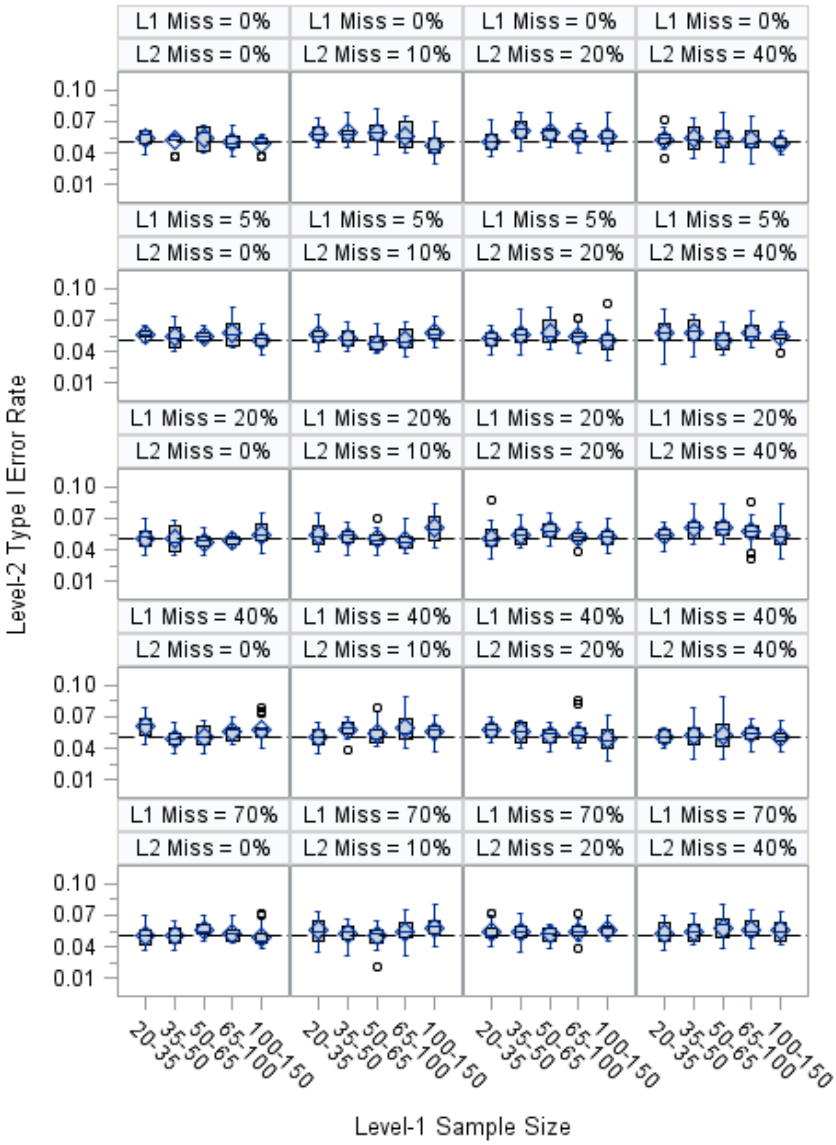

Figure 4.13 The Distribution of Level-2 Type I Error Rate by Level-1 Missingness, Level-2 Missingness and Level-1 Sample Size.

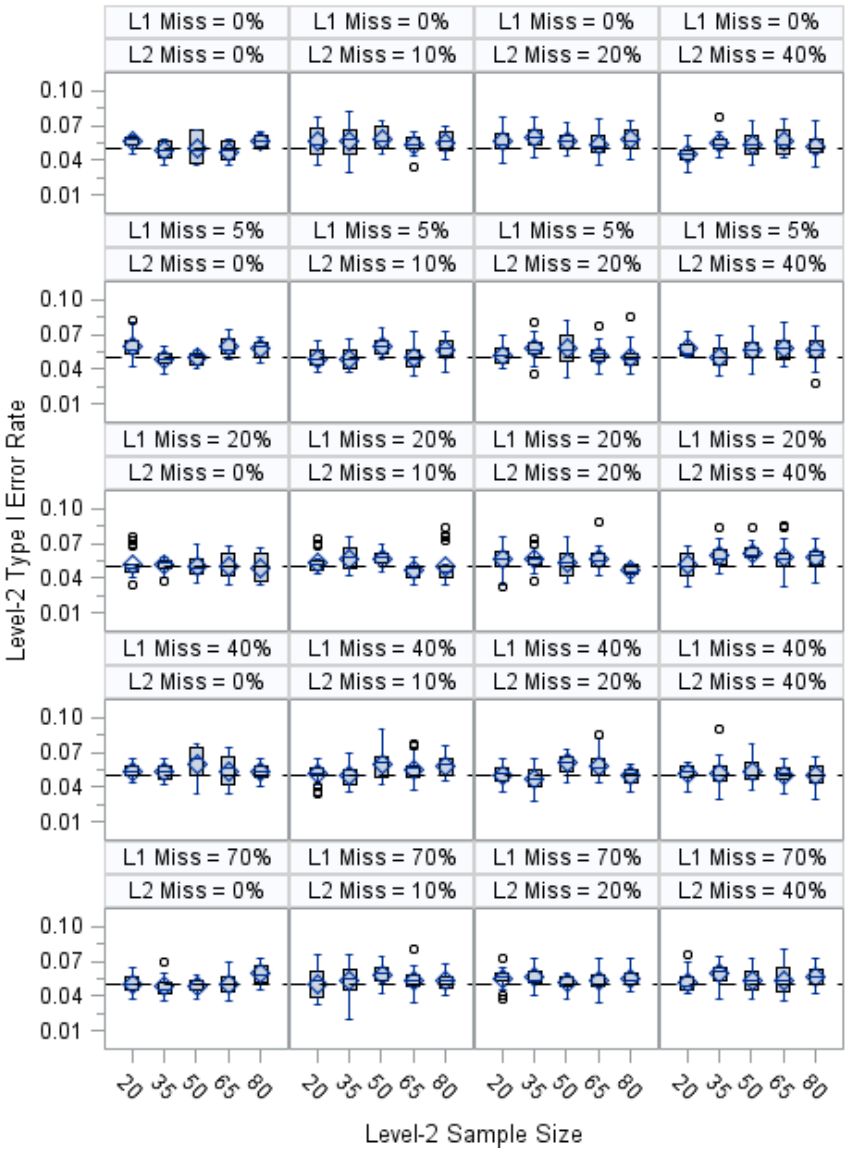

Figure 4.14 The Distribution of Level-2 Type I Error Rate by Level-2 Sample Size, Level-1 Missingness, and Level-2 Missingness.

Figure 4.15 shows the distribution of level-2 Type I error rate by MDT. The mean Type I error rate at Level-2 was slightly higher than the nominal .05 level, with a mean of .056 for the imputed conditions and a value of .052 for listwise deletion. Overall, however, they were very close to the nominal .05 level with similar variability.

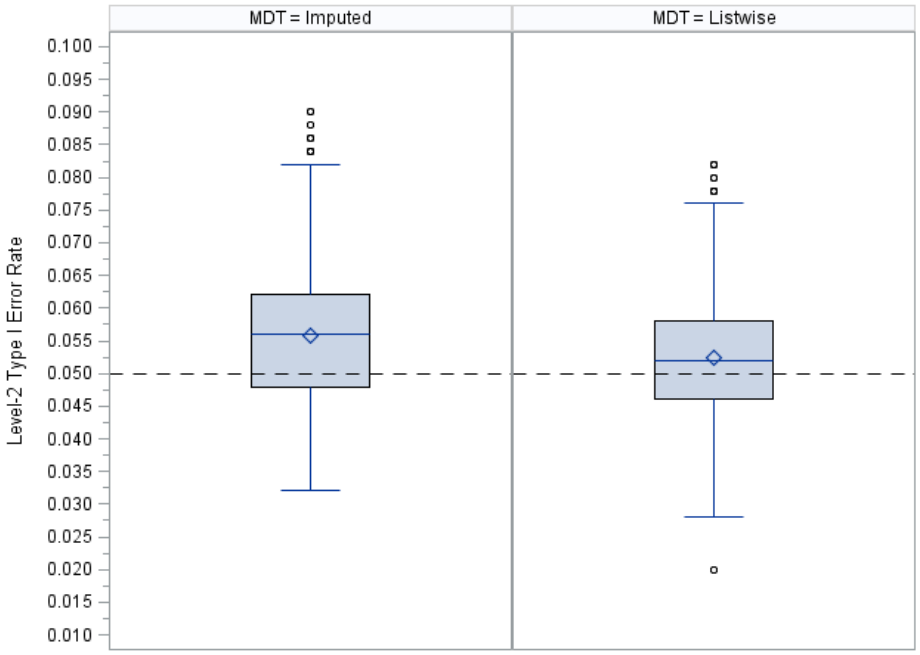

Figure 4.15 The Distribution of Level-2 Type I Error Rate by Missing Data Technique.

Overall Difference in Type I Error Rate. The average difference in Type I error from the complete data was -.002 (min = -.024, max =.018) indicting that the Type I error rate for the missing conditions was less than the Type I error rate of the complete cases. Important design factors for the difference in Type I error included level-1 missingness (η^2 = .779) and the interaction between MDT and level-1 missingness (η^2 = .043). Figure 4.16 shows that as level-1 missingness increased, the difference in Type I error rate decreased. For MLMI, mean differences were .000, -.001, -.003, -.007, and -.012 when level-1 missingness was 0%, 5%, 20%, 40%, and 70%, respectively. When data were listwise deleted, the mean differences in Type I error rates were .000, .000, -.003, -.009, and -.018 when level-1 missingness was 0%, 5%, 20%, 40%, and 70%, respectively. Listwise deletion provides a difference in Type I error rate similar to or lower than MLMI, and the difference between the two increased as level-1 missingness increased.

However, since MLMI was closest to the complete case, it seems that MLMI outperformed listwise deletion.

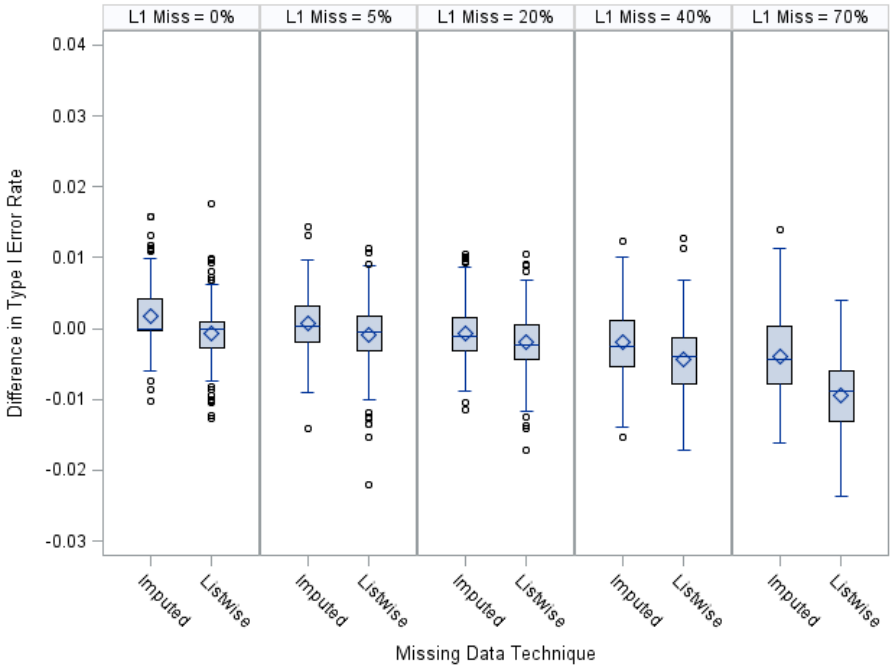

Figure 4.16 Distribution of the Difference in Type I Error Rate by Missing Data Technique and Level-1 Missingness.

Level-1 Difference in Type I Error Rate. At level-1, the mean difference in Type I error rate was -.006 (min = -.027, max =.007). Variability in the difference in Type I error rate was explained by level-1 missingness (η^2 = .805) and the interaction between MDT and level-1 missingness (η^2 = .043). Overall, the difference in Type I error from the missing case to the complete case was comparable for MLMI and listwise deletion until level-1 missingness was 40% or higher (see Figure 4.17). Once missingness was 40%, listwise deletion resulted in a further decrease in Type I error rate relative to the complete case than MLMI, to a difference as much as -.011 for MLMI and -.018 for listwise deletion.

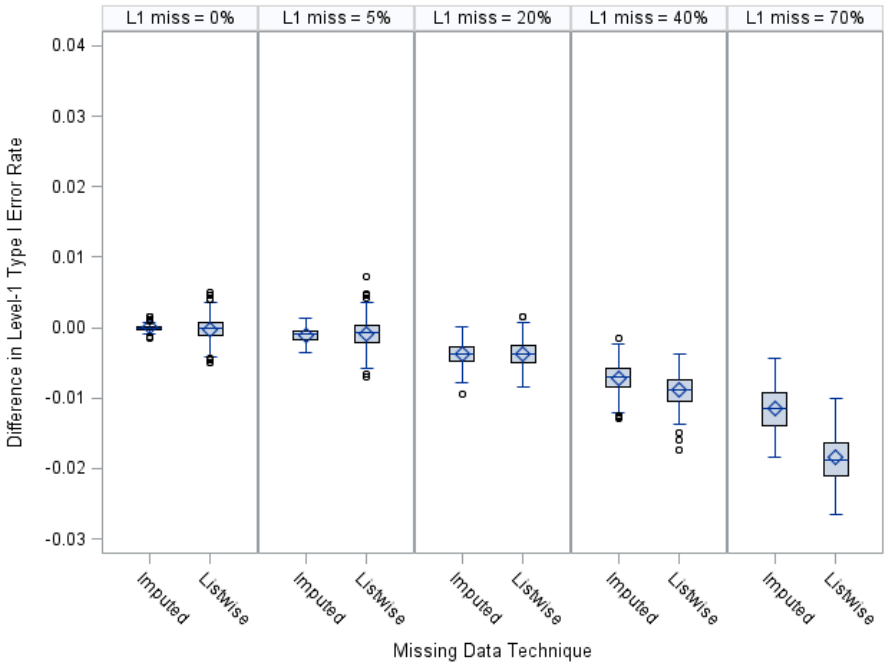

Figure 4.17 Distribution of the Level-1 Difference in Type I Error Rate by Missing Data Technique and Level-1 Missingness.

Level-2 Difference in Type I Error Rate. At level-2, the mean difference in Type I error rate was .001 (min = -.044, max =.036). The important design factor for difference in Type I error rate at level-2 included the three-way interaction between level-1 sample size, level-2 sample size, and level-1 missingness (η^2 = .061), the three-way interaction between level-2 sample size, level-1 missingness and level-2 missingness (η^2 = .057), the three-way interaction between level-1 sample size, level-1 missingness and level-2 missingness (η^2 = .051), and MDT (η^2 = .034), however the overall R-squared was only .488, suggesting that only 48% of the variance in difference in Type I error rate at level-2 was explained by the factors in this study, which is small for a simulation study. The small R-squared was a result of the lack of variability in differences in Type I error rate at level-2.

Figure 4.18 shows that Type I error rates at level-2 were similar to the Type I error rates of the complete data and were very close to the a priori established alpha of .05. While there are fluctuations in variability and values, no apparent trends seemed to exist across the levels of level-1 sample size, level-2 sample size, and level-1 missingness.

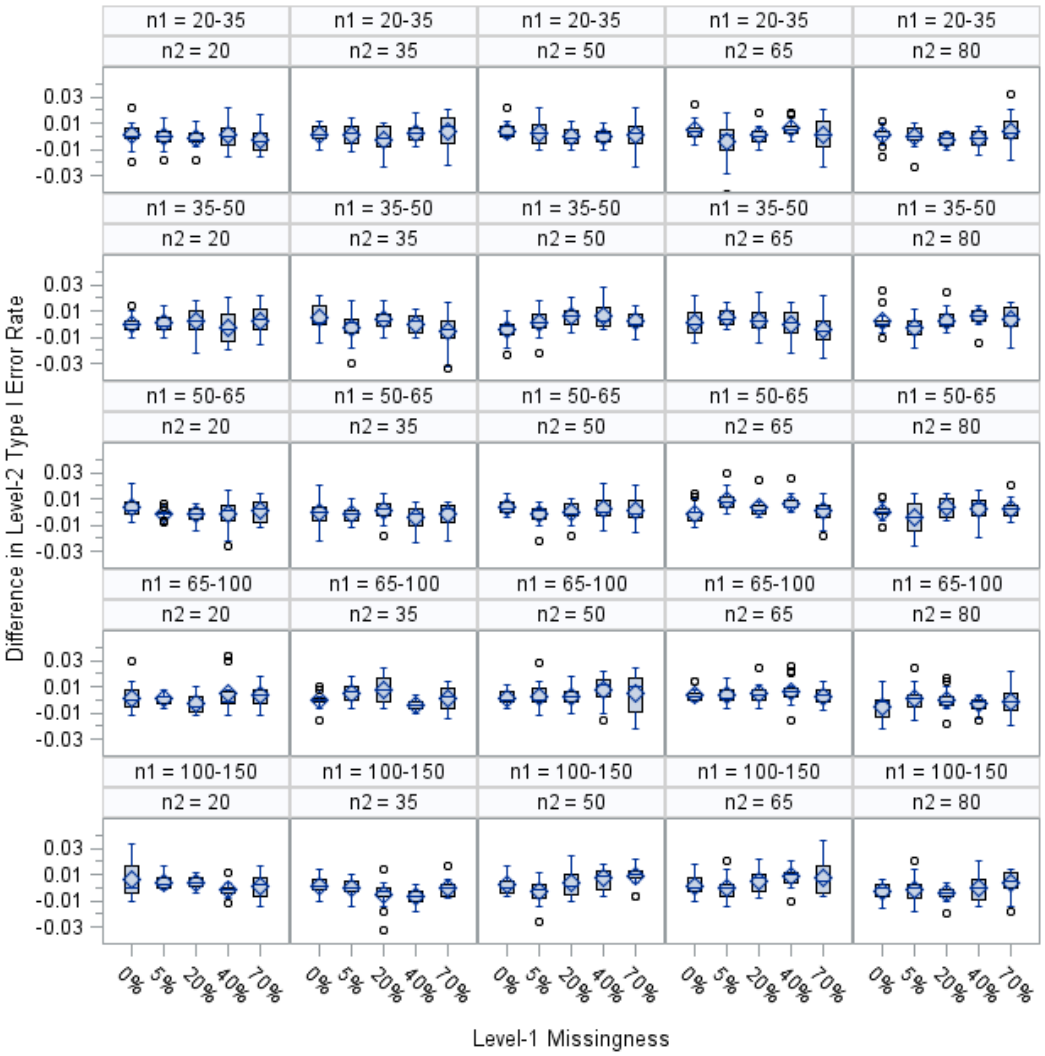

Figure 4.18 The Distribution of the Level-2 Difference in Type I Error Rates by Level-1 Sample Size, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.19 shows the three-way interaction between level-2 sample size, level-1 missingness and level-2 missingness. Overall, Type I error rates seem to be similar to the

complete case. While there are fluctuations in values and variability across these design factors, no apparent trends seem to exist.

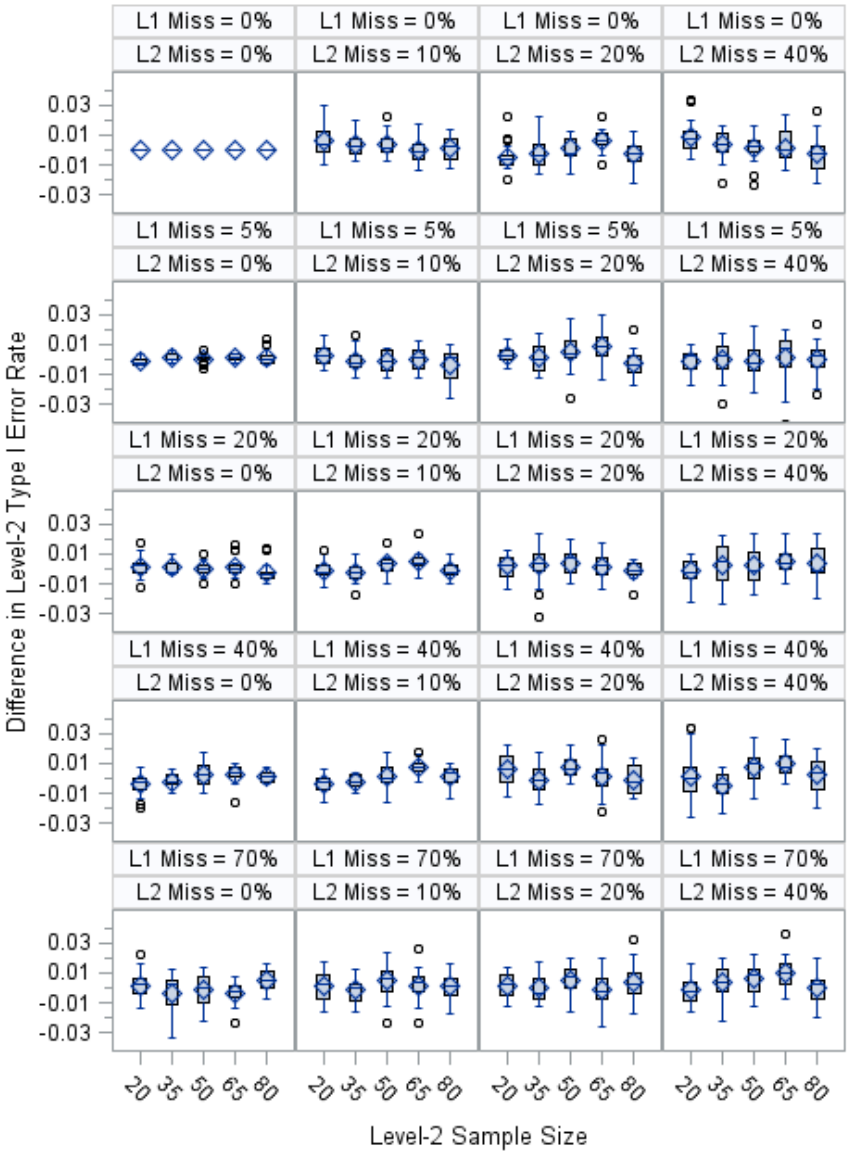

Figure 4.19 The Distribution of the Difference in Level-2 Type I Error Rate by Level-2 Sample Size, Level-1 Missingness, and Level-2 Missingness.

Figure 4.20 shows the three-way interaction between level-1 sample size, level-1 missingness and level-2 missingness. Similar to the previous two graphs, Type I error rates seem to be similar to the complete case. While there are fluctuations in values and variability across these design factors, no apparent trends seem to exist.

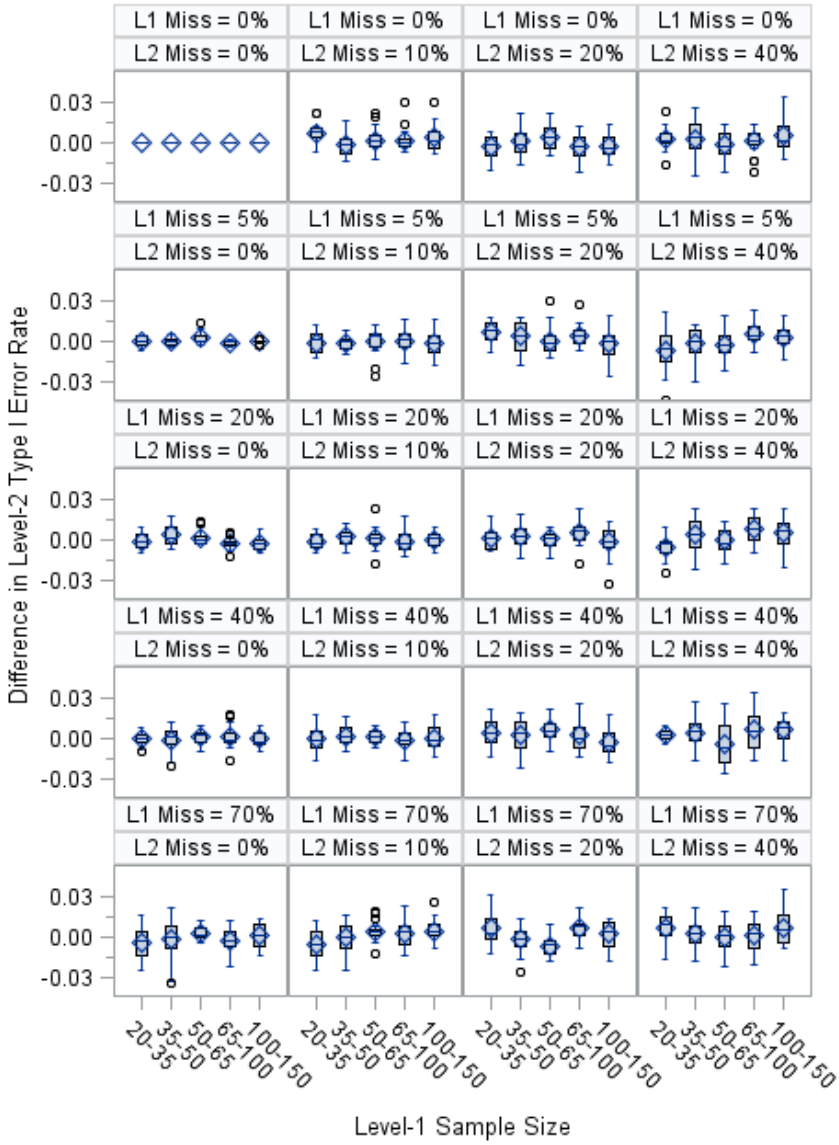

Figure 4.20 The Distribution of the Level-2 Difference in Type I Error Rate by Level-1 Sample Size, Level-1 Missingness, and Level-2 Missingness.

Figure 4.21 shows that while listwise deletion and MLMI had similar variability, on average level-2 Type I error rates were slightly higher for MLMI (with an average difference in Type I error rate of .003) compared to listwise (which had an average difference in Type I error rate of .000) providing evidence that MLMI slightly increases Type I error rate over the complete case.

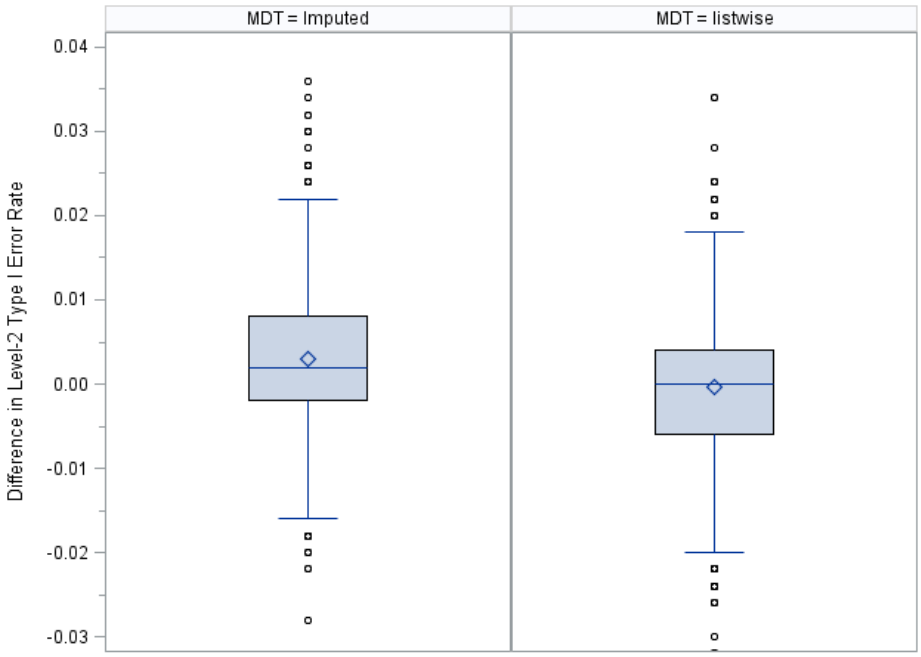

Figure 4.21 The Distribution of the Level-2 Difference in Type I Error Rate at Level-2 by Missing Data Technique.

Power

Overall Power. The mean overall power was .915 (min = .531, max = 1.000). Noteworthy amounts of variance were explained by level-2 sample size (η^2 = .879), level-2 missingness (η^2 = .058), MDT (η^2 = .019), the interaction between MDT and level-2 missingness (η^2 = .014), the interaction between level-2 sample size and level-2 missingness (η^2 = .013), and the interaction between MDT and level-2 sample size (η^2 = .004). Because the interactions encompass all of the main effects, the latter three interactions are summarized.

Figure 4.22 depicts the distribution of power by the interaction between MDT and level-2 missingness. As level-2 missingness increased, overall power levels decreased, however the magnitude of the decrease depended on the MDT used. Specifically, MLMI resulted in power of .939, .932, .927, and .909 when level-2 missingness was 0%, 10%,

20%, and 40%, respectively. Regardless of missingness, MLMI resulted in power that was always above the nominal level of .80. When listwise deletion was used, power was .934, .921, .903, and .851 when level-2 missingness was 0%, 10%, 20%, and 40%, respectively. Thus, while overall power did decrease as missingness increased, power never fell below the .80 nominal threshold.

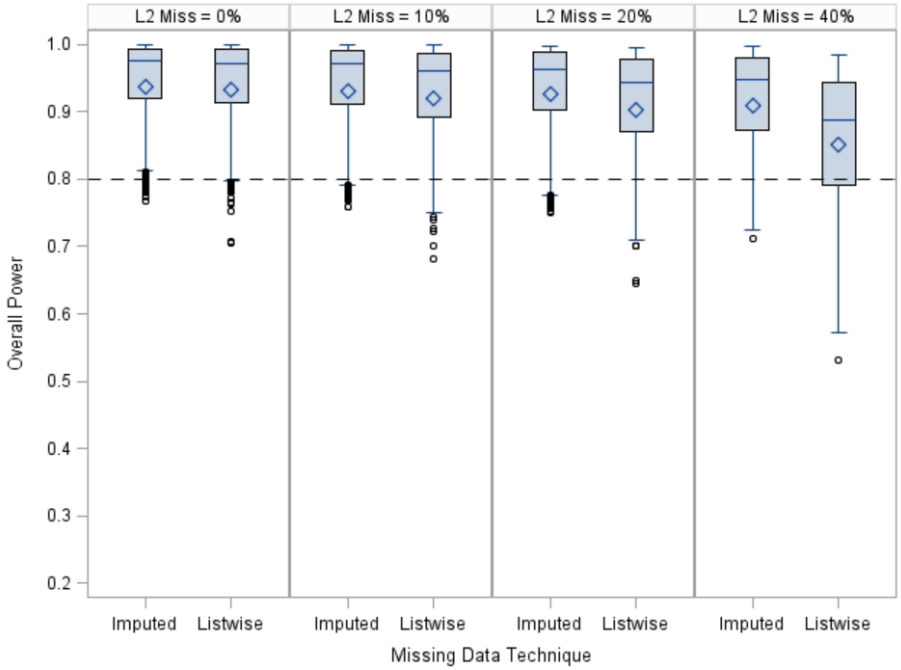

Figure 4.22 The Distribution of Overall Power by Missing Data Technique and Level-2 Missingness.

Figure 4.23 shows that when level-2 missingness was present, power decreased across all level-2 sample sizes. However, power stayed above the .80 threshold when level-2 sample size was 35, 50, 65, and 80 despite the level of missingness present. Only when level-2 sample size was 20 did power fall below the nominal .80 level, with power of .798, .774, .754, and .705 when missingness was 0%, 10%, 20% and 40% respectively.

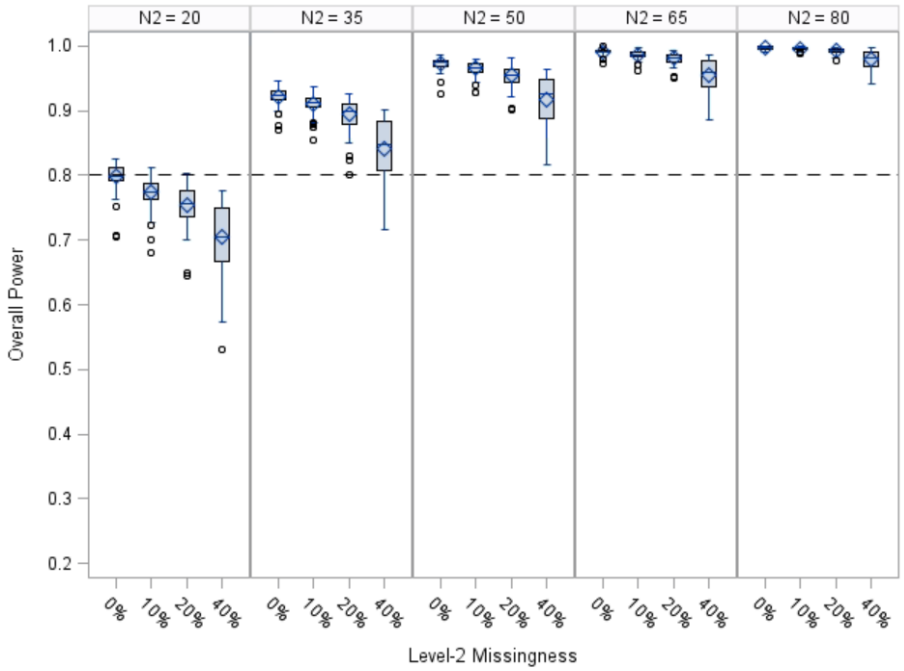

Figure 4.23 Distribution of Overall Power by Level-2 Sample Size and Level-2 Missingness.

Figure 4.24 shows the distribution of power by the interaction between MDT and level-2 sample size. In most cases, when listwise deletion or MLMI was used, power values were above the .80 nominal level, with listwise deletion being slightly lower than MLMI when sample size was 30 or higher. When level-2 sample size was 20, however, both MLMI and listwise deletion were below the nominal level with values of .778 for MLMI and .738 for listwise deletion.

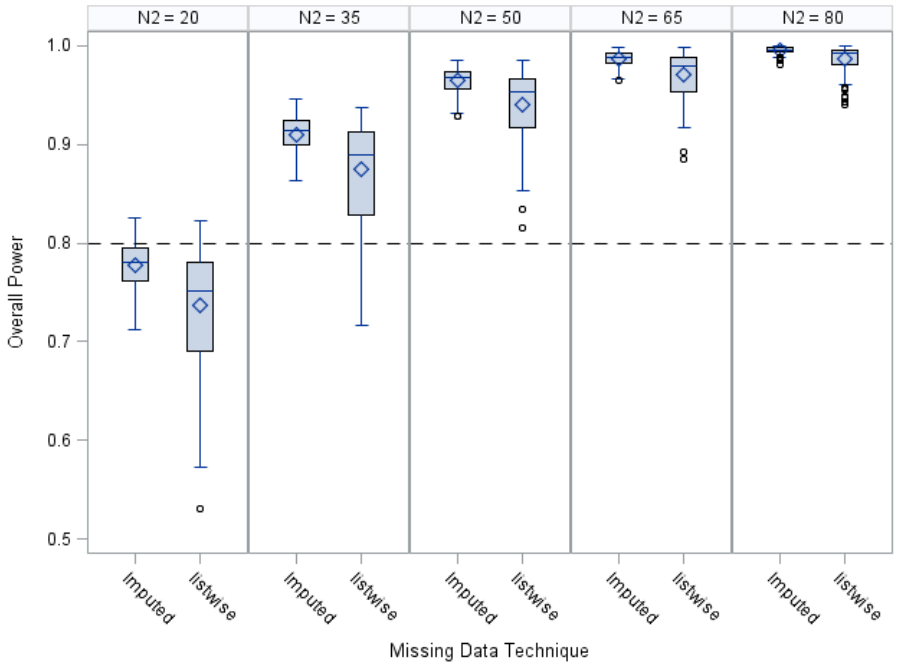

Figure 4.24 The Distribution of Overall Power by Missing Data Technique and Level-2 Sample Size.

Level-1 Power. The average value for power at level-1 was .999 (min = .862, max = 1.000). For power at level-1, the most important design factors were the interaction between level-1 sample size, level-2 sample size, and level-1 missingness (η^2 = .112) and the interaction between MDT, level-2 sample size, and level-1 missingness (η^2 = .066). Other design factors of interest are the interaction between level-2 sample size, level-1 missingness, and level-2 missingness (η^2 = .036), the interaction between MDT, level-2 sample size, and level-2 missingness (η^2 = .029), and the interaction between MDT, level-1 sample size, and level-1 missingness (η^2 = .026).

Figure 4.25 shows that power values at level-1 across all levels of level-1 sample size, level-2 sample size, and level-1 missingness were still far above the .80 recommended threshold. When level-1 and level-2 sample sizes were small (i.e., level-1 sample size was 20-35 and 35-50 and level-2 sample size was 20) and level-1

missingness was high (i.e., 70%) power did slightly decrease to .960. Thus, level-1 power is preserved even when sample size is small and percent of missingness is large.

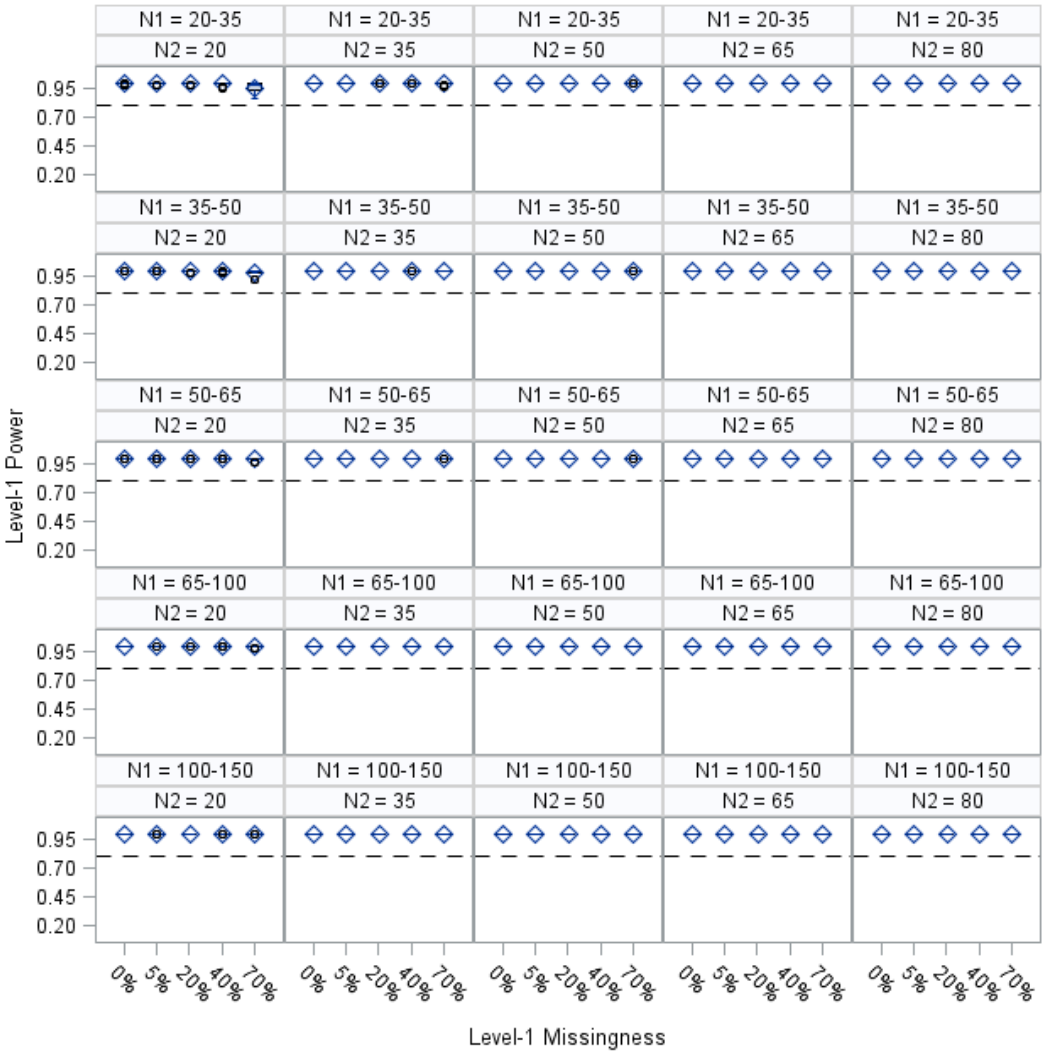

Figure 4.25 Distribution of the Level-1 Power by Level-1 Sample Size, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.26 shows that regardless of MDT, level-2 sample size, and level-1 missingness, power was higher than the nominal power of .80 with a mean across all design factor combination of .999. When level-2 sample size was small (i.e., 20) and level-1 missingness was high (i.e., 70%) values for listwise deletion did slightly decrease to .976, while MLMI remained at .998. Overall, however, both MDTs performed well.

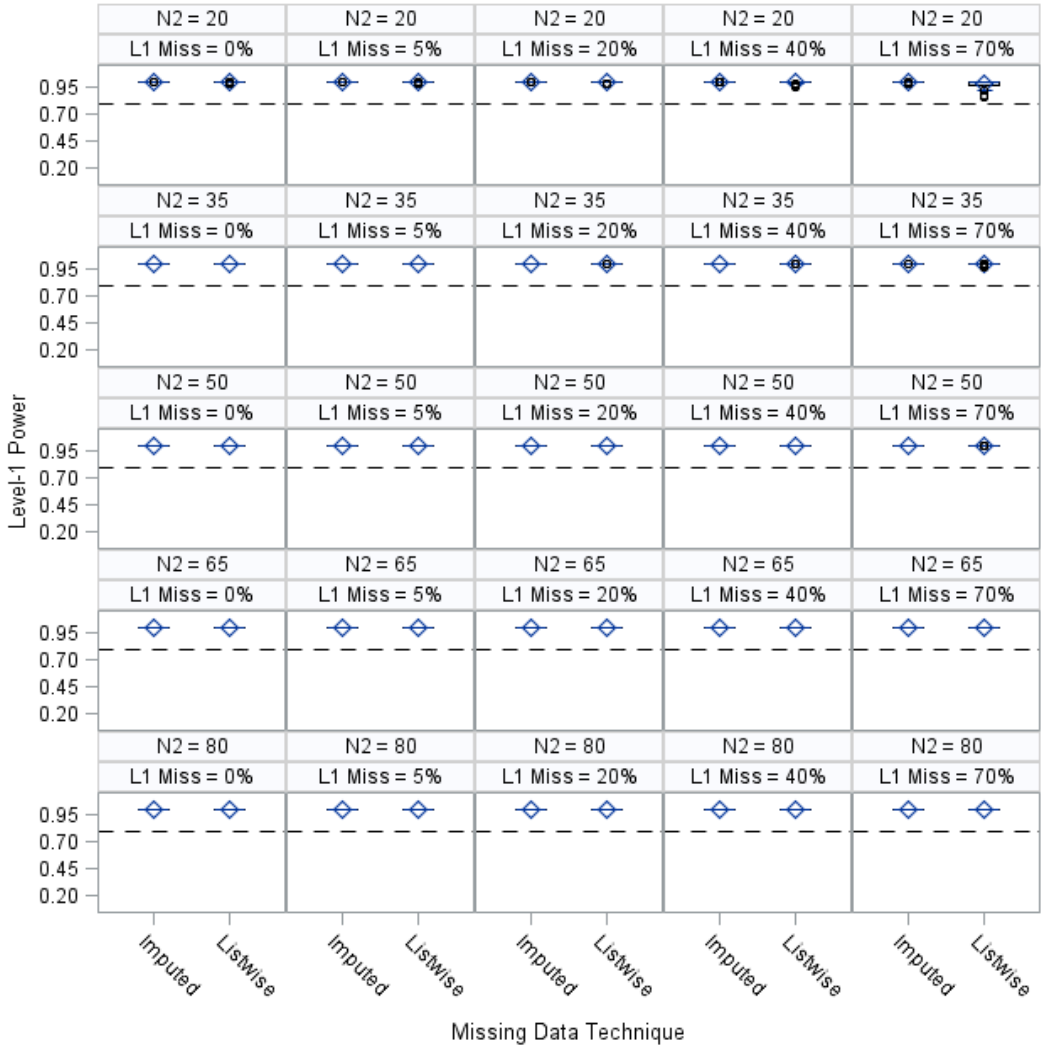

Figure 4.26 Distribution of Level-1 Power by Missing Data Technique, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.27 shows that regardless of level-2 sample size and missingness at both levels, level-1 power remained above the .80 nominal level. When level-1 missingness was high (i.e., 70%) and level-2 sample size was small (i.e., 20), power slightly decreased across all levels of level-2 missingness, however, none of the values fell below .860. Overall, power was retained across all levels of these design factors.

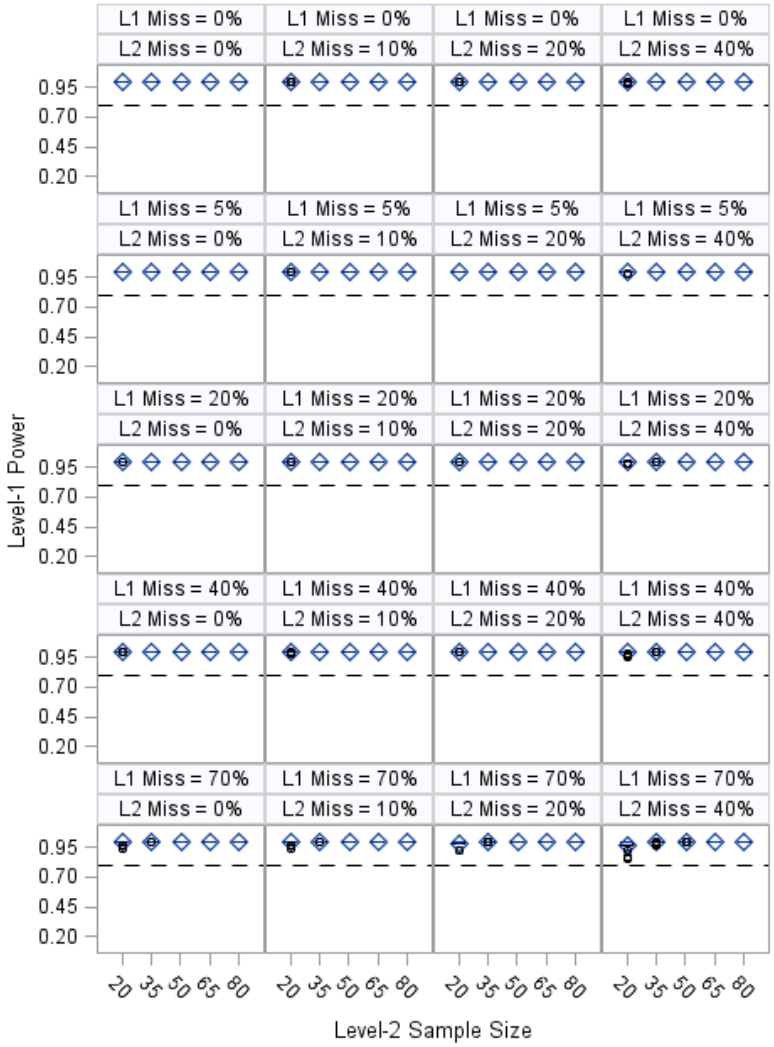

Figure 4.27 Distribution of Level-1 Power by Level-2 Sample Size, Level-1 Missingness, and Level-2 Missingness.

Figure 4.28 shows that both of the MDTs performed well regardless of level-2 sample size and level-2 missingness. Level-1 power slightly decreased when listwise deletion was used, level-2 sample size was small (i.e., 20), and level-2 missingness was large (i.e., 40%), however power only fell to .984. Across all combinations of these design factors, level-1 power was above the nominal .80 threshold, showing that both MDTs performed well.

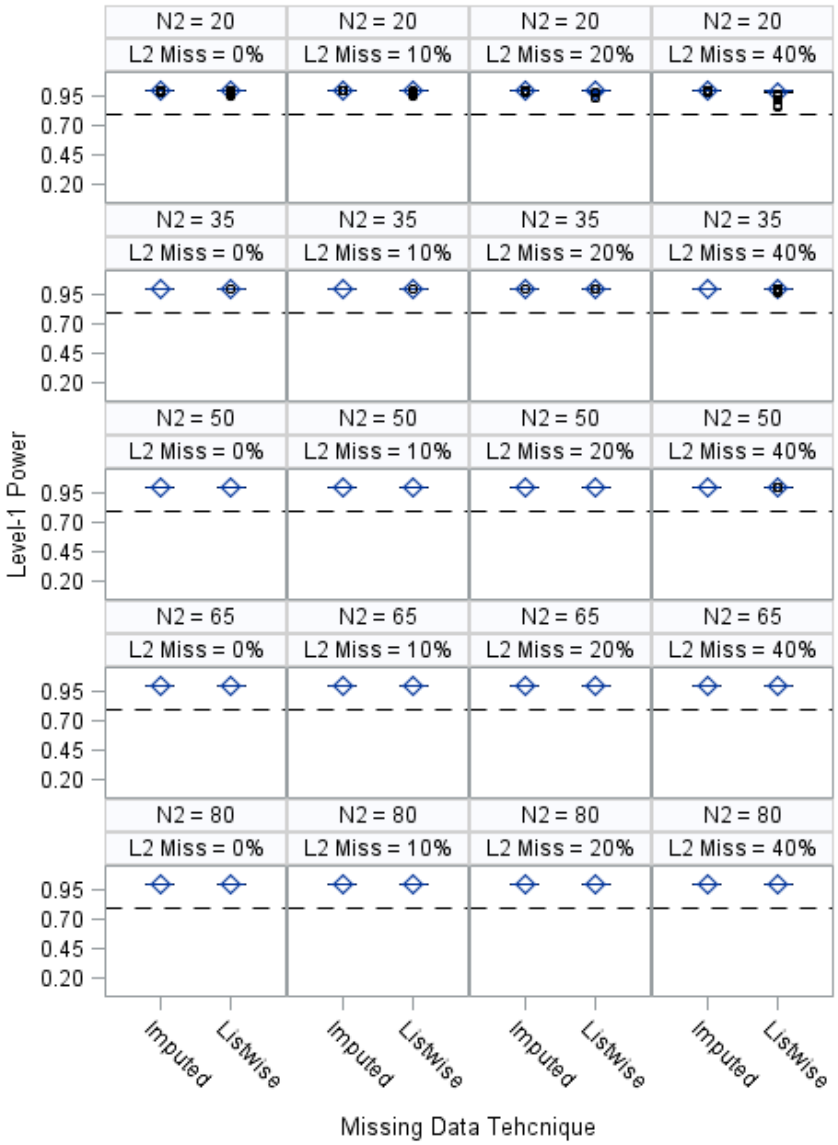

Figure 4.28 Distribution of Level-1 Power by MDT, Level-2 Sample Size, and Level-2 Missingness.

Lastly, Figure 4.29 shows that both MDTs performed well regardless of level-1 sample size and level-1 missingness. Power levels remained well above the .80 nominal level across all combinations of these design factors. A slight decrease did occur when level-1 sample sizes were smaller and level-1 missingness was large, however the minimum power value was .984, demonstrating that power was retained across both MDTs.

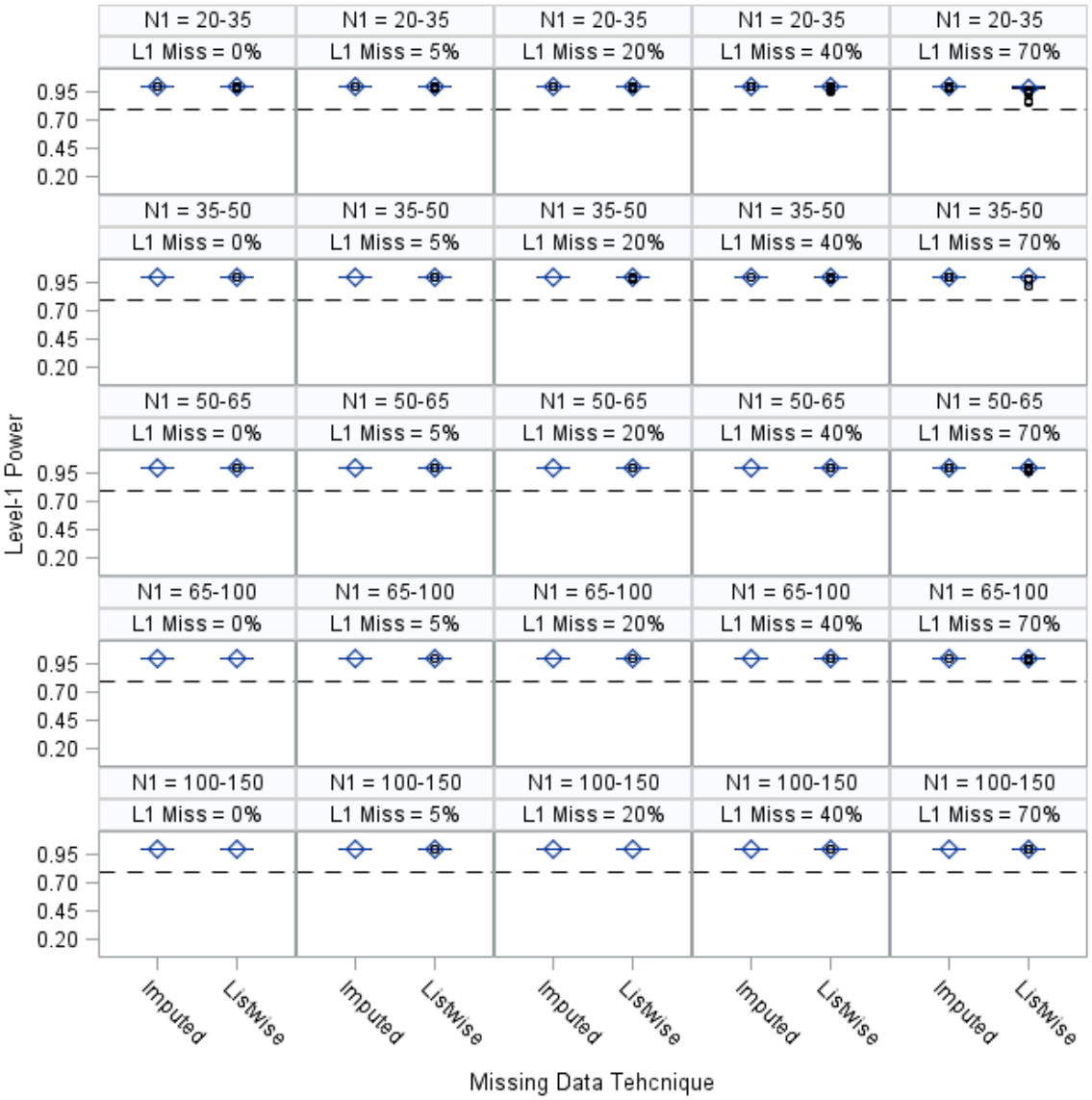

Figure 4.29 Distribution of Level-1 Power by Missing Data Technique, Level-1 Sample Size, and Level-1 Missingness.

Level-2 Power. Power at level-2 had a mean of .830 (min = .196, max = 1.000). For power at level-2, the most important design factors were level-2 sample size (η^2 = .884), level-2 missingness (η^2 =.058), MDT (η^2 =.019), the interaction between MDT and level-2 missingness (η^2 = .013), and the interaction between level-2 sample size and level-2 missingness (η^2 = .012), the interaction between MDT and level-2 sample size (η^2 = .003), the three-way interaction between MDT, level-2 sample size, and level-2

missingness (η^2 = .001). Since the three-way interaction encompasses all of the main effects and two-way interactions, only this interaction will be summarized.

Figure 4.30 depicts the three-way interaction of MDT, level-2 sample size, and level-2 missingness. Overall, as missingness increased, level-2 power decreased, and the decrease for listwise deletion was generally similar to or more extreme than MLMI. At a level-2 sample size of 20, power was below the .80 nominal level, even when there was no missingness. As missingness increased, power decreased regardless of which MDT was used, however power for listwise deletion was less than the power for MLMI, with power of .604, .554, and .497 at 0%, 10%, 20%, and 40% missingness, respectively, and power values for listwise deletion of .591, .527, .467, .340 at 0%, 10%, 20%, and 40% missingness, respectively.

When sample size was 35, power was above the .80 threshold when there was no missingness, and decreased as missingness increased. MLMI stayed at the .80 threshold or above when missingness was 20% or less with power values of .852, .839, .804, and .754 when missingness was 0%, 10%, 20%, and 40%, respectively. Listwise deletion resulted in power above .80 when missingness was 10% or less with power values of .838, .804, .754, and .604 when missingness was 0%, 10%, 20%, and 40%, respectively. When level-2 sample size was 50, and MLMI was used, power remained above the .80 nominal level across all levels of missingness with power values of .950, .942, .930, .896 at 0%, 10%, 20%, and 40% missingness, respectively. For listwise deletion, power was above the .80 nominal level when level-2 missingness was 20% or less, with power values of .942, .920, .886, and .775 at 0%, 10%, 20%, and 40% missingness, respectively. When level-2 sample size was 65 or higher, power levels across both MDTs and levels of

missingness were above the .80 level. When level-2 sample size was 65, power for MLMI was .985, .978, .973, and .954 when missingness was 0%, 10%, 20%, and 40%, respectively, while power for listwise deletion was .981, .968, .951, and .871 at 0%, 10%, 20%, and 40%, respectively. Lastly, when level-2 sample size was 80, power for MLMI was .995, .994, .990, and .983 at a missingness of 0%, 10%, 20%, and 40%, respectively, while power for listwise deletion was .993, .990, .980, and .933 at 0%, 10%, 20%, and 40% missingness, respectively.

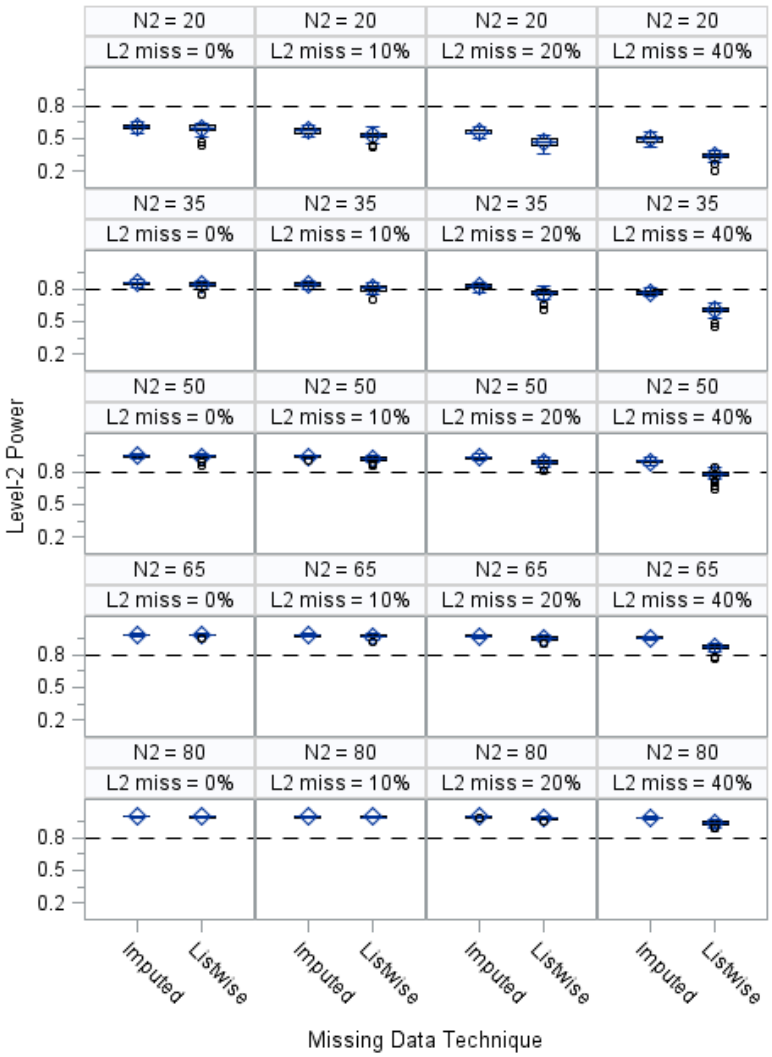

Figure 4.30 Distribution of Level-2 Power by Missing Data Technique, Level-2 Sample Size, and Level-2 Missingness.

Difference in Overall Power. The mean difference in power was -.023 (min = -.266, max = .009). Noteworthy amounts of variance was explained by, level-2 missingness (η^2 = .421), level-2 sample size (η^2 = .150), MDT (η^2 = .138), the interaction between MDT and level-2 missingness (η^2 = .098), the interaction between level-2 sample size and level-2 missingness (η^2 = .092), and the interaction between MDT and level-2 sample size (η^2 = .028). Because the interactions encompass all of the main effects, the latter three interactions will be summarized.

Figure 4.31 shows that listwise deletion and MLMI were both similar to the complete case when level-2 missingness was 0%, with a value of .000 for MLMI and a value of -.004 for listwise deletion. As the level of missingness increased, both listwise deletion and MLMI had less power than the complete case. The difference in power was -.005, -.011, and -.029 for MLMI at 10%, 20% and 40% level-2 missingness, respectively and the difference in power was -.017, -.035, and -.088 for listwise deletion at 10%. 20% and 40% level-2 missingness, respectively. However, as stated previously, for the most part, overall power was at or above .80 for both MLMI and listwise deletion.

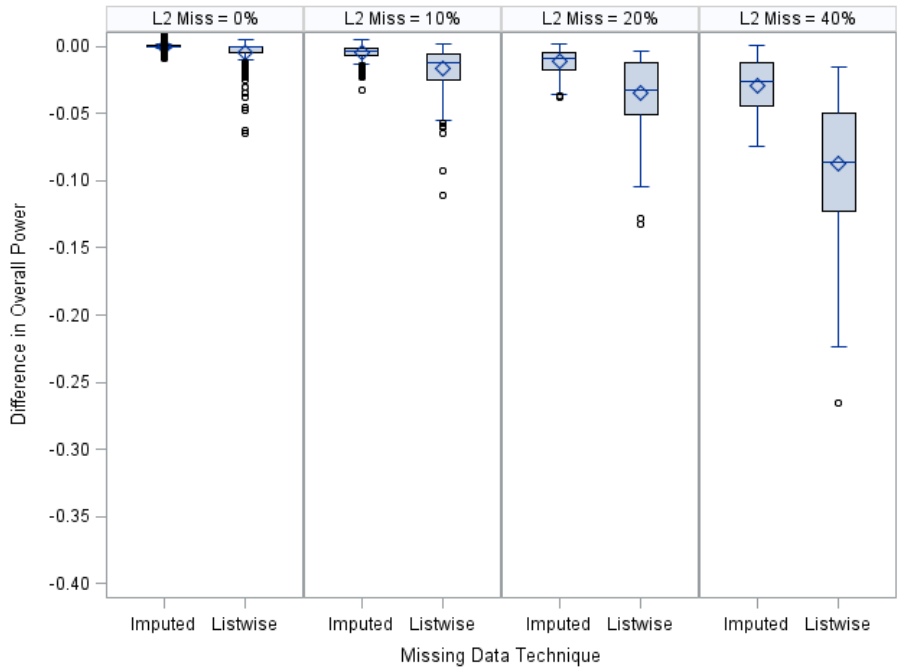

Figure 4.31 The Distribution of the Difference in Overall Power by Missing Data Technique and Level-2 Missingness.

Figure 4.32 shows that compared to the complete data power decreased as level-2 missingness increased. However, as level-2 sample size increased, the magnitude of the difference in power decreased. Specifically, when level-2 sample size was 20, power decreased by .004, .024, .045, and .094 when level-2 missingness was 0%, 10%, 20%, and 40%, respectively. When level-2 sample size was 35, power decreased by .003, .014, .033, and .086 when level-2 missingness was 0%, 10%, 20%, and 40%, respectively. When level-2 sample size was 50, power decreased by .001, .008, .022, and .057 when level-2 missingness was 0%, 10%, 20%, and 40% respectively. When level-2 sample size was 65, power decreased by .000, .005, .010, and .035 when level-2 missingness was 0%, 10%, 20%, and 40%, respectively. Lastly, when sample size was 80, power slightly decreased by .000, .002, .005, and .019.

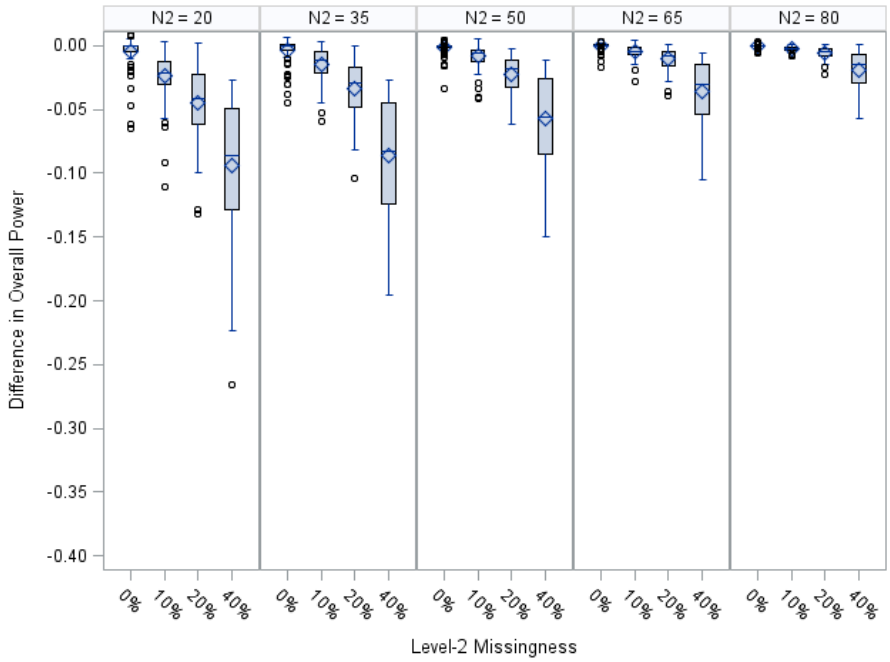

Figure 4.32 The Distribution of the Difference in Overall Power by Level-2 Sample Size and Level-2 Missingness.

Figure 4.33 shows that power decreased when level-2 sample size was 20 for both MLMI and listwise deletion, however the decrease was more severe when listwise deletion was used, where the difference in power was -.021 for MLMI and -.062 for listwise deletion. As sample size increased, power levels were more similar to the complete case. Specifically, when level-2 sample size was 35, MLMI resulted in a decrease of .017 and listwise deletion resulted in a decrease of .052. When level-2 sample size was 50, MLMI resulted in a decrease of .005 and listwise deletion resulted in a decrease of .03. When sample size was 65, MLMI resulted in a decrease of .005 and listwise resulted in a decrease of .020. Lastly, when level-2 sample size was 80, MLMI resulted in a decrease in power of .002 and listwise deletion resulted in a decrease in .011.

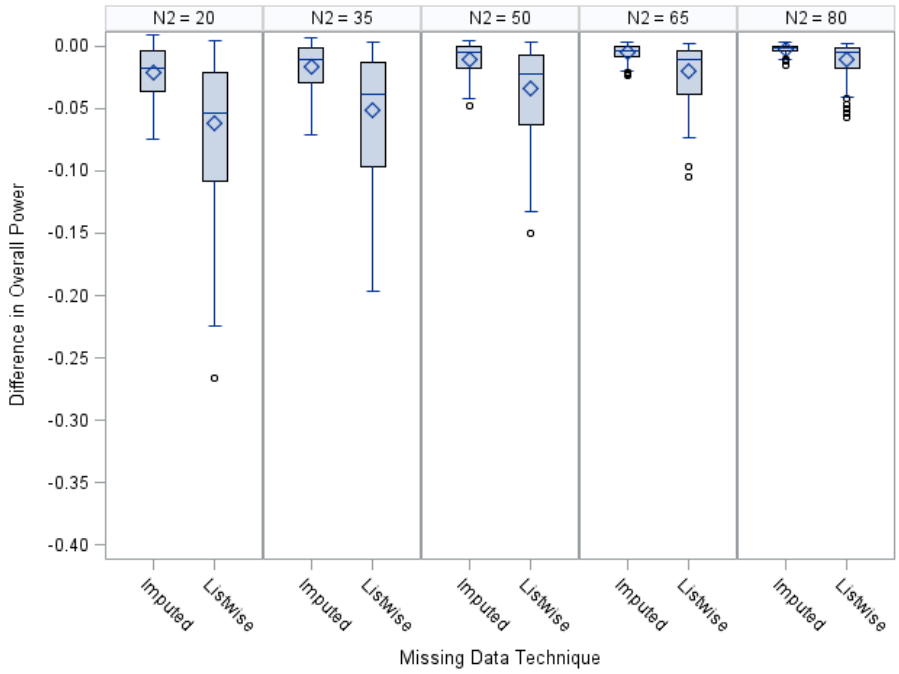

Figure 4.33 The Distribution of the Difference in Power by Missing Data Technique and Level-2 Sample Size.

Difference in Level-1 Power. The mean difference in power between the missing and complete case was close to zero for level-1 (M = 0.000, SD = .006, min = -0.138, max = .002). For power at level-1, the most important design factor was the interaction between level-1 sample size, level-2 sample size, and level-1 missingness (η^2 = .112), the interaction between MDT, level-2 sample size, and level-1 missingness (η^2 = .066), other design factors of interest were the interaction between level-2 sample size, level-1 missingness, and level-2 missingness (η^2 = .036), the interaction between MDT, level-2 sample size, and level-2 missingness (η^2 = .029), and the interaction between MDT, level-1 sample size, and level-1 missingness (η^2 = .023).

As expected, there is a decrease in power as missingness increased. Figure 4.34 shows that level-1 power decreased by as much as .138 when level-1 and level-2 sample size were small (i.e., level-1 sample size between 20-35 and level-2 sample size of 20)

and level-1 missingness was high (i.e., 70%). However, when level-1 sample size or level-2 sample size were large, the percent of missing data at level-1 had no impact on level-1 power.

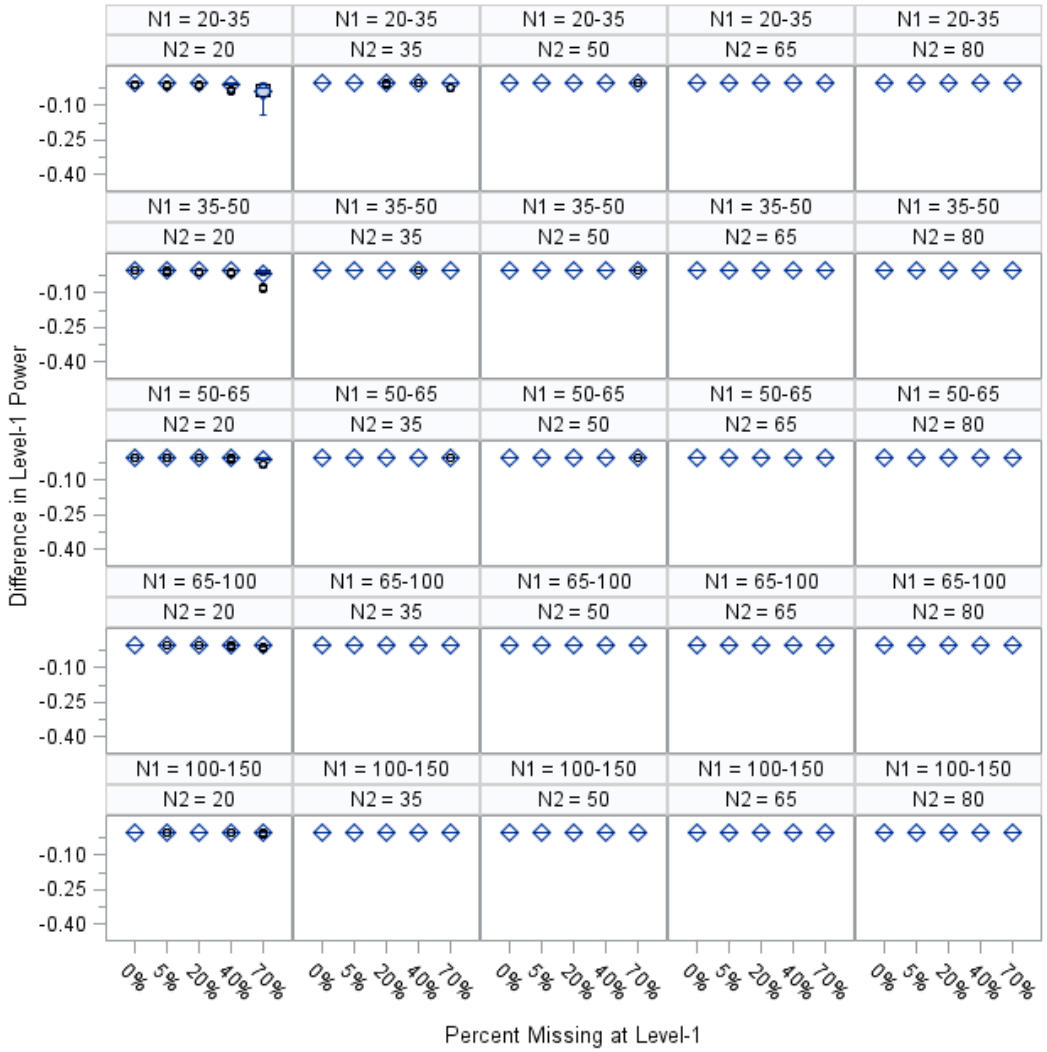

Figure 4.34 Distribution of the Difference in Level-1 Power by Level-1 Sample Size, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.35 also shows that power decreased when listwise was used, level-2 sample size was small, and level-1 missingness was large. Specifically, power decreased by -.024, but remained at .000 for every other combination of level-2 sample size. Power remained the same for MLMI across all combinations of design factors.

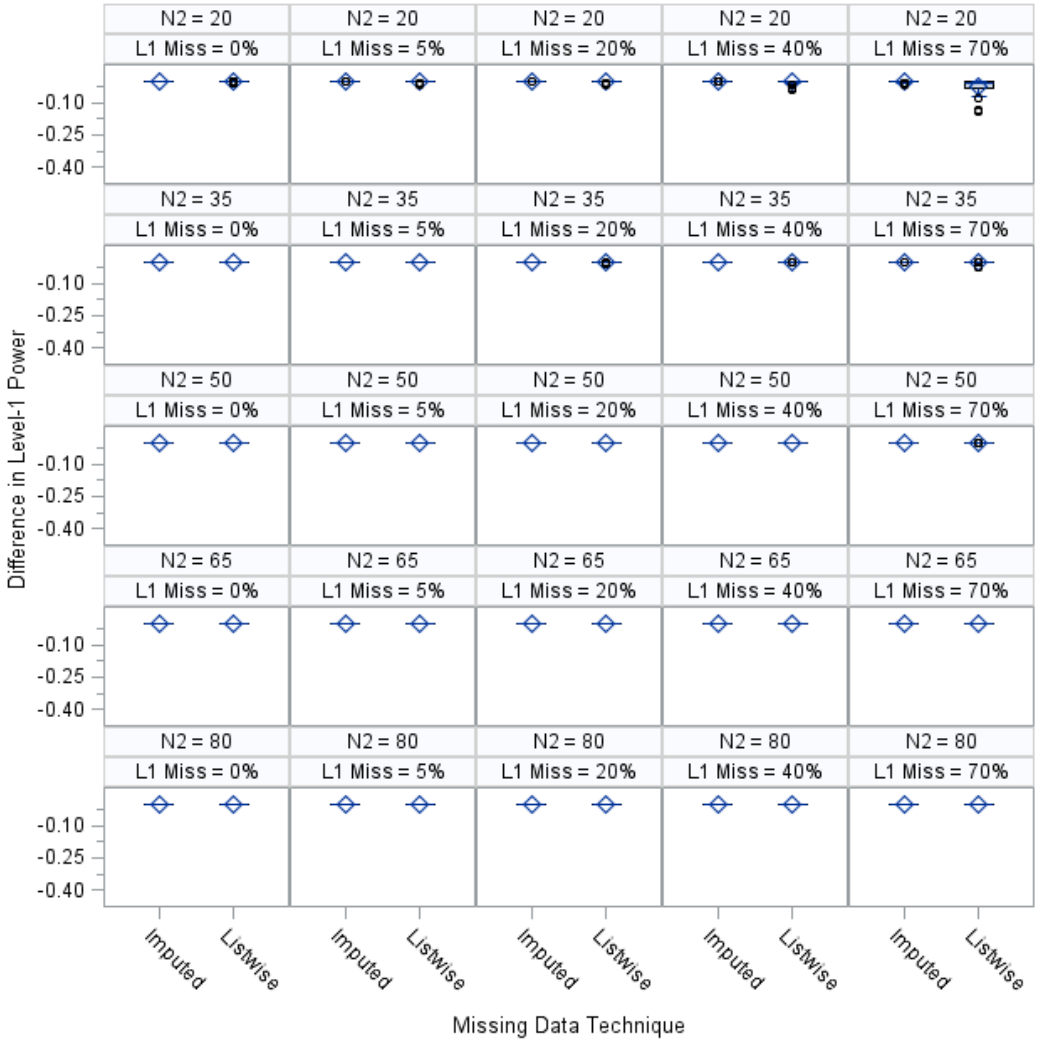

Figure 4.35 The Distribution of the Difference in Level-1 Power by Missing Data Technique, Level-2 Sample Size, and Level-1 Missingness.

Figure 4.36 shows that level-1 power also did not change drastically from the complete case, as the difference in power at level-1 stayed close to zero, especially when level-1 and level-2 missingness was small. Power slightly decreased when level-1 and level-2 missingness was large and level-2 sample size was small, with a maximum decrease of -.027.

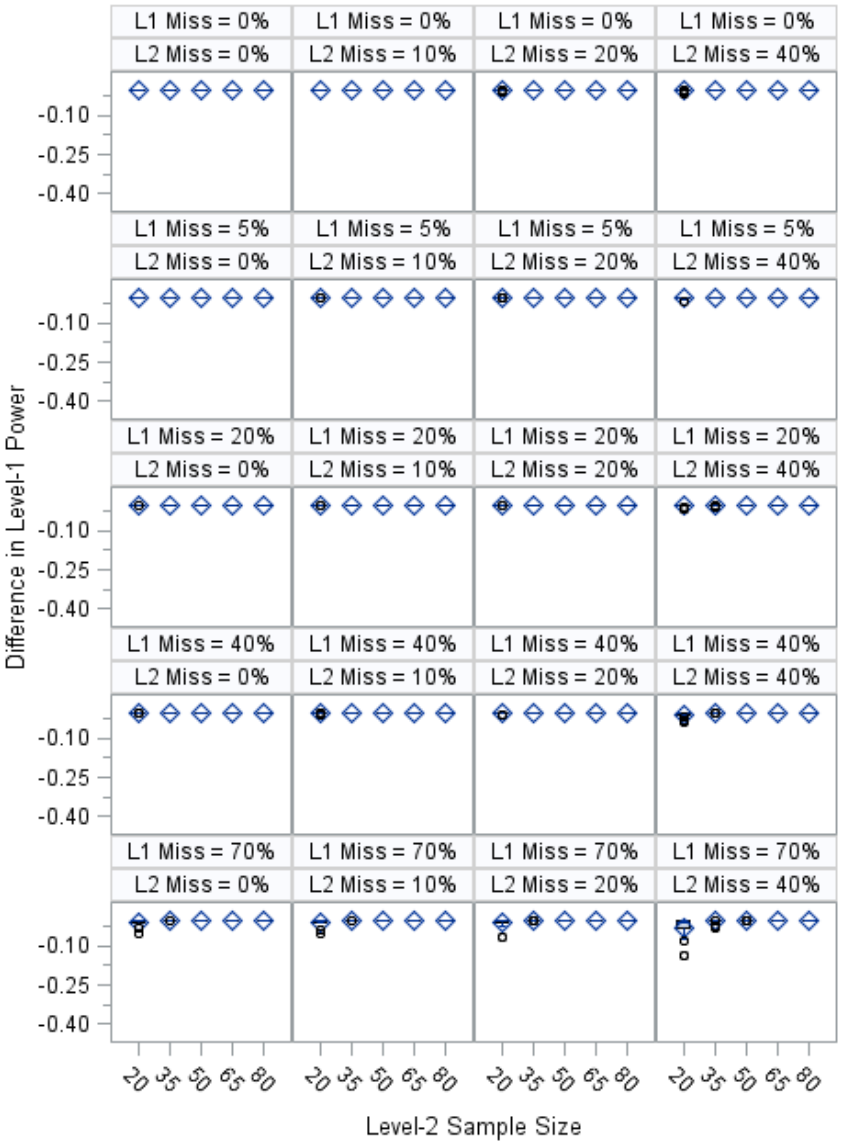

Figure 4.36 The Distribution of the Difference in Level-1 Power by Level-2 Sample Size Level-1 Missingness, Level-2 Missingness.

Figure 4.37 shows that power was similar to the complete case, especially when level-2 sample size was 35 and above. When sample size was 20, listwise deletion did slightly depart from the complete case, with a difference in power of -.002, -.003, -.005, and-.016 when level-2 missingness was 0%, 10%, 20%, and 40%, respectively. The average difference across MLMI was .000 across all level-2 sample size and level-2 missingness combinations.

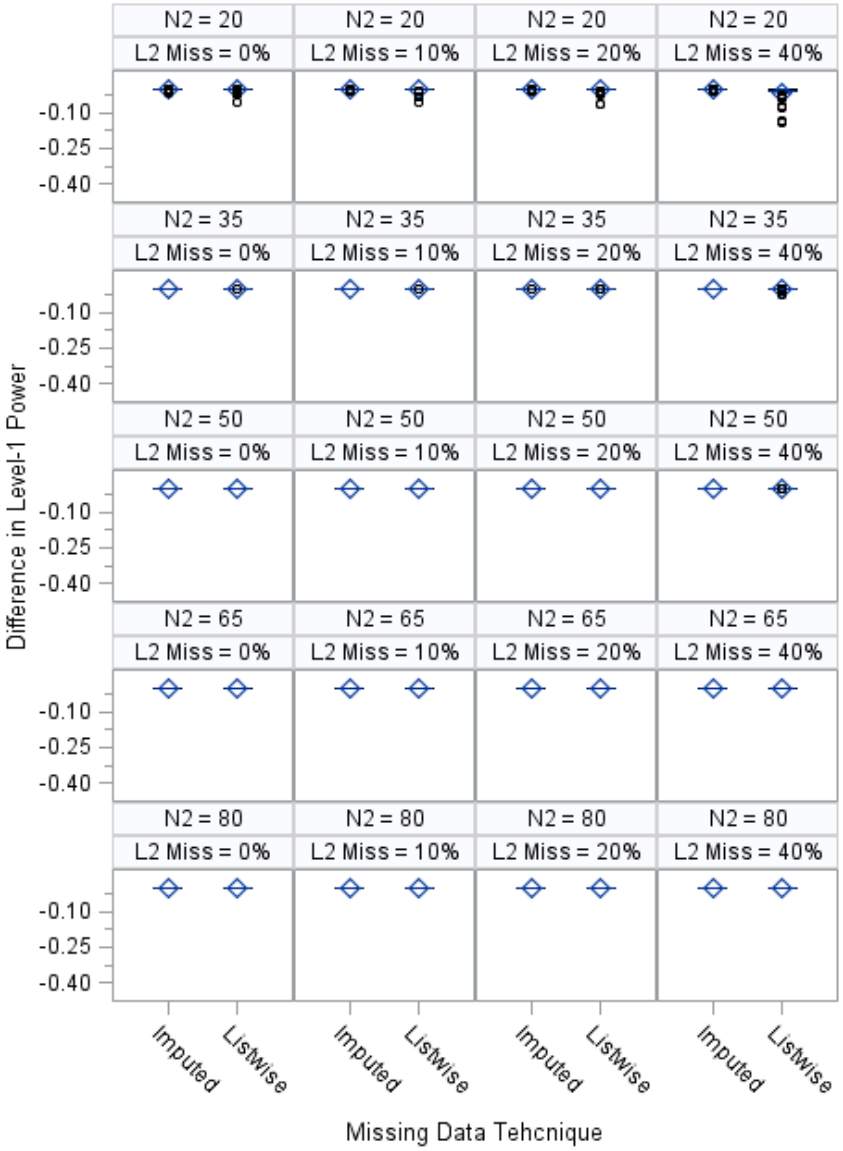

Figure 4.37 The Distribution of the Difference in Level-1 Power by Missing Data Technique, Level-2 Sample Size, and Level-2 Missingness.

Additionally, Figure 4.38 shows that while there was a slight difference from the complete case to the treated conditions, this only occurred when listwise deletion was used and level-1 missingness was 70%, with a difference in power of -.016, -.006, -.002, -.001, and -.000 when level-1 sample size was 20-35, 35-50, 50-65, 65-100, and 100-150, respectively.

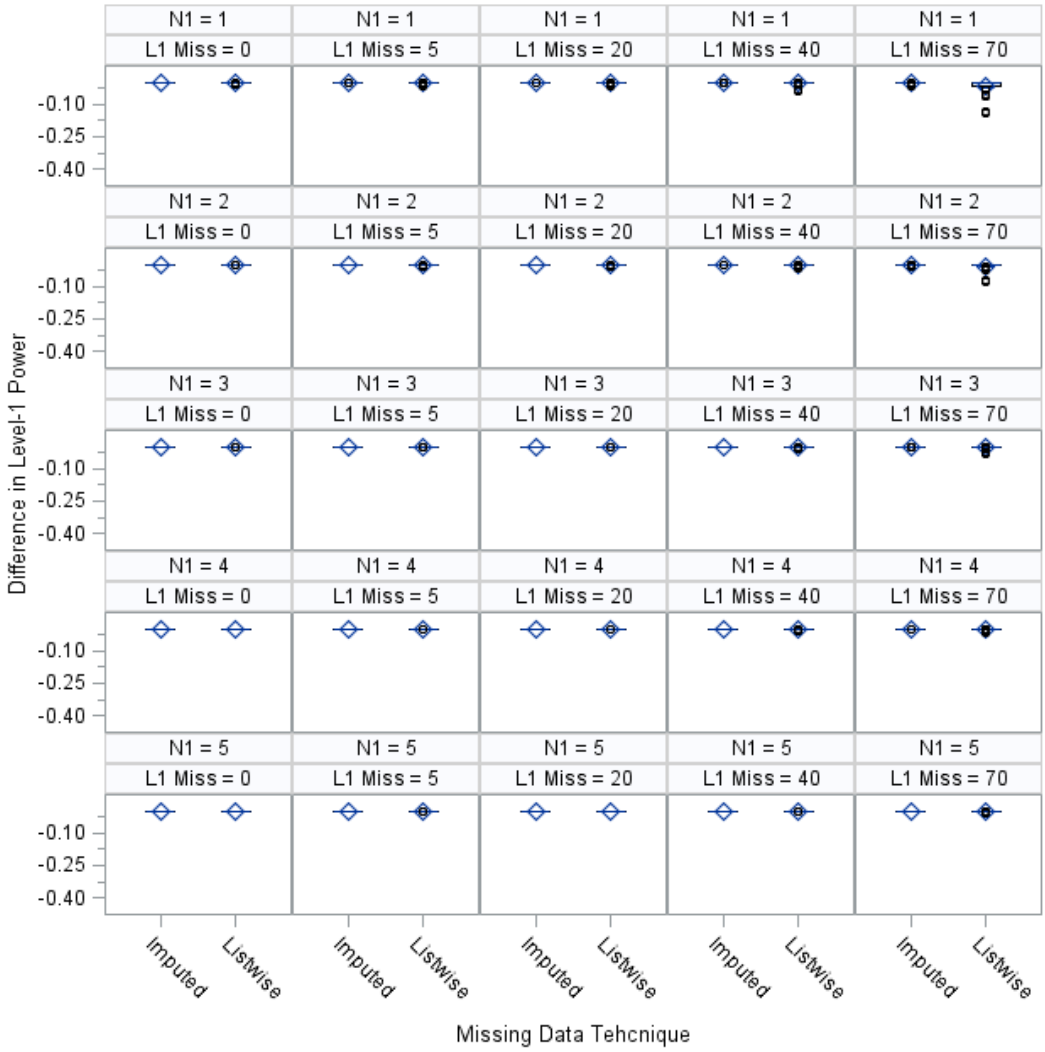

Figure 4.38 Distribution of the Difference in Level-1 Power by Missing Data Technique, Level-1 Sample Size, and Level-1 Missingness.

Difference in Level-2 Power. The mean difference in power between the missing and complete case was more pronounced for level-2 (M = .046, SD = .065, min = -.398, max = .024). For power at level-2, the most important design factors were level-2 missingness (η^2 =.439), level-2 sample size (η^2 = .148), MDT (η^2 =.140), the interaction between MDT and level-2 missingness (η^2 = .098), and the interaction between level-2 sample size and level-2 missingness (η^2 = .093), the interaction between MDT and level-2 sample size (η^2 = .025), and the three-way interaction between MDT, level-2 sample size,

and level-2 missingness (η^2 = .010). Since the three-way interaction encompasses all of the main effects and two-way interactions, only this interaction will be summarized.

Figure 4.39 shows the distribution of the difference in level-2 power by MDT, level-2 sample size and level-2 missingness. This figure shows that the difference between the complete case and the treated conditions increases as missingness increased, however as sample size becomes larger, the difference almost disappears. When level-2 sample size was 20, the difference was pronounced across all levels of missingness, with a difference in power for MLMI of .000, -.023, -.044, and -.101 at 0%, 10%, 20%, and 40% missingness, respectively. Listwise deletion at the same sample size produced a difference in power of -.013, -.068, -.129, and -.258 at 0%, 10%, 20%, and 40% missingness, respectively. However, when level-2 sample size was 80, MLMI produced a difference in power of .000, -.002, -.005, and -.013 when missingness was 0%, 10%, 20%, and 40%, respectively, while listwise deletion produced a difference in power of -.001, -.002, -.005, and -.013 at 0%, 10%, 20%, and 40% missingness, respectively.

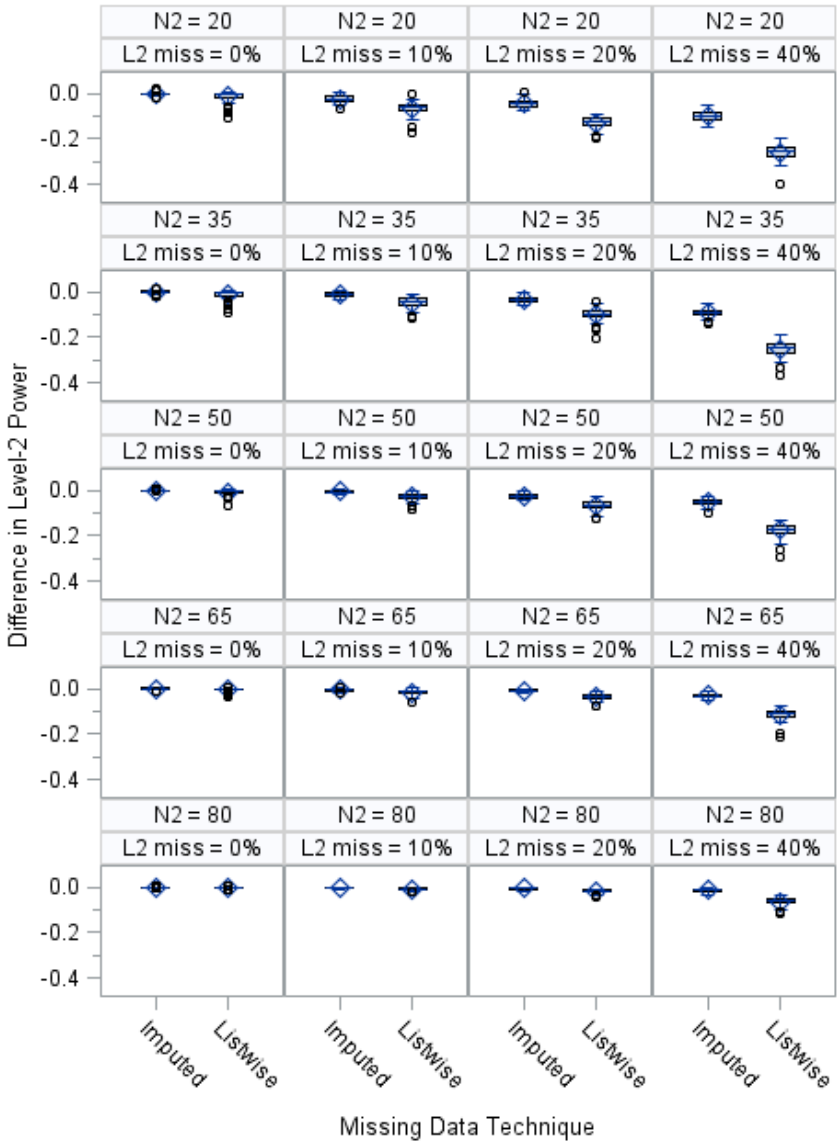

Figure 4.39 The Distribution of the Difference in Level-2 Power by Missing Data Technique, Level-2 Sample Size, and Level-2 Missingness.

Confidence Interval Coverage

Overall Confidence Interval Coverage. The overall mean confidence interval coverage (C.I. coverage) was .94 (min = .652 max = 1.00). A considerable amount of variance was explained by the interaction between MDT and level-1 missingness ($\eta 2$ = .302), level-1 missingness ($\eta 2$ = .253), level-1 sample size ($\eta 2$ = .079), level-2 missingness ($\eta 2$ = .063), MDT ($\eta 2$ = .060), the interaction between MDT and level-2

missingness (η2 = .045), the three-way interaction between MDT, level-1 sample size, and level-1 missingness (η2 = .032), and the interaction between MDT and level-2 sample size (η2 = .023). Because the latter three interactions encompass all of the main effects and interactions listed formerly, these three will be examined.

Figure 4.40 depicts the interaction between MDT and level-2 missingness. Overall, MLMI resulted in more variability in C.I. coverage values and much lower mean C.I. coverage values across all levels of level-2 missingness, with values of .917, .917, .916, .912 for 0%, 10%, 20%, and 40% level-2 missingness, respectively. Listwise deletion, however, resulted in higher C.I. coverage values when missingness was 20% or below, however values became similar to MLMI values when missingness was 40%, with mean C.I. coverage values of .997, .984, .957, and .903 for 0%, 10%, 20%, and 40% level-2 missingness.

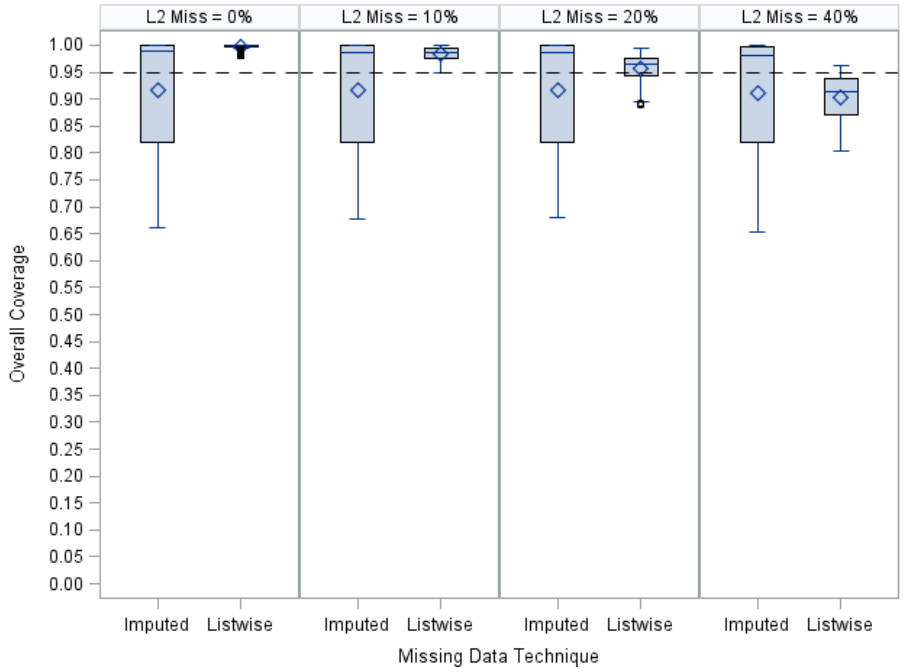

Figure 4.40 The Distribution of Confidence Interval Coverage by Missing Data Technique and Level-2 Missingness.

Figure 4.41 depicts the three-way interaction between MDT, level-1 sample size, and level-1 missingness. When sample size was 20-35, all C.I. coverage values for listwise deletion were above .95. For MLMI at the same sample size range, all C.I. coverage values were above .95, except for when missingness was 40% or greater, with C.I. coverage values of .939 and .848 when missingness was 40% and 70%, respectively. When sample size was 35-50, all values for listwise deletion were above .95. MLMI at the same sample size range falls below .95 when missingness was 40% or higher with the mean C.I. coverage value of .939 and .797 when missingness was 40% and 70%, respectively. When sample size was 50-65, again listwise deletion had C.I. coverage values above .95 across all levels of level-1 missingness. MLMI has C.I. coverage values above .95 as long as level-1 missingness was less than 40%, with C.I. coverage values falling to .838 and .771 with 40% and 70% missingness, respectively. When level-1 sample size was 65-100, C.I. coverage values for listwise deletion fell slightly below .95 for 0%-20% levels of missingness, but climbed slightly above .95 for 40% and 70% missingness with values of .946, .947, .944, .949, .955, and .959 for 0%, 5%, 20%, 40%, and 70% missingness, respectively. Again, at the same sample size level, MLMI had acceptable C.I. coverage values until missingness was 40% or higher, with values of .805 and .716 at 40% and 70% missing data, respectively. Lastly, when level-1 sample size was 100-150, listwise deletion was below the .95 level with values of .922, .922, .930, .937, and .948 with 0%, 5%, 20% 40%, and 70% missingness, respectively. At the same sample size range, MLMI had acceptable C.I. coverage values until missingness was 20% or higher with values of .888, .770, and .716 when missingness was 20%, 40%, and 70%, respectively.

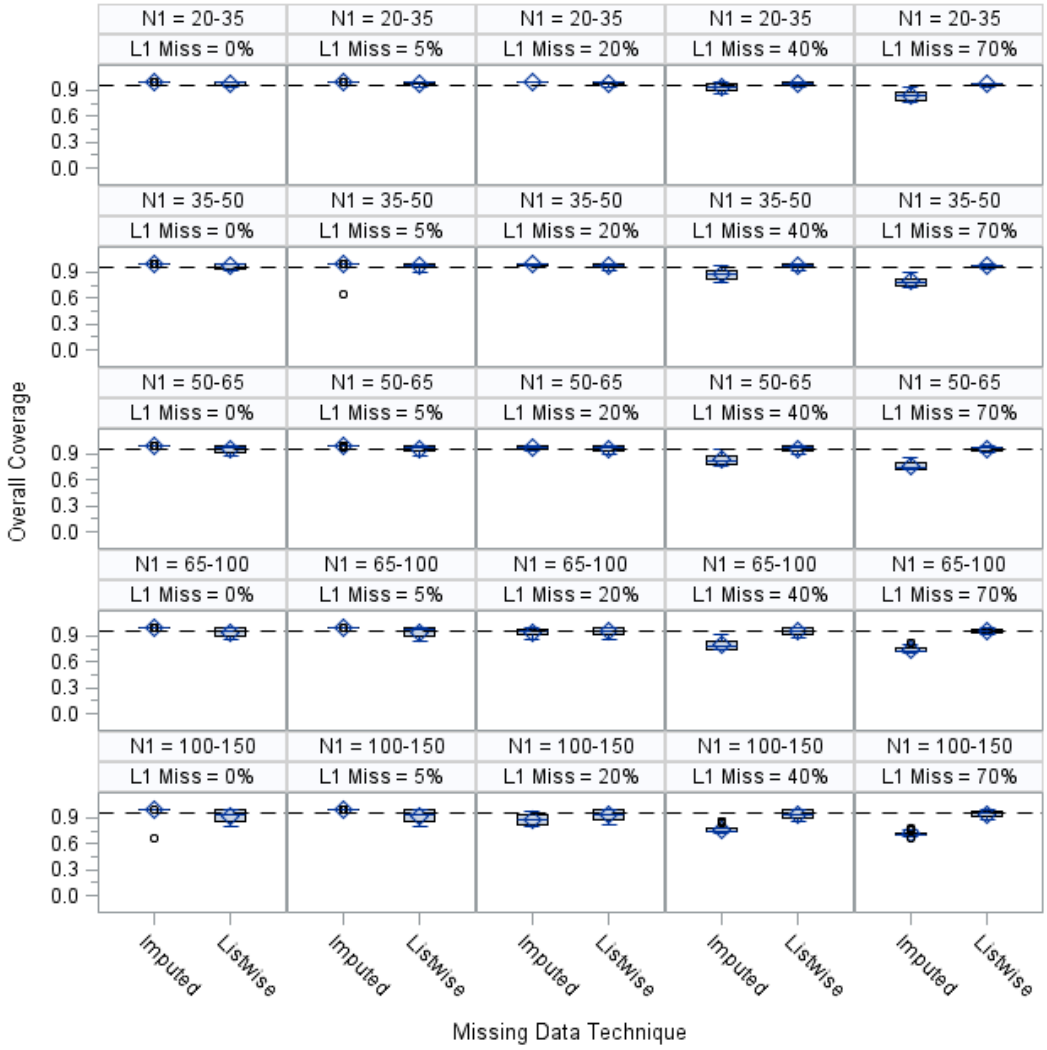

Figure 4.41 The Distribution of Confidence Interval Coverage by Missing Data Technique, Level-1 Sample Size, and Level-1 Missingness.

Figure 4.42 shows that listwise deletion stayed stable over level-2 sample size with a C.I. coverage value of .958, .962, .960, .961, and .960 when level-2 sample size was 20, 35, 50, 65, and 80, respectively. However, MLMI produced C.I. coverage values that decreased as level-2 sample size increased with values of .955, .930, .909, .897, and .885 when level-2 sample size was 20, 35, 50, 65, and 80, respectively. Thus, listwise deletion had acceptable levels across all levels of level-2 sample size whereas MLMI produced lower levels when level-2 sample size was 35 or larger.

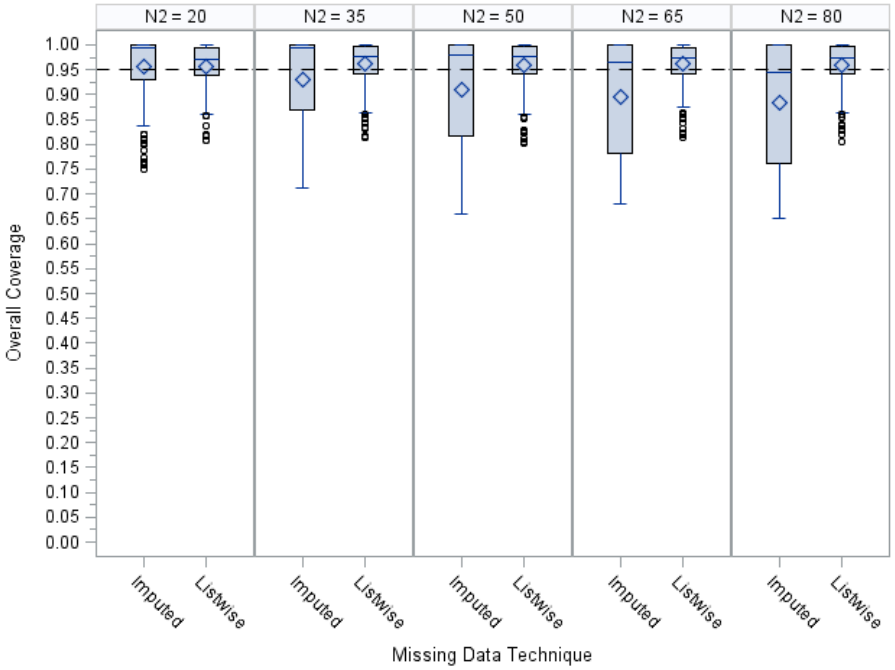

Figure 4.42 The Distribution of Confidence Interval Coverage by Missing Data Technique and Level-2 Sample Size.

Level-1 Confidence Interval Coverage. The mean C.I. coverage for level-1 was .807 (min = .000, max = 1.000). Notable amounts of level-1 C.I. coverage variance was explained by the interaction of MDT and level-1 missingness (η^2 = .310), level-1 missingness (η^2 = .270), MDT (η^2 = .136), level-1 sample size (η^2 = .055), the three-way interaction between MDT, level-1 sample size, and level-1 missingness (η^2 = .030), the interaction between MDT and level-2 sample size (η^2 = .026), level-2 sample size (η^2 = .024), and the three-way interaction between MDT, level-2 sample size, and level-1 missingness (η^2 = .022). Because the two three-way interactions encompassed all of the main effects and second order interactions, they will be summarized in this section.

Figure 4.43 depicts the three-way interaction between MDT, level-1 sample size, and level-1 missingness. This figure shows that when level-1 sample size and level-1 missingness were small (i.e., level-1 sample size was less than or equal to 35-50 and level-1 missingness was less than or equal to 20%), there was virtually no difference in

the C.I. coverage between the two MDTs. For listwise deletion, C.I. coverage stayed above 95% for all levels of missingness until level-1sample size was 65-100 or higher. Even when sample size was 65-100 or larger and missingness was high, C.I. coverage remained moderate, with the lowest value for listwise deletion being .847.

In contrast, MLMI produced estimates with less C.I. coverage, especially as the percent of missing data at level-1 increased. When level-1 sample size was 20 to 35, C.I. coverage fell to .767 and .443 when level-1 missingness was 40% and 70%, respectively. When level-1 sample size was 35 to 50, coverage stayed above 95% until missingness increased to 40% and 70% where coverage fell to .543 and .262, respectively. When level-1 sample size was 50- 65, coverage stayed above 95% until 20% missingness was reached and then fell to .906, .383, .173 coverage for 20%, 40% and 70% missingness, respectively. When level-1 sample size exceeded 60-65, coverage was above 95% until missingness was 20% and then sharply decreased to as much as .057 (in the case of level-1 sample size being 100-150 and level-1 missingness at 70%).

Listwise deletion produced estimates with higher coverage than MLMI when the percent of missing data was high at all levels of level-1 sample size. Coverage stayed above 95% at all levels of level-1 missingness when level-1 sample size was 20 to 35 and 35 to 50. When level-1 sample size was larger than 35 to 50, some decrease was seen in coverage as missingness increased, however the lowest coverage value was .845, and was much larger than MLMI coverage at the same levels. Thus, while coverage can be low, listwise deletion always resulted in similar or higher level-1 C.I. coverage than MLMI.

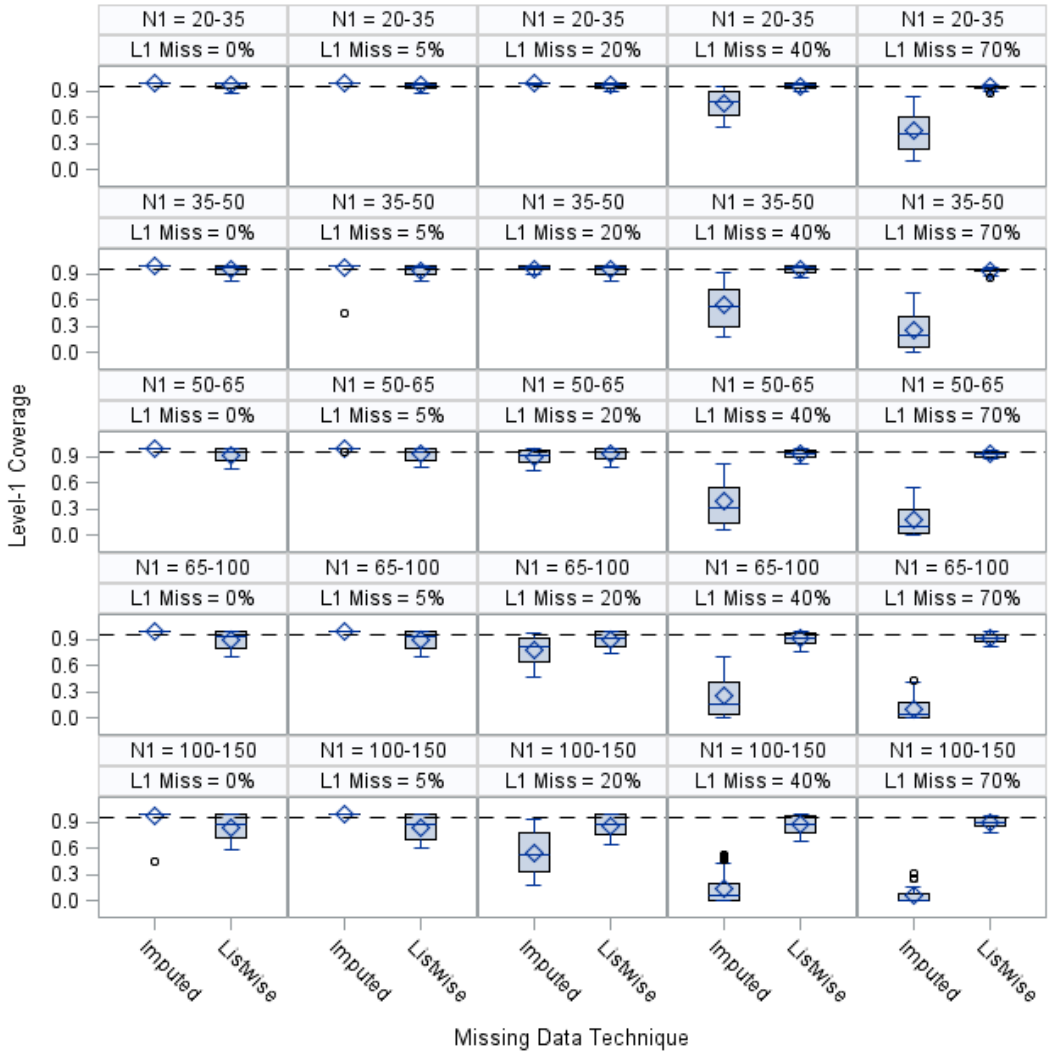

Figure 4.43 The Distribution of Level-1 Confidence Interval Coverage by Missing Data Technique, Level-1 Sample Size, and Level-1 Missingness.

Figure 4.44 shows that C.I. coverage for listwise deletion was slightly below .95 at every combination of MDT, level-2 sample size and level-1 missingness. However, C.I. coverage remained relatively high at every combination with a minimum C.I. coverage value of .912 and a maximum of .930. Although not as high as desirable, level-1 C.I. coverage did not seem to be impacted by level-1 missingness and level-2 sample size when listwise deletion was used. C.I. coverage was impacted by level-2 sample size and level-1 missingness when MLMI was used, where the C.I. coverage decreased as

level-1 missingness increased and the severity of the decrease depended on level-2 sample size.

Specifically, when level-2 sample size was 20, MLMI C.I. coverage was above .95 when level-1 missingness was 20% or below. C.I. coverage values decreased to .760 and .498 when level-1 missingness was 40% and 70%, respectively. When level-2 sample size was 35, C.I. coverage remained above .95 when missingness was 5% below, but decreased to .931, .532, and .263 when level-1 missingness was 20%, 40% and 70%, respectively. When level-2 sample size was 50, C.I. coverage was above .95 when missingness was 5% or below, and decreased to .851, .370, and .147 with 20%, 40%, and 70% missingness, respectively. When level-2 sample size was 65, again, C.I. coverage was above .95 when missingness was 5% or less, and decreased to .763, .252, and .083 when level-1 missingness was at 20%, 40%, and 70% respectively. Lastly, when level-2 sample size was 80, C.I. coverage remained about .950 when missingness was 5% or less. At 20%, 40%, and 70% level-1 missingness, values decreased to .676, .181, and .046, respectively.

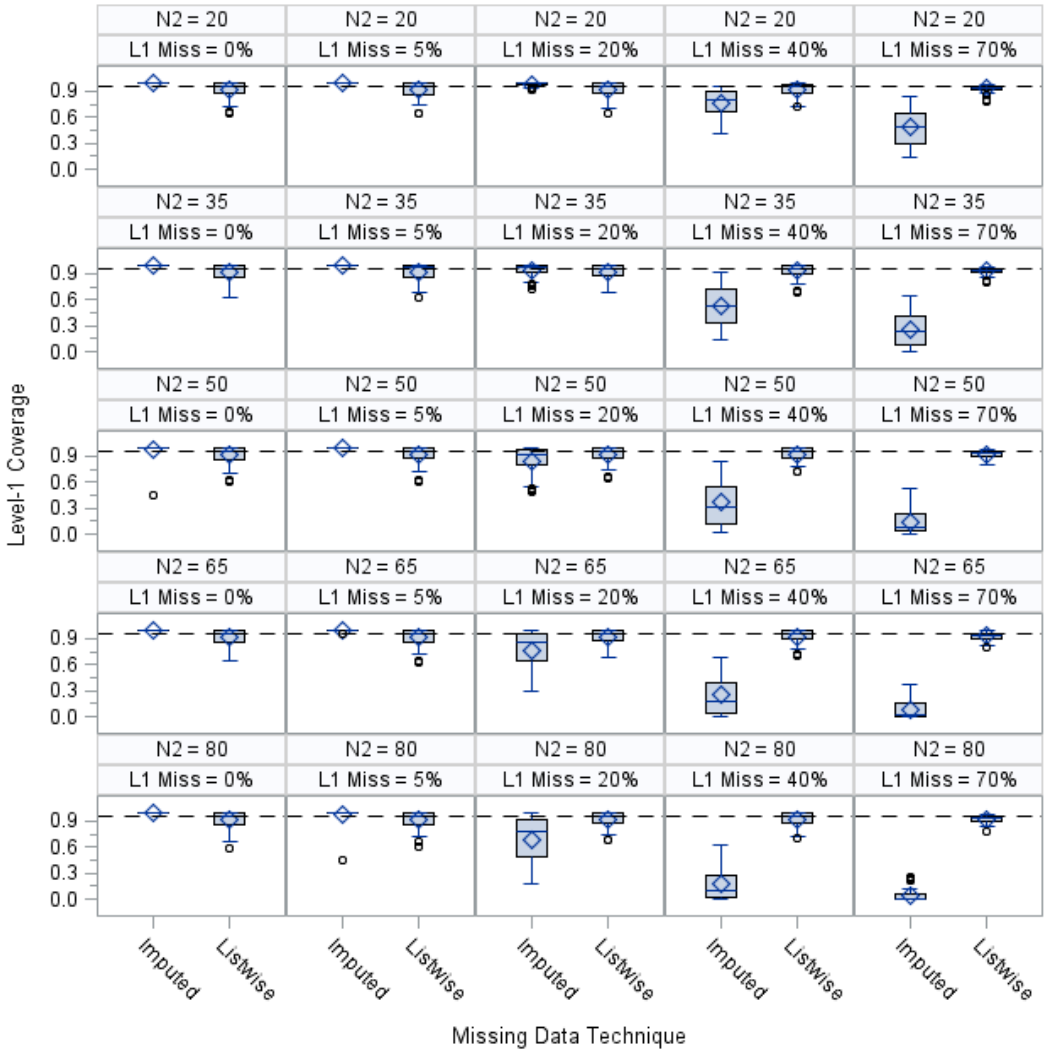

Figure 4.44 Distribution of Level-1 Confidence Interval Coverage by Missing Data Technique, Level-2 Sample Size and Level-1 Missingness.

Level-2 Confidence Interval Coverage. The mean confidence C.I. coverage for level-2 was .998. Notable amounts of variance occurred for C.I. coverage at level-2 for level-2 missingness (η^2 = .201), the interaction between level-2 sample size and level-2 missingness (η^2 = .028), the three-way interaction between level-1 sample size, level-1 missingness, and level-2 missingness (η^2 = .023), level-2 sample size (η^2 = .023), and the interaction between MDT and level-2 sample size (η^2 = .023). However, the overall R-squared for this outcome was only .493, which is low for a simulation study. To be

thorough, however, the interactions will be explored due to substantive interest in the design factors.

Figure 4.45 depicts level-2 C.I. coverage by level-2 missingness and level-2 sample size. When sample size was small, C.I. coverage slightly decreased when level-2 missingness was high, however, all values remained above the .95 threshold. This information combined with the overall R-square of the outcome, suggests that there was very little variability in level-2 C.I. coverage.

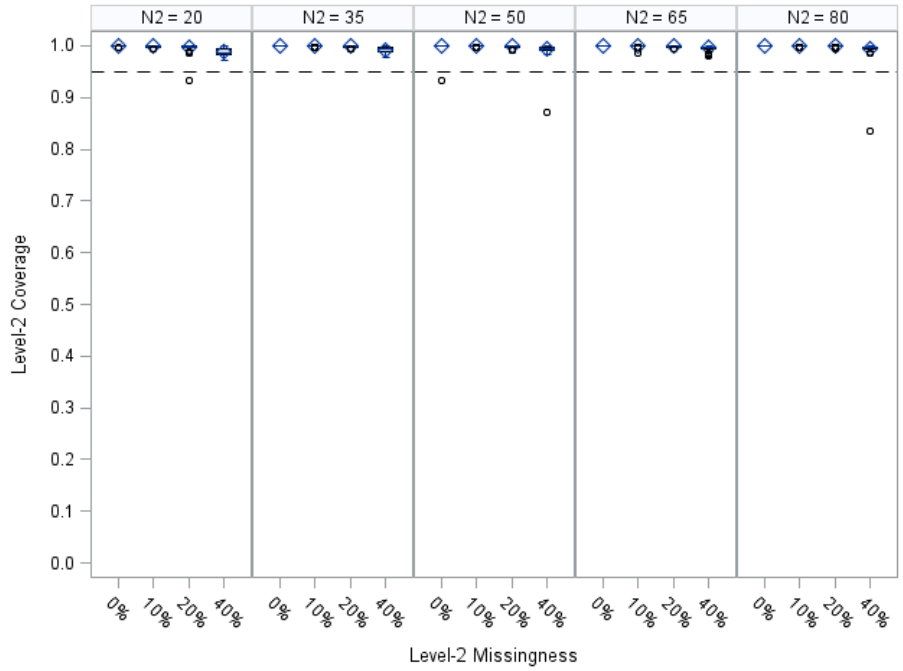

Figure 4.45 Distribution of Confidence Interval Coverage by Level-2 Sample Size and Level-2 Missingness.

Figure 4.46 depicts the three-way interaction between level-1 sample size, level-1 missingness and level-2 missingness. Overall, there was a slight decrease in level-2 C.I. coverage as missingness increased at each sample size level, however all values remained above the .95 threshold. This graph also suggests that there was very little variability in C.I. coverage values at level-2.

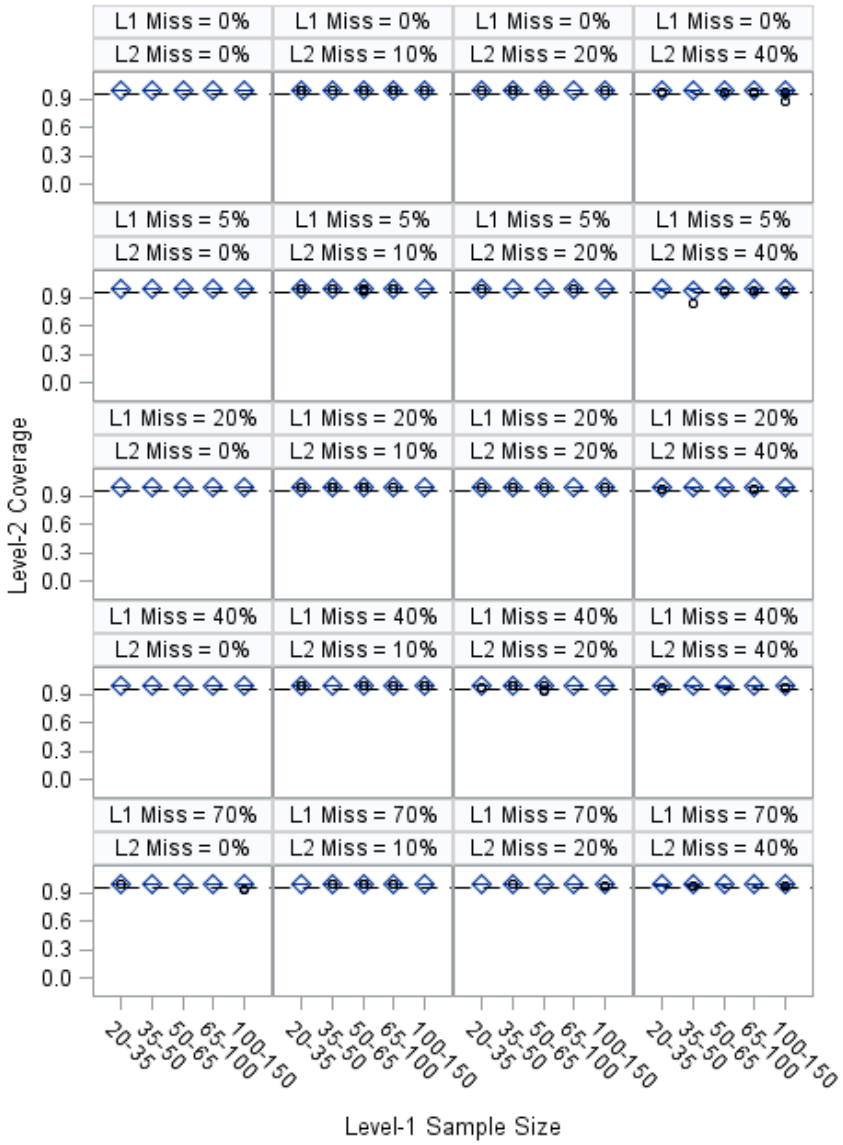

Figure 4.46 The Distribution of Level-2 Confidence Interval Coverage by Level-1 Sample Size, Level-1 Missingness, and Level-2 Missingness.

Figure 4.47 depicts the interaction between MDT and level-2 missingness. Overall there was a slight decrease in level-2 C.I. coverage values as level-2 missingness increased. Again, all C.I. coverage values remained above the .95 threshold across all combinations of design factor levels. This further shows that there was little variability to be explained in level-2 C.I. coverage values.

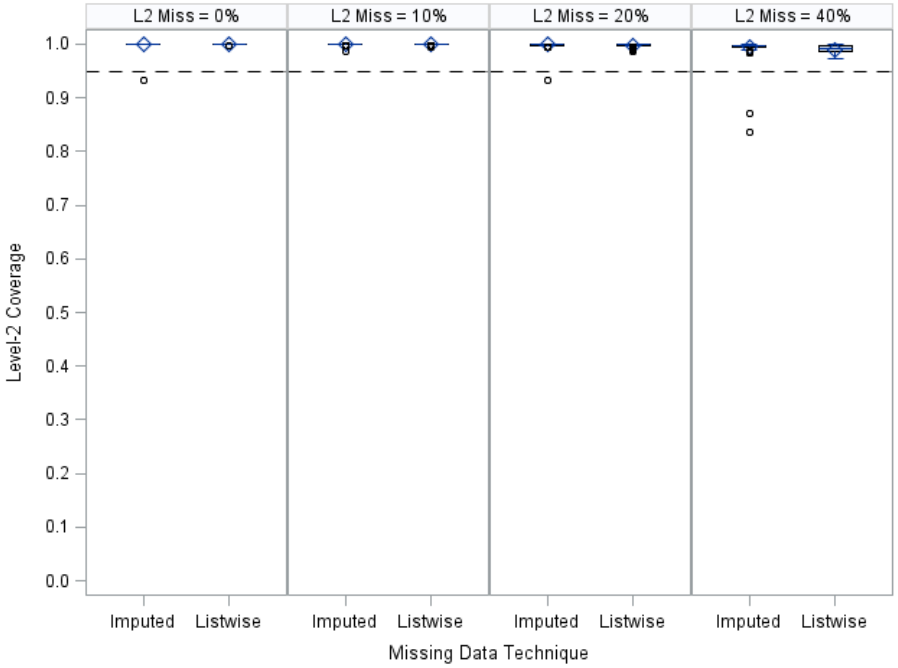

Figure 4.47 The Distribution of Level-2 Confidence Interval Coverage by Missing Data Technique and Level-2 Missingness.

Summary

The present study was intended to compare the performance of multilevel multiple imputation (MLMI) and listwise deletion in the context of linear two-level organizational models with continuous predictors. In this Chapter, I presented graphical results accompanied by text for practical and substantive design factors on bias, Type I error rate, power and C. I. coverage. Table 4.1 provides the eta-squared values for each of the outcomes by level and Table 4.2 provides a summary of the performance of listwise deletion and MLMI by each of these outcomes. In the next chapter, I will discuss how these results compare to previous literature and provide guidelines and recommendations for applied researchers who have missingness with clustered data.

Table 4.1. *Summary of Factors and Associated η^2 Values Covered in Chapter IV.*

Outcome	Level	η^2-1	η^2-2	η^2-3	η^2-4	η^2-5
Bias	Overall	MDT*L1 Miss = .280	MDT*L1 Miss*Mechanism = < .001	MDT*L2 Miss*Mechanism = < .001		
	L1	MDT*L1 Miss = .307				
	L2	MDT*L1 Miss = .239	MDT*L2 Miss = .008			
Type I Error	Overall	L1 Sample Size = .651	MDT*L1 Miss = .014			
	L1	L1 Sample Size = .661	MDT*L1 Miss = .014			
	L2	L1 Sample Size*L2 Sample Size*L2 Miss = .061	L1 Sample Size*L1 Miss*L2 Miss = .054	L2 Sample Size*L1 Miss*L2 Miss = .050	MDT = .028	
Difference in Type I Error	Overall	MDT*L1 Miss = .043				
	L1	MDT and L1 Miss = .043				
	L2	L1 Sample Size* L2 Sample Size* L1 Miss = .061	L2 Sample Size* L1 Miss*L2 Miss = .057	L1 Sample Size*L1 Miss*L2 Miss = .051	MDT = .034	
Power	Overall	MDT*L2 Miss = .014	L2 Sample Size*L2 Miss = .013	MDT*L2 Sample Size = .004		
	L1	L1 Sample Size*L2 Sample Size*L1 Miss = .112	MDT* L2 Sample Size* L1 Miss = .066	L2 Sample Size*L1 Miss*L2 Miss = .036	MDT*L2 Sample Size*L2 Miss = .029	MDT* L1 Sample Size*L1 Miss = 026
	L2	MDT*L2 Sample Size*L2 Miss = .001				
Difference in Power	Overall	MDT*L2 Miss = .098	L2 Sample Size*L2 Miss = .092	MDT and L2 Sample Size = .028		
	L1	L1 Sample Size*L2 Sample Size*L1 Miss = .112	MDT*L2 Sample Size* L1 Miss = .066	L2 Sample Size*L1 Miss*L2 Miss = .036	MDT*L2 Sample Size*L2 Miss = .029	MDT*L1 Sample Size*L1 Miss = .023
	L2	MDT* L2 Sample Size*L2 Miss = .010				
C. I. Coverage	Overall	MDT*L2Miss = .045	MDT*L1 Sample Size* L1 Miss = .032	MDT*L2 Sample Size = .023		
	L1	MDT*L1 Sample Size* L1 Miss = .030	MDT*L2 Sample Size* L1 Miss = .022			
	L2	L2 Sample Size*L2 Miss = .028	L1 Sample Size*L1 Miss* L2 Miss = .023	MDT*L2 Miss = .023		

Table 4.2. *Summary of Missing Data Treatment Results.*

Outcome	MDT	Performance
Bias	Listwise Deletion	Performed well at all levels
	MLMI	Yielded negative bias at level-1 and positive bias at level-2
Type I Error Rate	Listwise Deletion	Yielded conservative Type I error rates as missingness increased at level-1; Performed well at level-2
	MLMI	Yielded conservative Type I error rates at level-1 as missingness increases, but less than listwise deletion; Performed well at level-2
Power	Listwise Deletion	Performed well at level-1; Performed well at level-2 but power decreased as missingness increased (especially with small sample sizes) but overall power was retained as long as level-2 sample size was above 35
	MLMI	Performed well at level-1; Performed well at level-2 but power decrease as missingness increased (but less than listwise deletion) but overall power was retained
C.I. Coverage	Listwise Deletion	Performed well at all levels
	MLMI	YIelded extremely low C.I. coverage as level-1 sample size and level-1 missingness increased; Performed well at level-2

CHAPTER V

Discussion

The purpose of this study was to understand the influence of missing data technique, missing data mechanism, sample size, and amount of missingness on estimates of bias, power, Type I error rates, and confidence interval coverage in the context of linear two-level organizational models with continuous predictors. Additionally, one of the goals of this research was to be able to provide recommendations for applied researchers as to when multilevel multiple imputations (MLMI) and listwise deletion can be utilized.

This chapter briefly summarizes the results of the previous chapter and relates these findings back to previous literature. For findings that contradict previous research, I provide hypotheses as to why these differences could have occurred. The first and second section of this chapter discusses previous research on bias and power, respectively, as well as hypothesized reasons for the results. The third and fourth section summarize the findings on Type I error rates and confidence interval coverage along with hypotheses for findings. The fifth section uses the results from this study to provide recommendations for applied researchers. The sixth and final section addresses study limitations and suggestions for future studies.

Bias

Bias in the context of treating missing data has been well researched in non-hierarchical data. According to research in this area, bias can result when the wrong MDT is used, and the magnitude of bias is influenced by which technique is used, sample size, and amount of missingness, and missing data mechanism. Specifically, previous research has found that listwise deletion is effective if sample size is large, missingness is small, and data are MCAR, whereas imputation procedures are generally effective regardless of missing data mechanism, sample size, or missingness (Becker & Powers, 2001; Becker & Walstaf, 1990; Rubin, 1987).

With hierarchical data, however, previous research has found that bias is not an issue when using listwise deletion. In the context of MCAR, listwise deletion was found to outperform mean substitution and multiple imputation in the estimation of fixed and random effects (Gibson & Olejnik, 2003). Additionally, Kwon (2011) showed listwise deletion generally outperformed the other MDTs when data were MAR. He found that these methods produced "practically acceptable" bias in most fixed effects that were highly related to missingness. Cai (2008) also found that listwise was acceptable under MAR, especially when level-2 sample size was small. This study showed similar findings given the context of the study.

Results of this study showed that the interaction of MDT and level-1 missingness was the most important factor when it came to examining overall bias, bias at level-1 and bias at level-2. Bias did not seem to be an issue with listwise deletion given the context of this study, regardless of mechanism and missingness. However, bias can be a problem

with MLMI. MLMI produced negative overall bias, albeit very small, starting when missingness was 20% (-0.003), 40% (-0.006), and 70% (-0.010). Additionally, a moderate amount of bias occurred at level-1 when missingness was 20% (-0.0129), 40% (-0.0248), and 70% (-0.0399) and a small amount of positive bias was present at level-2 when missingness was 20% (0.0073), 40% (0.0128), and 70% (0.0203).

Before researchers can deem listwise deletion as effective at reducing bias, other design factors need to be manipulated that could impact bias. First, the gamma value, while based on previous literature, did represent a large effect size. Because of this, for the non-null missing variables, the effect could have been easier to detect and reproduce even when missingness was present. Thus, data simulated to have smaller gamma values could result in poor performance from listwise deletion. Second, the ICC could have had an impact on bias such that data with smaller ICCs may produce more bias as missingness increased (this could be the case with both MLMI and listwise deletion), specifically because the smaller the ICC the closer the data gets to having a non-heirarchical data structure, and thus, should be more similar to results from OLS. Third, the lack of bias with listwise deletion could be attributed to having positive bias on one of the vaeriables with missing data and negative bias on the second variable. In order to get the effect at each level, the two variables containing missingness at that level were averaged. Thus, using level-1 as an example, if X1 contained an average bias of -0.05 and X2 had an average bias of 0.05, these two effects would cancel each other out and it would seem that there was no bias present at the overall level.

In addition there could be some reasons why MLMI seemed to produce bias (i.e., little to no bias in some conditions and a moderate amount in other conditions). Most

likely, this is due to model size and the correlations among variables in the model. Previous research in OLS regression has determined that as model size increases, the more useful imputation procedures become (Enders, 2010; Gibson & Olejnik, 2003), as there is more information to explain misingness. Although this model was based on the typical number of predictors found in previous research, it is still a small model. Perhaps MLMI would have no bias or have less bias as model size becomes larger. Thus, MLMI needs to be further studied with various model sizes.

Lastly, the correlation among variables is crucial with imputation procedures, such that the more highly correlated variables are, the better the imputation procedure (Enders, 2010). This study used a moderate correlation (r = .25), and results could be different if the variables were more or less highly correlated.

Given the amount of emphasis in the missing data literature on missing data mechanisms, it was surprising to find that mechanism had no influence on overall bias, bias at level-1 or bias at level-2. Specifically, previous research showed that listwise deletion should perform well only when data are MCAR. The results of this study showed that listwise deletion performed well despite the underlying mechanism of the missingness. This could have happened for a couple of reasons. One reason could be that perhaps mechanism is not as important with hierarchical data. In addition to this study, previous studies have also demonstrated that listwise deletion has been shown to be effective with both MAR (Gibson & Olejnik, 2003) and MCAR missingness (Cai, 2008; Kwon, 2011). This could be attributed to the hierarchical nesting of the data. For example bias at level-1 may not be impacted by mechanism because these level-1 units are nested within level-2, thus we have more information to estimate these parameters than we do

with OLS regression. Second, it could be that mechanism does impact MLMs, but the impact was not detected in this study. Again, this could be because of the large effect size of the gamma used during simulation, or the ICC value. Using these as design factors in future studies can uncover more information about how listwise deletion impacts bias.

Power

Power has also received a lot of attention in previous literature on non-hierarchical data. Previous research concluded that listwise deletion can result in decreased power, while multiple imputation procedures retain power levels. Specifically, listwise deletion seems to result in decreased power when missingness is high and sample size is small, regardless of missing data mechanism. However, listwise deletion has been shown to retain satisfactory power when sample size is adequate and there is a small amount of missingness (Enders, 2010; Kromery & Hines, 1994). Because of this, researchers have generally recommended multiple imputation over listwise deletion with one of the reasons being that power is not impacted by the missingness when this procedure is used.

Research on hierarchical data has also been concerned with power. Similar to the non-hierarchical case, power has been shown to be an issue as the amount of missingness increases (Cai, 2008; Kwon, 2011). Using a deletion method in MLM can result in an even more dramatic reduction of the sample size, especially if data are missing at level-2 as all the corresponding level-1 units will subsequently be deleted. Because of this, previous literature has been the most concerned with level-2 missingness and disapproving of listwise deletion in MLMs.

This study showed that power decreased for both listwise deletion and MLMI, however the decrease for listwise deletion was more pronounced. However, power overall was retained above .80 for both MLMI and listwise deletion across most levels of missingness, especially when sample size at level-2 was large. While power did decrease overall, it was surprising that power for listwise deletion was, for majority of the conditions, retained. Overall, data characteristics that have been shown to impact power in MLMs are sample sizes, missingness, effect sizes, and ICC values. Notice that only two of these characteristics were manipulated in this study. Thus, manipulating the other two could produce completely different results. Namely, the gamma value used to simulate data, although based on previous literature, did represent a large effect size. Due to this, it was not too surprising that power remained high even with listwise deletion. Additionally, the ICC in this study was moderate. Data with smaller ICCs could be impacted completely different, such that a smaller ICC value could result in listwise deletion being less powerful and perhaps falling below the .80 threshold. These variables need to be manipulated in future studies in order to further understand the impact of listwise deletion on power.

Type I Error

Although less researched in the context of missing data, Type I error rate was examined in this study in order to determine if missing data impacted Type I error rates. This study showed that level-2 Type I error always remained adequate regardless of MDT, while Type I error rate at level-1 become overly conservative as missingness increased for both MDTs, but was more conservative for listwise deletion. This finding was very unexpected and has not been observed with previous research. I have only one

hypothesis about why this could have happened. Since level-1 rates were impacted but level-2 rates were not, and the impact was true for both MLMI and listwise deletion, perhaps the nested structure of the data caused Type I error rates to become conservative when missingness increased. That is, if a level-2 unit is deleted or imputed, all the corresponding level-1 units are also deleted or imputed with the same value, and thus level-1 can be more impacted by missingness than level-2.

More information needs to be collected about the impact of missingness on level-1 Type I error rate. While there was no evidence that Type I error was inflated, researchers should be mindful that Type I error rates could become overly conservative. This is important because the Type I error rate can impact power (among other things such as ICC and effect size). Specifically, an overly conservative Type I error rate can result in a decrease in power.

Confidence Interval Coverage

Confidence interval coverage (C. I. coverage) has also received very little attention in missing data research. This study found that listwise deletion clearly outperformed MLMI in level-1 C. I. coverage, but was adequate for both MDTs at level-2. Specifically, C. I. coverage produced from listwise deletion was adequate at both levels, yet seemed to decrease as missingness increased, which is not surprising. What was surprising, however, is the decrease in level-1 C. I. coverage when MLMI was used, especially how the coverage plummeted when sample size or missingness was large.

The findings observed here could have happened for a couple of reasons. First, the introduction of level-1 bias with MLMI could impact level-1 C. I. coverage. For

example, bias seemed to be introduced for MLMI (in some cases very little bias and in some cases a moderate amount of bias), which could have impacted level-1 C. I. coverage. For example, if MLMI produced bias to any extent in the estimates, the corresponding confidence interval around that estimate is less likely to contain the complete estimate value. For bias, the conditions that caused the most bias were when level-1 missingness was high. With C. I. coverage, this was also the case, which could provide initial evidence to support this hypothesis (although much more information is needed).

Further, MLMI could be producing low C. I. coverage because the model used did not have sufficient information to impute, especially with high missingness. For example, take the conditions where we have 70% of the data being imputed. There would need to be a lot of information explaining the missingness in the model in order to adequately impute the remaining 30% of the data. As stated previously, the model for this study was based upon literature, but a small model nonetheless. Previous research has found that the more information we have related to missingness, the better imputation methods work. Thus, perhaps a larger model with more information would result in adequate C. I. coverage, especially in conditions with high amounts of missingness.

Recommendations & Conclusions

In 1999 The American Psychological Association Task Force on Statistical Inference explicitly warned against the use of traditional MDTs, such as listwise deletion as several research studies had shown that the use of traditional MDTs could introduce bias into parameter estimates derived from a statistical model (Becker & Powers, 2001;

Becker & Wathlstaf, 1990; Rubin, 1987) and could result in a loss of information and statistical power (Anderson, Basilevsky, & Hum, 1983; Kim & Curry, 1977) when strict assumptions were not met.

While the statements from The American Psychological Association Task Force on Statistical Inference that explicitly warned against the use of traditional MDTs was based upon a plethora of literature suggesting that traditional techniques can cause issues with power and bias, it neglected to point researchers to literature that suggests that this is not true in all contexts. Previous research has shown that using a traditional method does not necessarily reduce statistical power or bias parameter estimates in every context. Research using traditional OLS regression techniques has shown that sample size and missingness are key characteristics in determining under what conditions each MDT should be used. For example, researchers conducting simulation studies using traditional OLS regression have come to the general consensus that pairwise and listwise deletion methods work well when sample sizes are large, and the missingness is small (Basilevsky et al. 1985; Kim & Curry, 1977; Roth & Swizer, 1995; Witta, 1992). Thus, in certain situations traditional methods can be appropriately applied despite the underlying missing data mechanism.

All of these studies, however, were completed on non-hierarchical data. Since multilevel models (MLMs) are an extension of single-level regression, it is plausible that results from single-level data extend to hierarchical data. However, these results suggest that listwise deletion and MLMI do not perform the same as they do with non-nested data. In fact, previous multilevel modeling simulation studies, as well as this study, show that the benefits of listwise appear to outweigh the disadvantages. First, based on the

results from this study, listwise deletion in MLMs did not introduce bias or inflated Type I error rates, and provided adequate C. I. coverage. Second, the procedure is very easy to implement and no subjective decisions need to be made prior to implementing the procedure. Third, this procedure is widely available in common statistical packages. One minor disadvantage of listwise deletion is that power does decrease at level-2. This disadvantage is deemed minor, however, because extreme missing data has to be present at level-2 in order for power to decrease below the nominal level of .80 based on a large effect size and moderate ICC value.

The disadvantages of MLMI, at least in the context of this study, appear to outweigh the advantages. The first disadvantage of MLMI is that is complicated to use. As stated in Chapter 2, several decisions need to be made such as specification of the imputation model, deciding how many imputation iterations need to be calculated, which algorithm to use, etc. Thus, two researchers could impute the same data set and come up with different estimates if they make different decisions during the imputation phase. Second, MLMI is not commonly available in most software packages. In addition, there are a lot of qualms with researchers about the different packages, what algorithms they use, and which levels can be imputed. Thus, software choice can lead two researchers to arrive at different conclusions.

Third, in the context of this study MLMI coverage of the estimated gamma from the complete case is equivalent to or worse than listwise deletion. As mentioned above, however, there could be some characteristics of this simulation study that impacted this that need to be further explored. Fourth, the imputation procedure is only as good as the imputation model specified. In this context, I had a correctly specified model due to

simulation. However, in real applied research, this is very unlikely. Having large imputation models is one way to offset this, however. Lastly, there could be an issue with bias in MLMI. Although more simulation studies are needed to better understand the potential impact that different ICC values, model size, and gamma values might have on bias, in the context of this simulation study, MLMI introduced bias in parameter estimates at both level-1 and level-2. While the bias was not consistent across design factors (e.g., for some conditions it was minimal and for others it was moderate), why would an applied researcher use a more complicated, potentially biased method over a simple non-biased method?

Overall, these results show that the blanket statement made by the American Psychological Association Task Force on Statistical Inference in 1999 is, perhaps, too general. In order to determine what MDT to use, researchers should take a more thorough look at missing data literature and simulation studies relevant to the design of their analyses instead of simply abandoning traditional techniques altogether. Specifically, in the context of linear two-level organizational multilevel models, with the characteristics of this study, listwise deletion seems to function rather well, and usually better than MLMI.

In addition to recommending listwise deletion over MLMI, these results show that when considering adequate power at level-2, researchers need to take into consideration both level-2 sample size as well as the level-2 missingness. Several sample size recommendation have been suggested in the multilevel modeling literature. The most common is 30 level-1 and 30 level-2 units (Maas & Hox, 2004; 2005; Pacagnella, 2011). However, Bell et al. (2014) have shown this recommendation to produce underpowered

analyses and suggests level-1 sample sizes greater than 20-40 and level-2 sample sizes greater than 30 to detect significant level-2 effects. This study showed that even with a level-2 sample size of 35 and 40% missingness, power fell below the nominal level of .80 for both MLMI and listwise deletion. When level-2 sample size was 50, 65, and 80, power was not an issue. Thus, with level-2 sample sizes in the 30-35 range, researchers need to be aware that having missing data could result in less than optimal statistical power regardless of MDT.

Limitations & Future Research

As with any research study, the generalizability of these results is limited to the design factors and facets manipulated in the study. Applied researchers wishing to use the guidelines set forth by this study need to be cautious and mindful of how close their research scenarios mirror the design and data of this simulation study.

This study examined a random intercept model with no cross-level interactions, as it was determined through examination of previous applied research to be the most common type of model. Because of this, results may not generalize to more complex models such as the random intercept and slopes model or models with cross-level interactions, which could be a good avenue for future research. The ability to estimate both fixed and random effects is a major benefit for multilevel models, thus researchers need to examine how the rate of non-positive definite G matrices, variance component bias, variance component power, and variance component coverage are impacted by these design factors.

In addition to more complex models with variance components, this study should also be extended to multilevel models with dichotomous outcomes. Previous research has shown that sample sizes need to be larger in order to have adequate power in models with dichotomous outcomes (Schoeneberger, in press). Since listwise deletion does cause a decrease in power, although slight in the case of models with continuous outcomes, the impact could be greater with multilevel logistic models.

The examination of model size could also be an avenue of research. Previous research has determined that as model size becomes larger, the more useful imputation procedures are (Enders, 2010; Gibson & Olejnik, 2003), as there is more information to include in the imputation model. In the original design of this study, model size was going to be manipulated in order to generally understand the utility of each of the MDT as model size increased. However, in order to keep this study informed by the literature, as was done with all of the study design factors, this factor was not included in the study given that the 39 articles examined when designing the study suggested that on average applied researchers typically use smaller models. However, this is still a worthy area for future research to examine.

This study used a specific ICC and gamma value that was identified to be representative of applied research. With that said, however, the gamma was quite large and the ICC value was moderately large. Little is known about how varying ICC values, effect size values, and missingness percentages can impact bias, Type I error, power, and C. I. coverage in MLMs. Thus, future studies should incorporate these as design factors by manipulating these values and examining if they have an impact on the outcomes used in this study.

Lastly, in order to examine a large number of design factors, results were manipulated such that more than on combination of design factors were aggregated and examined by outcome. Thus, within each of the plots in Chapter 4, there are various design factor combinations, and the mean across all of these conditions was used to evaluate performance. For example, a plot with a main effect of MDT would have a box plot for every condition where MLMI was used and every condition where listwise deletion was used. Note that within a boxplot, the only necessary common characteristic is that the condition used the same MDT. Some of the conditions within this boxplot had a level-2 sample size of 20, a level-1 sample size of 20-35, 0% missingness at level-1 and level-2 and data missing that was MAR while another had level-2 sample size of 80, level-1 sample size of 100-150, 70% missingness at level-1, 40% missingness at level-2, and data missing that was MCAR, or any other combination of level-1 and level-2 sample sizes, level-1 and level-2 missingness, and mechanism. Thus, the mean represents the average value of the outcome across many different related conditions. Note that in addition to examining the means, it could also be useful in future research to look at the conditions *within* the boxplot to help get a better sense of how these design factors impact each of the outcomes. Additionally, while the method used in this study was chosen such that a large amount of design factors could be manipulated in the same study, in order to truly assess the performance of each of the design factors, a study where each design factor is manipulated one at a time across multiple replications would be needed.

REFERENCES

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2002). Random-effects modeling of categorical response data. *Sociological Methodology, 30,* 27-80.

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications.

Anderson, A., Basilevsky, A., & Hum, D. (1983). *Missing data: A review of the literature*. In P. Rossi, J. Wright & A. Anderson (Eds.), Handbook of survey research (pp.415-494). New York: Academic Press.

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

Austin, P. C. (2005). Bias in penalizezd quasi-likelihood estimation in random effects logistic regression models when the random effects are not normally distributed. *Communications in Statistics- Simulation and Computation, 34,* 549-565.

Austin, P. C. (2007). A comparison of the statistical power of different methods for the analysis of cluster randomization with binary outcomes. *Statistics in Medicine, 26,* 3550-3565.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of cluster is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics, 6*(1), 1-30.

Basilevsky, A., Sabourin, D., Hum, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulation data as an experimental design. *Communications in Statistics, 14, 371-394.*

Becker, W. E., & Powers, J. R. (2001). Student performance, attrition, and class size given missing student data. *Economics of Education Review,* 20(4), 377-388.

Becker, W., & Walstad, W. (1990). Data loss from pretest to posttest as a sample selection problem. *Review of Economics and Statistics*, 72(1), 184–188

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimated in two-level linear models. *Methodology*.

Bell, B. A., Schoeneberger, J. A., Morgan, G. B., Kromrey, J. D., & Ferron, J. M. (2010). N <= 30: Impact of small level-1 and level-2 sample sizes on estimates in two-level multilevel models. Presentation at the American Educational Research Association Conference. Denver, CO.

Bell, B. A., Schoeneberger, J. A., Mogran, G. B., Zhu, M., Ferron, J. M., & Kromrey, J. D. (2011, May). Heriarchical vs. contextual models: Sample size, model complexity, and the 30/30 rule. Presentation at the Annual Modern Modeling Methods (M3) Conference. Storrs, CT.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches.* New York, NY: Russell Sage Foundation.

Bosker, R.J., Kremers, E. & Lugthart, E. (1990). School and instruction effects on mathematics achievement. *School Effectiveness and School Improvement*, 1 (4), 1-16.

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation odels with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal, 1*, 284-316.

Cai, X. (2008). *Missing data treatment of a level-2 variable in a 3-level hierarchical linear model.* (Order No. 3303463, Western Michigan University). ProQuest Dissertations and Theses, 129. Retrieved from http://search.proquest.com/docview/304446541?accountid=13965. (304446541).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Psychology Press.

Collins, C. M., Schafer, J. L. & Kam, C. M. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing data procedures. *Psychological Methods, 6,* 220-351.

Correti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal*, 44(2), 298-338.

Desimore, L. M., Smith, T., Baker, D., & Ueno, K. (2005). Assessing barriers to the reform of U.S. mathematics instruction from an international perspective. *American Educational Research Journal,* 42, 501-535.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Edward Arnold.

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

Enders, C. K. & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 430-457.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zharias, J. (2001). The enduring effects of small classes. *Teachers College Record,* 103, 145-183.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF deterction. *Journal of Educational Measurement, 47*(3), 299-317.

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

Glasser, M. (1964). Linear regression with missing observations among the independent variables. *Journal of the American Statistical Association, 59,* 834-844.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40,* 229-252.

Goddard, Y., Goddard, R., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record,* 109, 877-896.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* NY, NY: Oxford University Press

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed). London: Edward Arnold.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206-213.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal,* 42(2), 371-406.

Hill, P. W., & Rowe, K. J. (1996). Multilevel modeling in school effectiveness rsearch. *School Effectiveness and School Improvement, 7 (1),* 1-34.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Application.* New York, NY: Routledge.

Julian, M. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 325-352. doi: 10.1207/S15328007SEM0803_1

Kim, J. & Curry, J. (1977). The treatment of missing data in multivariate analyses. *Sociological Methods and Research, 6,* 215-240.

Kreft, I. G., & De Leeuw, J. (1998). *Introductin multilevel modeling.* Newbury Park, CA: Sage.

Kromery, J. D. & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Education and Psychological Measurement, 54,* 573-593.

Kwon, H. (2011). *A monte carlo study of missing data treatments for an incomplete level-2 variable in hierarchical linear models*.(Order No. 3476918, The Ohio State University). ProQuest Dissertations and Theses, 170. Retrieved from http://search.proquest.com/docview/898360287?accountid=13965. (898360287).

Kyriakides, L., Campbell, R. J., & Gatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers' comprehensive model of educational effectiveness. *School Effectiveness and School Improvememnt,* 11(4), 501-529.

Lamb, S., & Fullarton, S. (2002). Classroom and school factors affecting mathematics achievement: A comparative study of Australia and the United States using TIMSS. *Australia Journal of Education,* 46, 154-171.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87,* 1227-1237.

Longford, N. T. (1993). *Random coefficient models.* New York: Oxford university press, Inc.

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137.

Marsh, H.W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal, 5,* 22-36.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, 37, 153-184.

Mason, M. J. (1999). A review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement and Evaluation in Counseling and Development, 32,* 111-118.

Mistler, S. A. (2013a). *A SAS macro for applying multiple imputation to multilevel data.* Proceedings of the SAS Global Forum 2013, San Francisco, California: Contributed Paper (Statistics and Data Analysis) 438-2013.

Mistler, S. A. (2013b). A SAS macro computing pooled likelihood ratio tests with multiply inputed data. Proceedings of the SAS Global Forum 2013, San Francisco, California: Contributed Paper (Statistics and Data Analysis) 440-2013.

Moerbeck, M. (2004). The consequences of ignoring level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39,* 129-149. doi: 10.1207/s15327906mbr3901_5

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression. *BMC Medical Research Methodology, 7*(34), 1-10.

Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York, NY: Oxford University Press.

Nich, C. & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the alaysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology, 65,* 252-261. Doi: 10.1037//0022-006X.65.2.252

Opdenakker, M., & Van Damme, J. (2001). Relationship between school composistion and characteristics of school process and their effect on mathematics achievement *British Educational Research Journal,* 27, 407-432.

Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.). *Real data analysis* (pp. 31-78). Greenwich, CT: Information Age.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74(4),* 525-556.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2005). *HLM 6: Hierarchical linear and nonlinear modeling* [Computer Sotware]. Chicago, IL: Scientific Software International.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47,* 13-26.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 13-26.

Roth, P. L., & Swizer, F. S. (1995). A Monte Carolo analysis of missing data techniques in a HRM setting. *Journal of Management, 21,* 1003-1023.

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

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473-489.

SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall.

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

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MD: Houghton Mifflin.

Schoeneberger, J. (in press). The impact of sample size prevalence, estimation method, and other factors when estimating multilevel logistic models. *Journal of Experimental Education.*

Singer, J. (1998). Using SAS Proc Mixed to fit multilevel Models, Hierarchical Models, and Individual Growth Curves. *Journal of Educational and Behavioral Statistics, 24 (4),* 323-355.

Snijders, T. A. & Bosker, R. J. (1999) *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE publications.

Swoboda, C. C., & Kim, J.-S (2010, April). *Multiple imputation methods with multilevel data: A simulation study.* Paper presented at the Annual Meeting of American Educational Research Association. Denver, CO.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 528-540.

Theall, K. P., Scribner, R., Broyles, S., Yu, Q., Chotalia, J., Simonsen, N., Schonlau, M. & Carolin, B. P. (2011). Impact of small group size on neighborhood influence in multilevel models. *Journal of Epidemiology Community Health, 65,* 688-695.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18,* 681-694.

van Buuren, S., Groothuis-Oudshoorn, K. (2011a). Mice: Multivariate imputation by chained equations. R package version 2.9, URL http://CRAN.R-project.org/package=mice.

van Buuren, S., Groothuis-Oudshoorn, K. (2011b). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45* (3), 1-67.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Witta, E. L. (1992). Seven methods of handeling missing data using samples from a national database. *Dissertation Abstracts International, 53 (09),* (UMI No. 930567).

Wompold, B. E., & Serlin, R. C. (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425-433. doi: 10.1037//1082-989X.5.4.425

Wothke, W. (1993). Nonpositive definite matrices in structural equation modeling. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219-240). Mahwah, NJ: Erlbaum.

Xue, Y. (2002). The influence of early literatacy instruction on children's learning in kindergarten. *Dissertation Abstracts International,* 63 (07). (UMI No. 2058082).

Zhang, D. (2005). A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling. *Dissertation Abstracts International, 67 (08),* (UMI No. 3231609).

APPENDIX A – CODE FOR MCAR DELETION

```
/**************************MCAR missingness;*/
**Cutoffs for uniform distribution;
**Level-1;
*5% missingness;
%if &l1miss=5 %then %do; %let x1value = .975; %let x2value = .025;
%end;
*20% missingness;
%if &l1miss=20 %then %do; %let x1value = .90; %let x2value = .10; %end;
*40% missingness;
%if &l1miss=40 %then %do; %let x1value = .80; %let x2value = .20; %end;
*70% missingness;
%if &l1miss=70 %then %do; %let x1value = .65; %let x2value = .35; %end;

**Level-2;
*10% missingness;
%if &l2miss=10 %then %do; %let z1value = .95; %let z2value = .05; %end;
*20% missingness;
%if &l2miss=20 %then %do; %let z1value = .90; %let z2value = .10; %end;
*40% missingness;
%if &l2miss=40 %then %do; %let z1value = .80; %let z2value = .20; %end;

**Level-1;
%if &l1miss gt 0 %then %do;
proc sort data=diss15;
      by condition replicate;
data MCAR_&cond;
      set diss15;
      length miss_type $4;
      miss_type="MCAR";
      by condition replicate;
      call streaminit(0);
      Uniform_1=rand("Uniform");

      if Uniform_1 ge &x1value then x1delete = 1; *~2.5% out of X1;
      if Uniform_1 le &x2value then x2delete = 1; *~2.5% out of X2;
      if X1delete = 1 then x1_MCAR = .;
      if x1delete ne 1 then x1_MCAR = X1;
      if x2delete = 1 then x2_MCAR =.;
      if x2delete ne 1 then x2_MCAR = X2;
      run;
%end;
%else %if &l1miss = 0 %then %do;
data MCAR_&cond;
set diss15;
length miss_type $4;
      miss_type="MCAR";
      x1_MCAR = x1;
      x2_MCAR = x2;
```

```
      run;
      %end;

**Level-2;
%if &l2miss gt 0 %then %do;
proc sort data = MCAR_&cond;
by replicate l2_id;
run;
proc means data = MCAR_&cond noprint;
class replicate l2_id;
var z3 z1 z2; output out=MCAR_comp mean(z3)=z3 mean(z1)=z1 mean(z2)=z2;
run;

proc sort data = MCAR_comp; by replicate; run;
data MCAR_comp; set MCAR_comp;
if l2_id =. or replicate = . then delete;
by replicate;
call streaminit(0);
Uniform_2 = rand("Uniform");
if Uniform_2 ge &z1value then z1delete = 1;
if Uniform_2 le &z2value then z2delete = 1;
if Z1delete = 1 then z1_MCAR = .;
if Z1delete ne 1 then z1_MCAR = Z1;
if Z2delete = 1 then z2_MCAR = .;
if Z2delete ne 1 then z2_MCAR = Z2;
run;

data MCAR_&cond; merge MCAR_&cond MCAR_comp(keep = replicate l2_id
z1delete z2delete z1_MCAR z2_MCAR);
by replicate l2_id;
run;
%end;
%else %if &l2miss = 0 %then %do;
data MCAR_&cond; set MCAR_&cond;
z1_MCAR = z1;
z2_MCAR = z2;
run;
%end;
```

APPENDIX B – CODE FOR MAR DELETION

```
*****************************MAR missingness;
data diss15_MAR;
      set diss15;
      length miss_type $4;
      miss_type="MAR";
      run;

*****************MAR at level-1;
%if &l1miss gt 0 %then %do;
*restricting upper limit to 50%;
proc means data = diss15_mar noprint;
class replicate;
var x3; output out = lev1_comp p50=p_50z;
run;
data lev1_comp; set lev1_comp;
if replicate =. then delete;
drop _Type_ _freq_;
run;
proc sql;
create table lev1_comp2 as
select a.*, b.p_50z
from diss15_MAR a left join lev1_comp b
on a.replicate=b.replicate;
quit;
data lev1_comp3;
set lev1_comp2;
xx1=x1; if x3>p_50z then x1_0=0; else x1_0=1;
xx2=x2;
run;
**Creating datafile with sample sizes;
proc sql;
create table N1size as
select avg(replicate) as replicate2,
round((count(l1_id))*((&l1miss/100)/2)) as _Nsize_
from diss15_mar
group by replicate;
quit;

*datafile must have the same variable as in strata line of proc
surveyselect so renaming;
data rn1size;
set N1size;
rename replicate2=replicate;
run;

**selecting X1 for deletion;
proc surveyselect noprint data=lev1_comp3 sampsize=rN1size method=pps
out=l1x1select;
```

```
size x1_0;
strata replicate; run;

data lev1_comp4;
set lev1_comp2;
xx1=x1;
xx2=x2; if x3<p_50z then x2_0=0; else x2_0=1;
run;
**selecting x2 for deletion;
proc surveyselect noprint data=lev1_comp4 sampsize=rN1size method=pps
out=l1x2select;
size x2_0;
strata replicate; run;

**Combining datasets;
data l1x1select2; set l1x1select; xx1 =.; k1=0; keep k1 replicate l1_id
l2_id xx1 xx2;run;

data l1x2select2; set l1x2select; xx2 =.; k2=0; keep k2 replicate l1_id
l2_id xx1 xx2; run;

proc sort data = l1x1select2; by replicate l2_id l1_id;
proc sort data = l1x2select2; by replicate l2_id l1_id;
proc sort data = lev1_comp3; by replicate l2_id l1_id; run;

**Complete level-1 data set with missingness;
data lev1missing; merge l1x1select2 l1x2select2 lev1_comp3; by
replicate l2_id l1_id; run;


**Merging in level-2 data with complete set of generated data;
data level1data; merge diss15_MAR lev1missing(keep = replicate l1_id
l2_id k1 k2 xx1 xx2);
by replicate l2_id l1_id;
if k1 = . then k1=1;
if k2 = . then k2=1;
if k1 = 0 then xx1 = .;
if k2 = 0 then xx2 = .;
run;
%end;

%else %if &l1miss = 0 %then %do;
data level1data; set diss15_MAR;
xx1 = x1;
xx2= x2;
run;
%end;

%if &l2miss gt 0 %then %do;
**********************MAR at level-2;
proc sort data = diss15_MAR;
by replicate l2_id;
run;
proc means data = diss15_MAR noprint;
class replicate l2_id;
var z3 z1 z2; output out=b_comp mean(z3)=z3 mean(z1)=z1 mean(z2)=z2;
run;
```

```
data b_comp; set b_comp;
do i = 1 to &reps;
if replicate = i then t = i; end;
if l2_id =. or replicate = . then delete; run;

*restricting upper limit to 50%;
proc means data = b_comp noprint;
class replicate;
var z3; output out = b1_comp p50=p_50z;
run;

data b1_comp; set b1_comp;
if replicate = . then delete;
run;

data b2_comp; set b1_comp;
do i = 1 to &reps;
if replicate = i then t = i; end;
run;
data b3_comp; merge b_comp b2_comp;
by t; drop i _type_ _freq_; run;

data m1;
set b3_comp;
zz1=z1; if z3>p_50z then z0=0; else z0=1;
zz2=z2;
run;

proc surveyselect noprint data=m1 sampsize=%SYSEVALF(&n2*((&l2miss /
100)/2),integer) method=pps out=m3;
size z0;
strata replicate; run;

data m2;
set b3_comp;
zz1=z1;
zz2=z2; if z3<p_50z then z0=0; else z0=1;
run;
**selecting Z2 for deletion;
proc surveyselect noprint data=m2 sampsize=%SYSEVALF(&n2*((&l2miss /
100)/2),integer) method=pps out=m4;
strata replicate;
size z0;
run;

**Combining datasets;
data m5; set m3; zz1 =.; v1=0; keep v1 replicate l2_id zz1 zz2;
run;

data m6; set m4; zz2 =.; v2=0; keep v2 replicate l2_id zz1 zz2; run;

proc sort data = m5; by replicate l2_id;
proc sort data = m6; by replicate l2_id;
proc sort data = m1; by replicate l2_id; run;

**Complete level-2 data set with missingness;
```

```
data m7; merge m1 m5 m6; by replicate l2_id; run;
proc sort data = level1data; by replicate l2_id; run;

**Merging in level-2 data with complete set of generated data;
data Mar_&cond; merge level1data m7(keep = replicate l2_id v1 v2 zz1
zz2);
by replicate l2_id;
if v1 = . then v1=1;
if v2 = . then v2=1;
rename
xx1 = x1_MAR
xx2 = x2_MAR
zz1 = z1_MAR
zz2 = z2_MAR;
run;
%end;

%else %if &l2miss = 0 %then %do;
data MAR_&cond; set level1data;
rename
xx1 = x1_MAR
xx2 = x2_MAR;
z1_MAR = z1;
z2_MAR = z2;
run;
%end;
```

APPENDIX C – EXAMPLE MPLUS INPUT FILE

```
TITLE: This is an example of imputing a multilevel model in MPLUS. Mechanism = MCAR
       condition = 162 rep = 376
DATA:     FILE = impMCAR162_376.dat;
VARIABLE:  NAMES are rep cond l1_id l2_id x1_MCAR x2_MCAR x3 x4 z1_MCAR z2_MCAR z3 y1;
       USEVARIABLES are x1_MCAR x2_MCAR x3 x4 z1_MCAR z2_MCAR z3 y1;
       cluster = l2_id;
       between = z1_MCAR z2_MCAR z3;
       within = x1_MCAR x2_MCAR x3 x4;
       missing = ALL .;
       IDVARIABLE = l1_id;
ANALYSIS: Type= TWOLEVEL BASIC;
DATA IMPUTATION:
       IMPUTE = x1_MCAR x2_MCAR z1_MCAR z2_MCAR;
       NDATASETS=20;
       SAVE = I:\Dissertation\MCAR IMPUTE\impMCAR162_376_*.dat;
       thin = 500;
```