4-2008

Missing Data Treatment of a Level-2 Variable in a 3-Level Hierarchical Linear Model

Xiaofan Cai
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://scholarworks.wmich.edu/dissertations/756

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
MISSING DATA TREATMENT OF A LEVEL-2 VARIABLE IN A 3-LEVEL HIERARCHICAL LINEAR MODEL

by

Xiaofan Cai

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Educational Leadership, Research and Technology
Dr. Brooks Applegate, Advisor

Western Michigan University
Kalamazoo, Michigan
April 2008
MISSING DATA TREATMENT OF A LEVEL-2 VARIABLE IN A 3-LEVEL HIERARCHICAL LINEAR MODEL

Xiaofan Cai, Ph.D
Western Michigan University, 2008

Data used in educational research often come with a hierarchical structure such as students nested in classrooms and classrooms nested in schools. Hierarchical linear model (HLM) analysis allows applied researchers to incorporate the hierarchical structure of the data into data analysis to examine effects of variables at each level. However, problems such as missing data pose analytical challenges of biased estimation. With missing data occurring in level-2 variables in a 3-level HLM analysis, the choice of the missing data treatment may affect parameter estimation at all levels.

This Monte Carlo simulation study was designed to compare performance of six missing data treatment (MDT) methods—listwise deletion, mean substitution, restrictive Expectation-Maximization (EM), inclusive EM, restrictive multiple imputation (MI) and inclusive MI in generating unbiased estimates in a 3-level HLM model. An “intercept-only” 3-level HLM model was adopted. Missingness was generated as missing at random (MAR) for a level-2 predictor variable. The six MDTs were applied and the imputed datasets were analyzed using the same HLM model. Parameter estimates from the imputed datasets were compared to those obtained from
the complete datasets. The comparisons focused on the accuracy and precision of parameter estimates of fixed and random effects in the HLM model.

Results revealed that every MDT method produced more biases in the estimates with high proportion of missingness, and their performances improved as the level-2 sample size increased. Listwise deletion was a viable choice when level-2 sample size was small, it generated the most accurate but less precise estimates. With medium and large sample sizes, the restrictive EM method was effective in producing accurate and precise estimates for fixed effects parameters at all levels. The inclusive EM method outperformed all other methods in producing accurate and precise estimates for random effects. The two MI methods did not produce satisfactory estimates for level-2 fixed effects. However, the inclusive MI outperformed the restrictive MI on level-2 estimates of both fixed and random effects across the study conditions.

This study provides statistical evidence and practical recommendations for researchers who must consider different MDT methods when they encounter missing data in hierarchical data structures.
ACKNOWLEDGMENTS

The completion of this dissertation would not be possible without the help and support from faculty, friends and family. I would first like to express my deepest respect and gratitude to my advisor and dissertation committee chairperson Dr. Brooks Applegate, for his mentoring, guidance, support, patience and feedback throughout my doctoral study, and for challenging me to learn more than I thought I could. Special thanks go to my dissertation committee members: Dr. Susan Carlson for her expertise, valuable suggestions and insightful comments; Dr. Warren Lacefield for his valuable input, thought-provoking questions and encouragement.

I would like to extend my thanks to Dr. Paula Kohler, for supporting me and providing me critical resources to complete this study. I also would like to thank Dr. Jianping Shen and Dr. Ineke Way, for their encouragement and support throughout this process. Thanks to all the staff at Career Connections Research Center (CCRC), for their support, encouragement and friendship. Thanks to my fellow EMR graduate students, for their help and friendship.

Last but not least, I would like to give my special thanks and appreciation to my parents, Zhongmin Cai and, Erlian Zhong, all of my family members and my close friends who have given me love and support and have always believed in me.

Xiaofan Cai
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ ii

LIST OF TABLES .............................................................................................................. vii

LIST OF FIGURES ............................................................................................................ viii

CHAPTER

I. INTRODUCTION .......................................................................................................... 1
    Statement of Problem ................................................................................................. 1
    Three-level Hierarchical Linear Model .................................................................... 3
    Mechanism of Missing Data ..................................................................................... 4
    Methods of Missing Data Treatment (MDT) ............................................................. 5
        Listwise Deletion .................................................................................................. 5
        Mean Substitution ................................................................................................. 6
        Expectation-Maximization (EM) Algorithm Imputation ...................................... 6
        Multiple Imputation (MI) .................................................................................... 6
        Inclusive and Restrictive Imputation .................................................................... 7
    Experimental Conditions .......................................................................................... 7
    Level-2 Sample Size ................................................................................................. 7
    Proportion of Missingness ...................................................................................... 8
    Study Model and Research Questions ...................................................................... 8
    Significance of the Study .......................................................................................... 10
    Limitations of the Study .......................................................................................... 11
Table of Contents—continued

CHAPTER

Definitions .................................................................................................................. 12

II. THEORY AND LITERATURE REVIEW ..................................................... 15

The 3-level Hierarchical Linear Model .............................................................. 15

The Need for HLM Models ............................................................................. 15

The Fully Conditional Model ........................................................................ 16

The Conditional Model ("Intercept-only" or "Random Intercept" Model) .......... 18

The Advantages of HLM Model ..................................................................... 22

Mechanism of Missing Data .......................................................................... 23

Types of Missing Data ..................................................................................... 24

Methods of Missing Data Treatment .............................................................. 26

Listwise Deletion ............................................................................................. 26

Mean Substitution ............................................................................................. 27

Expectation-Maximization (EM) Algorithm ................................................. 27

Multiple Imputation (MI) .............................................................................. 28

Inclusive and Restrictive Strategies ................................................................. 31

Missing Data Treatment in HLM ................................................................. 32

III. METHODOLOGY ....................................................................................... 38

Study Design ..................................................................................................... 38

The 3-level Hierarchical Linear Model and Parameters .......................... 38

Experimental Design ...................................................................................... 41

Summary of Experimental Design ................................................................. 46
Table of Contents—continued

CHAPTER

Data Generation .......................................................... 47

Research Questions ....................................................... 51

Data Analysis .............................................................. 52

Analysis of Parameter Estimate Bias (PEB) ......................... 52

Analysis of Accuracy and Precision of the Estimates .......... 55

IV. RESULTS ................................................................... 57

Parameter Estimate Bias (PEB) .......................................... 57

PEB of Fixed Effects ....................................................... 58

PEB of Random Effects ................................................... 67

Summary of Parameter Estimate Bias ............................... 70

Accuracy and Precision of Parameter Estimates .................. 71

Accuracy and Precision of the Estimates of Fixed Effects .. 72

Accuracy and Precision of the Estimates of Random Effects .................................................................. 77

Further Investigation on Parameter Estimates .................... 79

Summary of Accuracy and Precision of the Estimates ........ 90

V. DISCUSSION ................................................................ 93

The Effect of Sample Size ................................................ 93

The Effect of Level-2 Total Sample Size ......................... 93

The Effect of Level-2 Group Size ..................................... 95

Performance of Missing Data Treatment (MDT) ................. 97

Listwise Deletion ........................................................... 97
## Table of Contents—continued

### CHAPTER

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Substitution and Restrictive EM Imputation</td>
<td>98</td>
</tr>
<tr>
<td>Inclusive EM Imputation</td>
<td>99</td>
</tr>
<tr>
<td>Restrictive MI and Inclusive MI</td>
<td>99</td>
</tr>
<tr>
<td>Restrictive and Inclusive Methods</td>
<td>100</td>
</tr>
<tr>
<td>Summary of Results and Conclusions</td>
<td>102</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>104</td>
</tr>
<tr>
<td>Future Research</td>
<td>105</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>106</td>
</tr>
</tbody>
</table>

### APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. SAS Program for Monte Carlo Simulation</td>
<td>119</td>
</tr>
<tr>
<td>B. Generalized Omega Squared Calculation</td>
<td>128</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1. Parameters for the random intercept model ........................................... 41
2. Design of sample size ........................................................................ 45
3. Design of the study ........................................................................... 46
4. ICC in conditional HLM model .............................................................. 47
5. Descriptive statistics of the generated data .......................................... 49
6. Parameters specified and actual parameter estimates generated in the data ............................................................................ 49
7. ANOVA (5x2x6) results for parameter estimate bias (PEB) ................. 59
8. ANOVA (5x2) results for parameter estimate bias (PEB) by MDT .......... 60
9. Mean values of parameter estimate bias (PEB) for fixed effects .......... 61
10. Mean values of parameter estimate bias (PEB) random effects .......... 62
11. Root-mean-square error (RMSE) for fixed effects estimates .............. 73
12. Root-mean-square error (RMSE) for random effects estimates .......... 74
13. Coverage and width of fixed effects estimates ..................................... 84
14. Coverage and width of random effects estimates ............................... 86
15. Comparison between the effects of total size and group size for level-3 estimates ........................................................................... 96
LIST OF FIGURES

1. PEB of parameter estimates (fixed effects) with 95% CI under 10% missingness ................................................................. 81

2. PEB of parameter estimates (fixed effects) with 95% CI under 40% missingness ................................................................. 82

3. PEB of parameter estimates (random effects) with 95% CI under 10% missingness ................................................................. 83

4. PEB of parameter estimates (random effects) with 95% CI under 40% missingness ................................................................. 83
CHAPTER I
INTRODUCTION

Statement of the Problem

In many educational and organizational settings the research involves hierarchical structures, for example, students are nested in classes or teachers and classes/teachers are nested in schools, thus hierarchical linear models or multilevel models are appropriate to incorporate the hierarchical structure into analyses. Since their introduction, hierarchical linear models (HLM) have been utilized in a large number of educational and behavioral research practices (Bryk & Raudenbush, 1987; Bosker, Kremers & Lugthart, 1990; Raudenbush & Willms, 1991; Raudenbush & Bryk, 1993; Muthén, 1994; Goldstein, 1997; Hargrove & Mao, 1997; Kyriakides, Campbell, & Gagatsis, 2000; Xue, 2002; Archibald, 2005; Franco, Lee & Satyro, 2005). One of the advantages of HLM is that it allows researchers to simultaneously investigate relationships within a particular level, as well as relationships between the levels. Moreover, HLM models allow researchers to partition out variance components at different levels with other predictor variables and to examine the variance proportions of each source.

Large-scale research projects, in which HLM analysis are appropriate, rarely occur with no missing data because they often require data to be collected from multiple informants. For example, in a hierarchical model with students nested in teachers, and teachers nested in schools (3-level HLM model), data need to be collected on students (e.g., students’ achievement), on teachers (e.g., teaching characteristics), and on schools (e.g., school characteristics). Thus, due to the multilevel data, there may be a high probability of missing data. If a school fails to record some students’ scores, then there
would be missing data at level-1; if a teacher does not return the survey by which
information about teacher characteristics needs to be collected, then we could have
missing data at level-2; if a school fails to report its student enrollment number or certain
characteristics of its student population, then the missing data would occur at level-3.
Thus, the problem of missing data is an issue researchers frequently encounter.

In the current practice of HLM analyses, the commonly used software such as
HLM 6 (Raudenbush, Bryk & Congdon, 2005) or SAS PROC MIXED (SAS Institute,
2004) requires that missing data need to be treated before data analysis, either by listwise
deletion or imputation. In many HLM studies, the dominant missing data method has
been listwise deletion (Bosker et al., 1990; Hill & Rowe, 1996; Kyriakides et al., 2000;
Marks, 2000; Finn, Gerber, Achilles, & Boyd-Zaharias, 2001; Opdenakker, 2001; Lamb,
2002; Xue, 2002; Archibald, 2005; Desimone, 2005; Goddard, 2007), only a couple of
HLM studies utilized the data imputation approach (Hill, Rowan & Ball, 2005; Correnti
& Rowan, 2007). In 1999, a report by APA Task Force on Statistical Inference
(Willkinson & Task Force on Statistical Inference, 1999) specifically pointed out that “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” (p. 598).

In a HLM model with missing data occurring at level-2, the dominant listwise
deletion strategy drops all the cases that are nested in the missing level-2 unit. This
deletion results in a marked reduction in the total sample size. However, if missing data
can be treated by imputation, a researcher could retain the original sample size and the
full statistical power of the design. A few studies (Prosser, 1991; Gibson & Olejnik, 2003;
Zhang, 2005; Dalton & Schulz, 2006) have focused on the comparison of missing data treatments (MDTs) with missingness at either level-1 or level-2 in 2-level HLM models. However, there is no study in the literature that addresses the performance of MDTs with missingness at level-2 variable in a 3-level HLM model. Thus, a question can be posed: in a 3-level HLM model with missing data at level-2, how does the choice of missing data treatment affect the parameter estimates in relation to other factors such as sample size or the amount of missing data?

Three-level Hierarchical Linear Model

Consider a 3-level HLM research design which investigates the effect of different instructional methods on student achievement. At level-1, there are student achievement scores, at level-2, students are nested in teachers, at level-3, teachers are nested within schools. The three-level HLM allows researchers to assess the variation between-students and between-classes, as well as the variation among schools; thus the variance is partitioned into three pieces—variation due to individual student differences, variation due to teacher differences and variation due to school differences.

The three-level hierarchical linear model adopted in the current study was a “random intercept” or “intercept only” model based on the “intercept-and-slopes-as outcomes” hierarchical model (Bryk & Raudenbush, 2002, p. 80). The “intercept only” model means only intercepts at each level are predicted by higher level predictors. The model has a structure of 3 levels: students, classes/teachers and schools, representing level-1, level-2, and level-3 respectively.
Mechanism of Missing Data

In the statistics literature on missing data, missingness is usually treated as a probability problem (Little & Rubin, 2002). According to Rubin (1976), there are generally three types of missingness based on the relationship of the missing data to the complete data.

**Missing Completely at Random (MCAR).** Missing data are MCAR when the distribution of missingness does not depend on the observed data, that is, the probability of an observation $X_i$ being missing is unrelated to the value of $X$, or unrelated to the value of any other variables in the study, and it is missing completely at random. Such a situation occurs if, for example, a question is accidentally, not printed on the survey questionnaire by a printing error, and thus there is missing data. If data are MCAR, then the same results should be obtained had there been no missing data. In practice, MCAR means that the analysis of available cases gives valid inferences at the expense of a loss of power.

**Missing at Random (MAR).** For data to be MAR, the probability of missingness does depend on observed data, but not on missing data. This means that an observation $X_i$ is missing is unrelated to the value of $X$ itself, but is related to the value of $Y$. In other words, missingness does not depend on the value of $X$ after controlling for another variable or variables (Howell, 2007). For example, in a survey of self-rated job satisfaction levels, new teachers may be more likely not to report their satisfaction. Then, the missing data on this variable is dependent on teachers' years of experience. Thus, MAR could be identified and controlled by another variable or variables. A special case of MAR is MCAR. If data are MAR, then there must be a relationship underlying the
missingness, and that relationship must be modeled with other observed data before data analysis.

*Missing Not at Random (MNAR).* When neither MCAR nor MAR can be established, the probability of an observation $X_i$ being missing does depend on $X_i$ itself, and the missing data are defined as MNAR. For example, in the job satisfaction survey, teachers with lower satisfaction levels may be more likely not to reveal their satisfaction level. Thus the mean rating of job satisfaction over available responses in the study would be an over-estimate of the true value if we would have obtained the complete data. When data are MNAR, more complex models need to be developed to estimate the missing values (Howell, 2007) and this lies beyond the scope of this study.

**Methods of Missing Data Treatment (MDT)**

Researchers have explored various strategies of dealing with missing data (Little & Rubin, 1987; Allison, 2001). The most commonly practiced MDT methods include listwise deletion, mean imputation, Expectation-Maximization (EM) imputation (Dempster, Laird, & Rubin, 1977) and multiple imputation (MI) (Rubin, 1987; Schafer, 1997a).

**Listwise Deletion**

Listwise deletion, or case deletion, is the most common approach for treating missing data. It is used by default in many statistical programs, and thus by many researchers. The analysis procedure drops the whole case if there is missing data on any variable in the analysis, and thus only the remaining “complete” cases are analyzed. The main advantage of listwise deletion is its simplicity. The disadvantage lies in the loss of
information and statistical power. If the percentage of the missing data is small and MCAR can hold, this method can be quite efficient.

**Mean Substitution**

In the mean imputation procedure, missing values are replaced by the average (mean) of the observed values for that variable. By substituting the mean value for the missing value, the original sample size remains unchanged, however it results in reduced variance estimates in the affected variables and in any covariance estimate with another variable. This method requires MCAR.

**Expectation-Maximization (EM) Algorithm Imputation**

Developed by Dempster, Laird, and Rubin (1977), the EM algorithm is a maximum likelihood (ML) based method to impute missing data. It is an iterative process where the E-step calculates the expected log-likelihood value for complete data based on the observed data, and the M-step computes the ML estimate to maximize the log-likelihood value from the E-step. The process is repeated until the change in the estimated values is negligible and the iterations converge (Schafer & Olsen, 1998).

**Multiple Imputation (MI)**

The basic idea of multiple imputation developed by Rubin (1987) is to treat missing data as a random variable and replaced the missing data points with more than one value. In MI, each missing value is replaced by a list of \( m > 1 \) simulated values. By substituting the simulated value for the corresponding missing value, MI method produces \( m \) alternative versions of the complete data set. Each of the \( m \) data sets is then treated as a complete data set and analyzed under the same data analysis model. Then, "the various results from the complete-data analyses are then combined to obtain overall..."
estimates and standard errors that reflect finite sample variation” (Schafer & Graham, 2002).

Inclusive and Restrictive Imputation

 Several studies (Dalton & Schulz, 2006; Collins, Schafer & Kam, 2001; Schafer & Olsen, 1998) suggest that the variables used in MI process will affect the quality of the imputed data. Collins, Schafer and Kam (2001) found that excluding variables that should be included in MI generated biased parameter estimates. The rationale is that an imputation model should preserve the variance and covariance structure of the data that will be incorporated into subsequent analyses, including the associations and relationships among variables (Durrant, 2005). This study implemented both EM and MI methods using inclusive and restrictive strategies. Inclusive strategy involves using numerous variables in the imputation process and restrictive strategy uses few variables in the imputation process. In the current study, the inclusive approach utilized both level-1 and level-2 variables in the imputation process, while the restrictive approach only included level-2 variables in the imputation process for level-2 missing data. The restrictive and inclusive imputation approaches were only applied in the EM and MI methods.

Experimental Conditions

Level-2 Sample Size

Sample size is always a crucial factor in research studies because as the sample size increases, the power of the statistical analysis increases, and the variance estimates become less biased. In HLM studies, the occurrence of missing data would substantially change the sample size if listwise deletion were used. With MAR missingness at level-2,
listwise deletion would reduce the level-2 sample size as well as that of level-1 because of the nested structure. Additionally, the estimates based on the reduced level-2 sample size would also bias the estimates at level-3.

The present study was conducted under three level-2 sample size conditions, 60, 180 and 360 to reflect small, medium and large level-2 sample sizes found in the HLM literature with similar nested data structures (Goddard, 2007; Archibald, 2005; Lamb, 2002; Finn et al., 2001; Opdenakker, 2001; Kyriakides et al., 2000). Since the level-2 total size is a product of level-3 total size and level-3 group size, therefore, two ways of sample size design were used—A) to fix the level-3 group size with varying level-3 total size or B) to fix the level-3 total size with varying level-3 group size. The two ways of design resulted in five sample size conditions (see Chapter 3 for design details).

Proportion of Missingness

The percentage of missing data can impact the performance of missing data treatments. Research studies on missing data treatments employed various levels of missingness, ranging from 5% (Prosser, 1991) to 50% (Collins, Schafer & Kam, 2001; Dalton & Schulz, 2006). The proportions of missingness in this study were set at 10% and 40%, representing low and high levels of missingness and MAR missingness was produced.

Study Model and Research Questions

The purpose of this study was to compare the estimation bias generated from the six missing data treatment methods in the context of 3-level HLM models with MAR missing data at one level-2 variable. The hierarchical model was based on data from the student (level-1), class (level-2) and school levels (level-3). Following the notation of
Raudenbush & Bryk (2002), the "intercept-only" or "random intercepts" HLM model employed in the study is presented below (see Chapter 3 for detailed descriptions).

Level-1:
\[ Y_{ijk} = \pi_{0jk} + \pi_{1jk} a_{ijk} + e_{ijk} \]

Level-2:
\[ \pi_{0jk} = \beta_{00k} + \beta_{01k} X_{1jk} + \beta_{02k} X_{2jk} + r_{0jk} \]
\[ \pi_{1jk} = \beta_{10k} \]

Level-3:
\[ \beta_{00k} = \gamma_{000} + \gamma_{001} W_{1k} + u_{00k} \]
\[ \beta_{01k} = \gamma_{010} \]
\[ \beta_{02k} = \gamma_{020} \]
\[ \beta_{10k} = \gamma_{100} \]

Combined model:
\[ Y_{ijk} = \gamma_{000} + \gamma_{001} W_{1k} + \gamma_{010} X_{1jk} + \gamma_{020} X_{2jk} + \gamma_{100} a_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \]

Six missing data treatment (MDT) methods were studied—listwise deletion (LD), mean substitution (MS), restrictive EM (REM), inclusive EM (IEM), restrictive MI (RMI) and inclusive MI (IMI). The five fixed effects (\( \gamma_{000}, \gamma_{001}, \gamma_{010}, \gamma_{020}, \) and \( \gamma_{100} \)) and three random effects (\( u_{00k}, r_{0jk}, \) and \( e_{ijk} \)) parameter estimates produced by each MDT were examined.

The research questions addressed are:

1. When compared with complete data, do the parameter estimates of the fixed effects calculated from six MDTs (i.e., listwise deletion, mean substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield statistically significant difference on the amount of bias under different experimental conditions (sample size and proportion of missing data)?
2. When compared with complete data, do the parameter estimates of the random effects calculated from six MDTs (i.e., listwise deletion, mean substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield statistically significant different amount of bias under different experimental conditions (sample size & proportion of missing data)?

3. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of fixed effects that are accurate and precise when compared with parameter estimates from complete data?

4. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of random effects that are accurate and precise when compared with parameter estimates from complete data?

Significance of the Study

Data from educational research usually come with a hierarchical structure such as students nested in classes, and classes nested in schools. HLM analyses allow applied researchers to incorporate the hierarchical structure into the data analysis and to examine the effects of variables at each level. As the awareness of the advantages of this kind of analysis has gained popularity among researchers in the field, the number of studies that are employing HLM analysis has increased. However, problems such as missing data pose challenges in terms of producing biased estimates. In the 3-level HLM literature, data of level-2 variables are commonly collected through surveys or questionnaires which may involve a different data collection process from that of students (i.e., level-1) and
schools (i.e., level-3) data. With missing data occurring at level-2 in a 3-level HLM analysis, the choice of missing data treatment may affect parameter estimates of every level. This simulation study is designed to investigate the performance of six missing data treatment methods (listwise deletion, mean imputation, REM, IEM, RMI and IMI) for modeling parameter estimates as a function of level-2 sample size and percentage of missing data. For researchers who use HLM models, the findings of the study will provide statistical evidence related to different methods for dealing with missing data. The study also provides practical recommendations for researchers who will need to decide what MDT method to follow under different conditions when they encounter missing data issues in HLM analysis.

Limitations of the Study

There are several limitations of the study that must be acknowledged. First, this study only investigated limited numbers of sample sizes and amounts of missing data, which might not represent all situations in real practice. Second, the 3-level HLM model adopted in the study is a random intercept model. Cross-level interaction effects were not included in the model. Third, all the sample size conditions are balanced designs which might not be the case with real data. Finally, univariate normality is assumed for all predictor variables, additionally, multivariate normality is assumed in the imputation procedures. With the violation of multivariate normality, some imputation methods might perform slightly different. Therefore, the results of the study should be taken with caution when applying to different situations, with violation of multivariate normality or to more complex models.
Definitions

Hierarchical Linear Model (HLM)

An analysis strategy that allows for nested, multilevel, or hierarchical data structures, also referred to as multi-level linear models. For example, units at one level, like individuals, can be regarded as groups or clusters within units at a higher level, like communities.

Intercept-as-outcome Model

HLM models in which only the intercept is predicted from higher-level variables. Differences in intercepts represent mean differences in the dependent variable that can be predicted from independent variables from multiple levels.

Missing Completely at Random (MCAR)

The missing data for a variable Y are “missing completely at random” if the probability of having a missing value for Y is unrelated to the value of Y itself or to any other variables in the data set. More simply, this means that the missing data values are a simple random sample of all data values.

Missing at Random (MAR)

The missing data for a variable Y are “missing at random” if the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis. MCAR is a special type of MAR.

Missing not at Random (MNAR)

The missing data for a variable Y are “missing not at random” if the probability of missing data on Y is related to the value of Y itself even if we control for other variables in the analysis.
**Listwise Deletion**

It is also referred as complete case analysis or case deletion. Assuming incomplete cases are similar to complete cases, it restricts the analysis to those subjects with no missing data on variables of interest.

**Mean Substitution**

It is the technique that replaces missing data with the mean of non-missing values. It will underestimate the standard deviation and standard errors of the imputed variable.

**EM (Expectation-Maximization) Algorithm Imputation**

The EM algorithm, introduced by Dempster, Laird, and Rubin in 1977, is a technique that finds maximum likelihood estimates in parametric models for incomplete data. It is used to locate the mode or modes of the likelihood function or of the posterior density. By using E-steps (expectation) and M-steps (maximization) alternatively and iteratively, one is sure to get the maxima under some proper conditions.

**Multiple Imputation (MI)**

Developed by Rubin (1987), multiple imputation is a Monte Carlo technique in which the missing values are replaced by m>1 simulated versions (Schafer, 1999). In multiple imputation, missing values for any variable are predicted using existing values. The predicted values are substituted for the missing values, resulting in a full imputed data set. The process is repeated multiple times, producing multiple imputed datasets. Statistical analysis is performed on each imputed dataset as complete dataset, and an overall analysis is conducted to combine the multiple analysis results.
Accuracy

It refers to the difference between the estimate and the true parameter. In this study, it is calculated as the amount of deviation between the average of the estimates and the true parameter.

Precision

It refers to the amount of dispersion of the estimates scattered around the true parameter. An index of precision is the standard deviation of all the estimates generated under a particular condition.
CHAPTER II
THEORY AND LITRATURE REVIEW

The 3-level Hierarchical Linear Model

The Need for HLM Models

Students within same class tend to be more similar to each other than students randomly sampled from the entire population. As members of a school or class influence one another, they develop shared understandings, educational practices, and mechanisms of decision making (Bidwell & Kasarda, 1997). In the aggregate, these relations partly define the characteristics as a unique group that affects each person in the group. For example, students in a particular fifth-grade classroom are more similar to each other than to students randomly sampled from other schools, or from the national population of fifth-graders. Students within a particular classroom tend to come from a community or community segment that is more homogeneous in terms of morals and values, and even educational preparation than the population as a whole. Further, students within a particular classroom share the experience of being in the same environment—the same teacher, physical environment, and similar experiences, which may lead to increased homogeneity over time.

Individuals who are drawn from an institution, such as a classroom, school, business, or health care unit, will be more homogeneous than if individuals were randomly sampled from a larger population. Because these individuals tend to share similar characteristics (e.g., environmental, background, experiential, demographic, or otherwise), observations based on these individuals are not fully independent. However, most analytic techniques require independence of observations as a primary assumption.
for the analysis. Because this assumption is violated in the presence of hierarchical data, ordinary least squares (OLS) regression (e.g., ANOVA, multiple regression) produces standard errors that are too small unless design effects are incorporated into the analysis. In turn, this leads to a higher probability of rejection of a null hypothesis than if: (a) an appropriate statistical analysis is performed, or (b) the data included are truly independent observations (Bryk and Raudenbush, 2002; Draper, 1995).

The Fully Unconditional Model

In the fully unconditional HLM model, the estimation starts without predictor variables at any level. It is used to assess the initial proportion of variance at each level. Consider a simple 3-level model where students are nested in classrooms within schools, and the dependent variable is students' math achievement score. The 3-level model partitions the variance of the outcome (dependent) variable into three components: level-1 variance component—among students within classes, level-2 variance component—among classes within schools, and level-3 variance component—among schools. The following presentation of HLM is drawn from Bryk and Raudenbush (2002). The notations $i, j,$ and $k$ denote students (level-1), classrooms (level-2), and schools (level-3) where

- $i = 1, 2, \ldots, n_{jk}$ students within classroom $j$ in school $k;
- j = 1, 2, \ldots, J_k$ classrooms within school $k$; and
- $k = 1, 2, \ldots, K$ schools.

The model at each level and the combined model would be:

Level-1 (student level):  
\[ y_{ijk} = \pi_{0jk} + e_{ijk} \quad [2.1] \]

Level-2 (class level):  
\[ \pi_{0jk} = \beta_{00k} + r_{0jk} \quad [2.2] \]
Level-3 (school level):

\[ \beta_{00k} = \gamma_{000} + u_{00k} \]  \hspace{1cm} [2.3]

Combined:

\[ y_{jk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{ijk} \]  \hspace{1cm} [2.4]

Where

- \( y_{jk} \) — math achievement score for student \( i \) in class \( j \) and school \( k \).
- \( \pi_{0jk} \) — mean math score of class \( j \) in school \( k \).
- \( \beta_{00k} \) — average math score across classes in school \( k \).
- \( \gamma_{000} \) — grand mean of the math score across all schools.
- \( e_{ijk} \) — the random "student effect" which is normally distributed with a mean of 0 and variance of \( \sigma^2 \).
- \( r_{0jk} \) — the random "class effect" which is normally distributed with a mean of 0 and variance of \( \tau_{\pi} \).
- \( u_{00k} \) — the random "school effect" which is normally distributed with a mean of 0 and variance of \( \tau_{\beta} \).

At level-1 (student level, or within-classroom level), the math achievement for each child is modeled as a function of a classroom mean (\( \pi_{0jk} \)) plus a random error (\( e_{ijk} \)).

At level-2 (classroom level or between-class-within-school level), each classroom mean (\( \pi_{0jk} \)) serves as an outcome variable varying around school mean (\( \beta_{00k} \)) with random errors (\( r_{0jk} \)). At level-3 (school level or between-school level), the school means (\( \beta_{00k} \)) are modeled as randomly varying with (\( u_{00k} \)) around a grand mean (\( \gamma_{000} \)). When Equations 2.1 to 2.3 are combined, we get Equation 2.4 which represents the simplest 3-
level fully unconditional HLM models and allows estimation associated with the three levels—students, classes and schools.

The Conditional Model ("Intercept-only" or "Random Intercept" Model)

When we hypothesize that part of the variability at each level could be explained by relevant variables, for example, student characteristics, classroom characteristics and/or school characteristics, we will want to include predictor variables at each level, then we will have a conditional 3-level HLM model. The model presented in the next section uses $a_{ijk}$, $X_{ijk}$ and $W_{jk}$ representing one student, one classroom, and one school characteristic variable, respectively.

Level-1 model—student level (the between-student-within-classroom model). The level-1 model of the 3-level HLM analysis is a within-classroom model to estimate the relationships between students' outcomes and their predictor variables (e.g., students' motivation) as well as the variance among the students within classroom:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} a_{ijk} + \epsilon_{ijk}$$  \[2.5\]

$Y_{ijk}$—math achievement score for student $i$ in classroom $j$ and school $k$.

$\pi_{0jk}$—intercept for classroom $j$ and school $k$, i.e., the average math score in class $j$ adjusted for student characteristic $a_{ijk}$ (motivation).

$\pi_{1jk}$—coefficient of the association between $a_{ijk}$ (motivation) and the math score in classroom $j$ and school $k$.

$\epsilon_{ijk}$—random student (level-1) effect, i.e., residual of student $ijk$'s score from predicted score based on the model, which is normally distributed with $\epsilon_{ijk}$ [0~$N(\sigma^2)$].

Level-2 model—classroom level (the between-classroom-within-school model).

The level-2 of HLM analyses is about between-classroom model (level-2) in which we
can evaluate effects of classroom characteristics (e.g., class climate or teacher evaluation score) on the average classroom score (the intercept at level-1). Thus, the classroom mean scores are modeled as function of teacher and/or classroom characteristics. The class-level model can be presented by the following equation where we have $X_{1,jk}$ as classroom variable.

$$\pi_{0,jk} = \beta_{00k} + \beta_{01k} X_{1,jk} + r_{0,jk} \quad [2.6]$$

$$\pi_{1,jk} = \beta_{10k} \quad [2.7]$$

where

$\beta_{00k}$ — average math score for school $k$.

$\beta_{01k}$ — average $X_{1,jk}$ effect in school $k$, allowed to vary among classrooms but is not predicted by any classroom variable.

$r_{0,jk}$ — random class effect within school, which is normally distributed with $r_{0,jk} [0 \sim N(\tau_a)]$.

As shown in this model, only one random effect is to be estimated—the level-1 intercept ($\pi_{0,jk}$), that is, the average class math scores are allowed to vary within schools. The student characteristic variable at level-1 (motivation) is estimated as a fixed effect, that is, the between-classroom variances of their relationships to the math score are constrained to zero. Thus, in this intercept-only model, we only investigate whether the average class math score ($\pi_{0,jk}$) varies among classes within each school.

**Level-3 model—school level** (the between-school model). At this level, the average class scores are modeled as a function of the school characteristic $W_{ik}$ (e.g., school size). In the school-level model, $\beta_{00k}$ and $\beta_{10k}$ are predicted at the school level, and they
become outcome variables, but only average school scores ($\beta_{00k}$) are allowed to vary across schools, while $\beta_{10k}$ is constrained to be constant, which means we assume the $X_{1jk}$ effect on class average is the same across schools.

$$\beta_{00k} = \gamma_{000} + \gamma_{001} W_{ik} + u_{00k}$$  \hspace{1cm} [2.8]

$$\beta_{01k} = \gamma_{010}$$  \hspace{1cm} [2.9]

$$\beta_{10k} = \gamma_{100}$$  \hspace{1cm} [2.10]

$\gamma_{000}$ — grand mean of math score among schools.

$\gamma_{001}$ — coefficient of the association between school average score and $W_{ik}$ (e.g., school size).

$u_{00k}$ — random school effect which is normally distributed with $u_{00k} \sim N(0, \tau_{00})$.

The combined model. When Equations 2.5 to 2.10 all combined, we get Equation 2.11 which represents a combined 3-level conditional HLM model with $a_{ijk}$, $X_{ijk}$ and $W_{ik}$ as single student, classroom and school variables at level-1, level-2 and level-3 respectively.

$$Y_{jk} = \gamma_{000} + \gamma_{001} W_{ik} + \gamma_{010} X_{ijk} + \gamma_{100} a_{ijk} + u_{00k} + u_{01k} + e_{ijk}$$  \hspace{1cm} [2.11]

The intercept-only model as presented in Equation 2.11 assumes that only intercept varies among classes (level-2) and schools (level-3). In this intercept-only model, the intercept at level-1 serves as the outcome variable at level-2, and the intercept at level-2 serves as the outcome variable at level-3. Thus, only the intercept at lower level will be predicted by higher level variables, that is, the level-1 intercept is regressed on level-2 variables, and the level-2 intercept is regressed on level-3 variables. All
regression coefficients except the intercepts are constrained to be constant within classes (level-2) and within schools (level-3).

Random effects and intra-class correlation (ICC). This 3-level model partitions the total variability in the outcome variable \( Y_{ijk} \) into three variance components—a) among students within classrooms (level-1)—\( \sigma^2 \), b) among classrooms within schools (level-2)—\( \tau_\pi \), c) among schools (level-3)—\( \tau_\beta \).

With estimates of these variance components, we are able to calculate the proportion of variance within classrooms, among classrooms within schools, and among schools (Raudenbush & Bryk 2002, p. 230):

\[
\frac{\sigma^2}{\sigma^2 + \tau_\pi + \tau_\beta} \text{ is the proportion of variance within classrooms;} \quad [2.12]
\]

\[
\frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta} \text{ is the proportion of variance among classrooms within schools;} \quad [2.13]
\]

\[
\frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta} \text{ is the proportion of variance among schools.} \quad [2.14]
\]

In an intercept-only two-level HLM model, the intraclass correlation (ICC) is an indication of the proportion of variance at the second level, and Hox (2002) suggests that it can be interpreted as the expected correlation between two randomly chosen individuals within the same group. He then notes that in three-level hierarchical models, one of the methods of calculating ICC at class and school levels as found in Davis & Scott (1995) is:

\[
\rho_{\text{class}} = \frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta} \quad [2.15]
\]

and \[
\rho_{\text{school}} = \frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta} \quad [2.16]
\]
Notice that Equation 2.13 and 2.15 are exactly the same, and so are Equation 2.14 and 2.16. Thus, in this study, “proportion of variance” and “ICC” would be used interchangeably. \( \rho_{\text{class}} \) indicates the proportion of class-level variance compared to the estimated total variance, and \( \rho_{\text{school}} \) indicates the proportion of school-level variance compared to the estimated total variance. They represent the likelihood that two elements in the same level have the same value. Specifically, \( \rho_{\text{class}} \) indicates the homogeneity of students in the classroom, as the value closes to zero, students in the same classroom tend to be more similar. \( \rho_{\text{school}} \) indicates the homogeneity of students in one school, as the value goes to zero, students in the same school tend to be more homogenous.

The Advantages of HLM Model

This “intercept-as-outcome” model allows researchers to investigate the effect of teacher or class characteristics, as well as school characteristics, on students’ average achievement. The model is appropriate for answering questions like: “Do some classes have higher average achievement scores than other classes in the same school?” “Do some schools have higher average achievement scores than other schools?” and “Do some class or school variables have different effects on students’ achievement?” Additionally, researchers are able to partition the variance into 3 difference pieces and find out which source contributes the largest variance even after controlling for student-, classroom- and school-level factors.

Though it is not within the scope of this study, another advantage of the HLM model is that the slope coefficients (e.g., \( \pi_{1, \cdot}, \beta_{1, \cdot} \), etc.) may also be assumed to vary across classes or schools. With the slope variation, we would hypothesize that a class characteristic (e.g., teacher attitude) is affecting students in the same classroom.
differently or a school characteristic (e.g., school size) is influencing students from the same school differently on their achievement. In other words, it would answer the question: "Is there a difference in the association between a class (or school) characteristic and achievement of students in the same class (or school)?"

Moreover, HLM analyses provide other benefits, such as easy modeling of cross-level interactions, which allows for more interesting questions to be asked of the data. It would allow researchers to answer a question like: "Is the association of teacher attitudes and achievement stronger in some schools than in others?" Thus, it tries to explore the effects of cross-level interactions between the characteristics of classes and the characteristics of schools.

In summary, HLM models have several advantages in educational studies: 1) they take into consideration of homogeneity and non-independence of observations due to the nested structure of the data; 2) they enable decomposition of any observed relationship into its between- and within-level components; 3) they allow for adjusting regression coefficients for other confounding variables and assessing the differential effect of a variable within a given level; and 4) they provide flexibility in evaluating cross-level interaction effect between variables of different levels.

Mechanism of Missing Data

In the statistics literature on missing data, missingness is usually treated as a probability problem (Little & Rubin, 2002). We can define \( R \) as the missingness in a given data set with missing data. The distribution of \( R \) could be mathematically described to capture the relationships between the missingness and the values that are actually missing. Statistically, we could consider \( R \) as a set of random variables having a joint
probability distribution (Schafer, 2002). In the literature, the distribution of missingness
or the probabilities of missingness is usually referred to as the probability distribution of
$R$. There are generally three types of missingness based on the relationship of the missing
data to the complete data (Rubin, 1976).

**Types of Missing Data**

The missing data mechanism can be denoted as the probability that a set of values
that are missing given the values taken by the observed and missing observations.

$$P(R | Y_{obs}, Y_{mis})$$

Where: $Y_{obs}$ — observed data

$Y_{mis}$ — missing data

$R$ — distribution of missingness

**Missing Completely at Random (MCAR).** When the distribution of missingness or
the probability of an observation being missing does not depend on the observed data or
unobserved data, which is denoted as

$$P(R | Y_{com}) = P(R) \quad \text{or} \quad P(R | Y_{obs}, Y_{mis}) = P(R)$$

Where: $Y_{com} = \{ Y_{obs}, Y_{mis} \}$ and $Y_{com}$ — complete data

Then, the missing data is missing completely at random (MCAR). That is to say, a
missing response occurs purely by chance. For example, a laboratory sample is
accidentally dropped and the observation is missing. Technically, missing data are said to
be MCAR if the probability of a missing response is independent of all the (measured and
unmeasured) characteristics of the individuals under study.

**Missing at Random (MAR).** Rubin (1976) defined missing data as MAR if

$$P(R | Y_{com}) = P(R | Y_{obs})$$
which means that when the data are missing at random, the probability of missingness does depend on observed data, but not on missing data. In other words, missingness does depend on other observed characteristics of the individuals, but missingness does not depend on the missing values. For example, people with lower educational levels would be more likely to not report their income in an interview. Thus, MAR could be identified and controlled by another variable or variables. MCAR is a special case of MAR where MCAR has a higher level of randomness.

**Missing Not at Random (MNAR).** When MCAR and MAR do not hold and the distribution of missingness does depend on $Y_{\text{miss}}$, the missing data are defined as MNAR. For example, in a survey study of health status, if we ask about blood pressure, people with higher blood pressure would be less likely to report the information.

If we have collected data on two variables—$X$ and $Y$, and we have data on variable $X$ for all participants while variable $Y$ has some missing data. If the probability that $Y$ is missing for a participant does not depend on his/her own values of either $X$ or $Y$, in other words, the observed values of $Y$ are a true random sample of all $Y$ values with no underlying bias, the missing data would be MCAR. If the probability that $Y$ is missing depends on $X$ but not $Y$, then the missing data would be MAR. That is, the observed $Y$ values represent a random sample of the actual $Y$ values for each value of $X$, but the observed values for $Y$ do not necessarily represent a true random sample of all $Y$ values. MNAR would be the case if the probability of missingness depends on $Y$, that is, the $Y$ value of a participant is missing due to certain characteristics of the $Y$ variable itself.
Methods of Missing Data Treatment

Researchers have explored the methods of treating missing data. The most commonly practiced missing data treatment methods include listwise deletion, mean imputation, regression estimates imputation, maximum likelihood imputation based on EM algorithm (ML) and multiple imputation (MI).

**Listwise Deletion**

Also known as case deletion, it is used by default in many statistical programs and thus by many researchers. This procedure deletes the whole case record if it has any missing data point relevant to an analysis. Performing complete-case analyses under MCAR, we would obtain consistent results as if there had been no missing data. In practice, if missing data are MCAR, the analysis of cases with complete data will give valid inferences with some loss in statistical power. The main advantages of listwise deletion are simplicity and consistency (Prosser, 1991). Simplicity lies in the operation of the analysis which could be directly conducted on the data after listwise deletion. Consistency follows from the use of the original case for the analyses rather than imputed information (Little & Rubin, 1987).

The drawback of the method is the loss of information and statistical power, especially with large-scale data sets with many variables, where the waste of information could be considerable. However, under MCAR assumption, no bias is introduced into estimation, but standard errors are larger than they would be if there were no missing data. Thus, if the percentage of the missing data is small, and MCAR can be assumed, this method can be quite efficient.
Mean Substitution

In the mean substitution procedure, missing values are replaced by the average (mean) of the observed values for that variable. The rationale is that “in the case of the normal distribution, the sample mean provides an optimal estimate of the most probable value” (Anderson, Basilevsky & Hum 1983, p. 456). This method requires MCAR missing. If data are MCAR, we could assume that the mean of the observed values is the expected value of the true value, thus replacing the missing data by that mean would not change the expected value for the variable.

The substitution of a mean value for missing value creates several problems in the analysis phase. First, mean substitution reduces the variance and the intercorrelation between variables. Second, mean substitution reduces the amount of observed covariation between predictor and outcome variables, distorts their linear relationships, and makes effects difficult to detect (Prosser, 1991; Marini, Olsen & Rubin, 1980).

Expectation-Maximization (EM) Algorithm

The Expectation-Maximization algorithm (Dempster, Laird & Rubin, 1977; Little & Rubin, 1990; Rubin, 1991) is a maximum likelihood estimation (MLE) of the missing values. It examines the relationships among variables intended for use in an analysis, and imputes values for the missing data using iterative approaches (Little & Rubin, 1987). Appropriate for MAR data, the EM involves E-step (expectation) and M-step (maximization). Given a set of parameter estimates, such as a mean vector and covariance matrix for a multivariate normal distribution, the E-step calculates the conditional expectation of the “complete-data” log likelihood based on the available complete cases and the parameter estimates. Then, given a “complete-data” log
likelihood value, the M-step finds the parameter estimates to maximize the log likelihood from the E-step. These two steps are repeated until the iteration convergence criterion is met (e.g., .0001) (Schafer & Olsen, 1998).

The advantage of the EM algorithm is that it solves a difficult incomplete-data problem by constructing two easy steps—computing the conditional-expectation of the log-likelihood with the incomplete data, and finding the maximizer of this expected likelihood. It takes the advantage of relationships in the data and uses that information to iteratively estimate the missing data.

In recent years, the EM method has been incorporated into many statistical software packages. The EM procedure is available in SPSS 10.0 and up and the program AMOS (Arbuckle, 2005) and LISREL (Jöreskog & Sörbom, 2001) compute ML estimates for structural equation modeling (SEM) data. Other programs, such as EMCOV (Graham & Hofer, 1991), Mx (Neale, Boker, Xie & Maes, 2003), Mplus (Muthén & Muthén, 2006) and MI procedure (SAS Institute, 2004) also offer EM estimation for missing data imputation.

**Multiple Imputation (MI)**

Multiple imputation, proposed by Rubin (1978, 1987), is a Monte Carlo technique to treat missing data (Schafer, 1999a). In MI, under multivariate normal distribution and assuming missing data are at least MAR, each missing value is replaced by a list of $m > 1$ simulated values. By substituting a simulated value from a predictive distribution for the missing values, the MI method produces $m$ alternative versions of the complete data set. Each of the $m$ data sets is then treated as a complete data set and analyzed under the same model. The various results from those "complete data versions are then combined to
obtain overall estimates and standard errors that reflect finite sample variation" (Schafer & Graham, 2002).

Schafer (1997a) suggested that for data with an arbitrary missing pattern a Markov Chain Monte Carlo (MCMC) method, an iterative imputation process, is used to impute the missing values. “MCMC is a collection of methods of simulating random draws from a predictive distribution via Markov Chains” (Schafer, 1999a, p. 8). Markov Chain is a series of random variables with a sequence. The sequence is formed in a way that the distribution of each element depends on the value of its previous one. In the MI process, “MCMC works to create a small number of independent draws of $Y_{mis}$ from a predictive distribution, and then these random draws are used for imputation” (Schafer, 1999a, p. 8). It involves two steps. In Imputation Step (I-step), MCMC, from a starting value\(^1\), simulates the missing values for each observation independently based on estimated mean vector and covariance matrix. In Posterior Step (P-step), it re-computes mean vector and covariance matrix with the simulated value from the I-step. The new estimates are then used in the next I-step. The two steps are iterated to create a Markov Chain long enough for a reliable distribution (Schafer 1997a; Schafer, 1999a) and the simulated value from the final iteration will be used for the missing value. By default, SAS PROC MI completes 200 iterations before the first imputation and 100 iterations between imputations (SAS Institute, 2004).

With $m$ imputations, the desired analysis can be conducted on each of the $m$ imputed data set. Then, a single parameter estimate will be produced by averaging the values from the $m$ data sets and standard error of the parameter estimated will be

\(^1\) EM posterior mode provides a good starting value (Schafer, 1997a, p. 87)
calculated as well. For a parameter estimate \( Q \), the combined point estimate for \( Q \) from multiple imputation is the average of the \( m \) complete-data estimates:

\[
\hat{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i
\]

The total variance of \( \hat{Q} \) is given by the formula (Rubin 1987):

\[
T = \bar{U} + (1 + \frac{1}{m})B
\]

Where: \( \bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i \)

\[
B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2
\]

As shown in the formula above, the total variance of \( \hat{Q} \) consists of two components—within imputation variance (\( \bar{U} \)) and between-imputation variance (\( B \)). Thus, the uncertainty is accounted for by different versions of the missing data and the variability between the imputed datasets (Wayman, 2003).

Software that can perform MI includes NORM (Schafer, 1999b) which conducts MI for incomplete data with arbitrary patterns of missing values, SAS PROC MI and MIANALYZE (SAS Institute, 2004), PAN (Schafer, 1997b) which performs MI for longitudinal data and clustered data, AMOS 7.0 (Arbuckle, 2006) which conducts MI for continuous and ordered categorical data, WinMICE (Jacobusse, Van Buuren & Groothuis-Oudshoorn, 2005) and MLwiN (Rasbash, Steele, Browne, & Prosser, 2004) which can deal with multilevel imputations.
Inclusive and Restrictive Strategies

For the EM and MI methods mentioned above, it is desirable to include as much information as the original dataset may contain into the imputation process, which can improve the performance of these missing data methods. Collins, Schafer and Kam (2001) defined restrictive strategy as including few or no auxiliary variables, and inclusive strategy as including numerous auxiliary variables in missing data imputation. Auxiliary variables refer to the variables included in the missing data procedure. As they pointed out that the restrictive strategy “might omit important correlates of missingness, whereas the inclusive one might err on the side of too many (variables)” (p. 332).

In the study of comparing inclusive and restrictive strategies with ML and MI, Collins, Schafer and Kam (2001) found that with 25% MAR missing, omitting a variable which correlates .9 with the cause of missingness resulted in substantial problems with large standardized differences between the estimates and the true parameters, large root-mean-square error (RMSE), and less coverage of nominal 95% intervals. Thus, they recommended adopting an inclusive approach because important causes of missingness are less likely to be omitted.

In HLM or multilevel studies, the missing data imputation process might include variables from each level (Carpenter & Goldstein, 2004; Jacobusse et al., 2005), which is the inclusive strategy in the context of hierarchical structures. Dalton and Schultz (2006) included inclusive EM imputation in comparisons of MDTs for missing data at level-2. Their inclusive strategy utilized one level-1 predictor and one level-2 predictor in the imputation process. Gibson and Olejnik (2003) noted that one of the limitations in their study is their use of only level-2 predictors in the imputation process while both level-1
and level-2 variables were included in their HLM analyses. Carpenter (2004) even suggested that the "imputation model should include all the structure that a researcher wishes to investigate in the model of interest." For example, covariates and interactions that are to be included in the data analysis on the full data set should be introduced in the imputation model, otherwise the imputed observations might not have this structure. The software currently could perform multilevel imputation including Schafer's PAN which is suitable for repeated hierarchical data, WinMICE (Jacobusse, Van Buuren & Groothuis-Oudshoorn, 2005) which is limited to linear models with a maximum of two levels, and MLwiN (Rasbash, Steele, Browne, & Prosser, 2004) which can impute multilevel missing data up to four levels.

**Missing Data Treatment in HLM**

In HLM studies, missing data can be very problematic. Specifically if missing data occur at higher levels (e.g., level-2) and a listwise deletion is adopted, then not only are the level-2 cases with missing data omitted, but all the nested level-1 cases are deleted, too. Thus, listwise deletion would affect the parameter estimation at every level—because of the change in sample size. For example, if we have 10 students nested in each of 100 classes and those 100 classes equally nested in 25 schools, the total sample size of students is 1000. With 30% missing data at class level, listwise deletion will result in 70 classes at level-2, the total number of students at level-1 would then decrease to 700. It might or might not reduce the number of schools in the study, but it certainly reduces the number of classes nested in each school and shrinks the variance within the school.

In the real world, missing data can easily occur in HLM studies and all of the HLM software require that the missing data be deleted or imputed before conducting the
analysis. If the amount of missing data is small and a researcher can assume the missingness is not related with the variable itself or other variables, the common practice of listwise deletion might not be a problem. However, if the percentage of missing data is not small and/or there is evidence that the variable with missing data is related with some other variables (i.e., missing data are MAR), data imputation needs to be considered. Listwise deletion with MAR missing would result in reduced statistical power due to the decrease in sample size and biased parameter estimates due to loss of information (Roth, 1994).

Peugh and Enders (2004) reported that among 545 articles published in the top professional journals of educational and psychological research in 2003, nearly half of them (42%, 229 articles) had missing data in the analyses. Among them, listwise deletion is the dominant method used to deal with missing data. Only 6 articles used ML estimation or MI to treat missing data. When the problem of missing data is presented in the HLM analysis, listwise deletion is also the most convenient method used by researchers since it is the default in computer software. Several HLM studies employed listwise deletion method (Bosker et al., 1990; Kyriakides et al., 2000; Marks, 2000; Finn et al. 2001; Opdenakker, 2001; Lamb, 2002; Xue, 2002; Archibald, 2005; Desimone, 2005; Goddard, 2007). The unfortunate side of listwise deletion is obvious that it changes the sample size as well as the sample composition (i.e., representativeness). Stated by Little and Rubin (2002), “since researchers are interested in making inferences about the entire target population, rather than the portion of the target population that we have information on all variables, the strategy of “complete case analysis” (listwise deletion) is generally inappropriate” (p. 6).
With the increasing attention to the missing data effect, some researchers in recent HLM studies have adopted imputation methods in treating missing data. In Hill, Rowan, & Ball’s study (2005) mean imputation is used with missing data on level-2 (teacher level) variables. Correnti and Rowan (2007) employed MI to treat the level-2 (class level) missing data where they utilized 80 level-2 variables in their imputation process.

Several missing data studies found in the literature used Monte Carlo simulation technique to compare different missing data treatments (MDTs). Among them, there were several studies that focused on the MDTs in single-level model (Roth & Switzer, 1995; Witta, 1992; Schafer & Granham, 2002; Collins, Schafer & Kam, 2001). In those studies, sample size, percent of missing data and missing data treatment were the major study variables. The findings of those studies were consistent in that when the sample size was large enough (e.g., at least 10 observations per predictor) and the percent of missing data is small (10%), listwise deletion performed well in estimating regression coefficients (Roth & Switzer, 1995; Witta, 1992), but when the sample size was not large (N=50) and the percent of missing data was high (30%), imputation methods such as regression imputation, maximum likelihood and multiple imputation tended to produce better estimates (Raymond & Roberts, 1987; Roth & Switzer, 1995). Raymond and Roberts (1987) concluded that imputation by regression procedures produced less biased estimates when the percent of missing data is more than 15% and when the sample size is less than 60 in the context of multiple regression analysis. Donner and Rosner (1987) reported that imputation using maximum likelihood procedures outperformed listwise deletion and regression methods in a univariate linear regression model. Kromrey and Hines (1994) investigated the performance of MDTs in regression analysis with MNAR.
They found that mean substitution, simple and multiple regression imputation produced biased estimates of $R^2$ and regression weights, while listwise and pairwise deletion generated accurate parameter estimation up to 30% of missingness.

Graham, Hofer and MacKinnon (1996) compared the performance of EM imputation, raw ML imputation, mean substitution and pairwise, deletion in treating all types of missingness. They found that the maximum likelihood and multiple imputation procedures clearly outperformed traditional or standard approaches such as listwise deletion and mean substitution under any type of missingness condition. They noted that ML approaches did not generate substantial bias under MCAR, MAR, and even under MNAR condition. Schafer and Graham (2002) reported that listwise deletion is unbiased in restoring the data under MCAR, but it generated substantial bias under MAR and MNAR. ML imputation and MI generated similar estimates, they did not produced substantial bias, though with some deviation, under MCAR or MAR, but they did not perform well under MNAR.

Since the HLM analyses are appropriate when the data have a multilevel or nested structure, researchers have become interested in studying missing data treatments in HLM analyses in recent years. However, there are a limited number of studies that address the performance of MDTs in HLM analysis. Prosser (1990) compared MDTs in a 2-level HLM model with MCAR missingness at a level-1 predictor. He found that listwise deletion is a better choice over mean substitution and regression imputation in terms of estimation bias in both fixed and random effects.

A couple of studies focused on the treatment of missing data in HLM models with missingness at higher levels (i.e., missingness does not occur at level-1). Using Monte
Carlo simulation techniques, Gibson & Olejnik (2003) analyzed a 2-level HLM model and compared MDTs in treating level-2 MCAR missingness. They reported that listwise deletion and EM imputation performed satisfactorily for fixed effects estimation, while mean substitution and MI were not effective in estimating parameters of fixed effects. With respect to random effects, only listwise deletion performed well except when the level-2 sample size was small (N=30) and the percentage of missingness is high (40%).

Based on the findings, they suggested that if MCAR could be assumed, researchers can rely on the default procedure of listwise deletion for estimation of fixed effects, and it is also recommended over imputation methods if the random effects were to be interpreted. However, they also recommended further analyses with missingness as MAR.

In another Monte Carlo study, Zhang (2005) examined the effects of the percentage of missing data, non-normality and sample size in 2-level HLM and SEM estimation with MAR missing. With hierarchical models, she found that nonnormality, severe or not, did not significantly affect the imputation performance even though multivariate normality is assumed. Multiple imputation is generally effective dealing with MAR missing data, but not under small sample size condition. With a higher proportion (30%) of missing data, MI tended to generate imputed data more deviant from the original complete data, and thus showed greater estimation bias.

Dalton and Schulz (2006) investigated the performance of listwise deletion, regression imputation, restrictive EM imputation and inclusive EM imputation in a 2-level HLM model with MAR missingness. They reported that with missingness on a level-2 predictor, listwise deletion did not perform well even with 10% missingness, and showed moderate bias in all parameter estimates at 30% and 50% missingness. The
restrictive EM imputation method outperformed mean replacement, listwise deletion and regression imputation in fully recovering the data, and in estimating both fixed and random effects and their standard errors, even with high level (50%) of missingness. However, the inclusive EM method generated more bias than the restrictive model. Thus they suggested that when missingness is encountered on a level-2 predictor, researchers should not rely on listwise deletion, but consider utilizing a restrictive EM imputation procedure based on level-2 data.

In the literature, missing data treatments (MDT) have been investigated fairly extensively in the context of single-level analysis such as regression analysis. However, the effects of MDTs in the context of HLM are somewhat less explored, and all the studies were focusing on 2-level models. There has not been a study that examined the effect of MDTs in a 3-level hierarchical model. The current study was designed to investigate the effects of MDTs and their relation to other factors—sample size and percentage of missing data, in a 3-level hierarchical model. This study provides empirical evidence for researchers to consider when they encounter missing data in HLM analysis.
This Monte Carlo simulation study was designed to compare the performance of six different missing data treatment (MDT) methods, namely listwise deletion, mean imputation, restrictive EM, inclusive EM, restrictive MI and inclusive MI for generating unbiased estimates in a 3-level HLM model. The imputed datasets were analyzed under the same HLM model. Parameter estimates from those imputed datasets were compared to those obtained from the original complete datasets. The comparisons focused on the bias of parameter estimates of fixed effects and random effects (i.e., variance components). The missingness was MAR occurring on a level-2 predictor variable in a 3-level HLM model.

The 3-level HLM Model and Parameters

*Intercept only as outcome model.* In the “intercept-only as outcome model” (or “random intercept” model), only the intercept at a lower level is predicted by higher level variables, that is, level-1 intercept is regressed on level-2 variables and level-2 intercept is regressed on level-3 variables. All regression coefficients except the intercepts is constrained to be constant within classes (level-2) and schools (level-3). There is one predictor variable at level-1—\( a_i \); two predictor variables at level-2—\( X_1 \) and \( X_2 \); and one predictor variable at level-3—\( W_j \). So the models at each level and the combined model are:

**Level-1:** \[ Y_{ijk} = \pi_{0jk} + \pi_{1jk} a_{ijk} + e_{ijk} \]  

\[ 3.1 \]
Where:

\( Y_{ijk} \) is the outcome/dependent variable (e.g., math achievement score) for student \( i \) in class \( j \) in school \( k \).

\( \pi_{0jk} \) is the mean score of class \( j \) in school \( k \).

\( \pi_{1jk} \) is the corresponding level-1 coefficient that indicates the direction and strength of association between each student characteristic (i.e., \( a_i \)) and the outcome variable \( Y_{ijk} \).

\( e_{ijk} \) is the random "student effect." It is the residual at level-1 which is normally distributed with a mean of 0 and variance of \( \sigma^2 \).

Level-2

\[
\pi_{0jk} = \beta_{00k} + \beta_{01k} X_{1jk} + \beta_{02k} X_{2jk} + r_{0jk}
\]

[3.2]

Where:

\( \beta_{00k} \) is the average outcome for school \( k \).

\( \beta_{01k} - \beta_{02k} \) are the corresponding level-2 coefficients that indicate the direction and strength of association between each class characteristic (i.e., \( X_1, X_2 \)) and the average score in class \( jk \) (\( \pi_{0jk} \)), again they will be constrained as constants.

\( \beta_{10k} \) is the cross-level effect coefficients between level-1 and level-2. Here, it is constrained to be constant which means the class characteristics do not have any differential effect on how \( a_i \) associate with the outcome (i.e., \( \pi_{1jk} \)).

\( r_{0jk} \) is the random "class effect" within school \( k \) which is normally distributed with a mean of 0 and variance \( \tau_\pi \).
Level-3: $\beta_{00k} = \gamma_{000} + \gamma_{001} W_{1k} + u_{00k}$  \[3.3\]

$\beta_{01k} = \gamma_{010}$

$\beta_{02k} = \gamma_{020}$

$\beta_{10k} = \gamma_{100}$

Where:

$\gamma_{000}$ is the grand mean of all the schools.

$\gamma_{001}$ is the corresponding level-3 coefficient that represents the direction and strength of association between school characteristics (i.e., $W_i$) and the grand mean ($\gamma_{000}$).

$\gamma_{010} - \gamma_{020}$ are the cross-level effect coefficients between level-2 and level-3. They are also constrained to be constants which means the school characteristics $W_i$ does not have any differential effect on how $X_1$ and $X_2$ associate with the average class score (i.e., $\beta_{01k}$ and $\beta_{02k}$).

$\gamma_{100}$ is the cross-level effect coefficient between level-1 and level-3. It is also constrained to be constant which means the school characteristic $W_i$ does not have any differential effect on how $a_I$ associates with the outcome (i.e., $\pi_{1jk}$).

$u_{00k}$ is the random “school effect” among schools which is normally distributed with a mean of 0 and variance of $\tau_\theta$.

Combined model:

$Y_{ijk} = \gamma_{000} + \gamma_{001} W_{1k} + \gamma_{010} X_{1jk} + \gamma_{020} X_{2jk} + \gamma_{100} a_{1jk} + u_{00k} + r_{0jk} + e_{ijk}$  \[3.4\]

Therefore in this 3-level random intercept model, there are total of eight parameters. One intercept, four parameters of fixed effects and three parameters of
random effects, specifically, fixed effects parameters are \( \gamma_{100} \) for level-1 variable \( a_i \), \( \gamma_{010} \) and \( \gamma_{020} \) for level-2 variables \( X_i \) and \( X_2 \), \( \gamma_{001} \) for level-3 variable \( W_i \). random effects parameters \( (\sigma^2, \tau, \text{ and } \tau_\beta) \) are the variance components at student level \( (e_{jk}) \), at class level \( (r_{0,k}) \) and at school level \( (u_{00k}) \). Table 1 presents the eight parameters estimated in the HLM analysis.

Table 1.

<table>
<thead>
<tr>
<th>Parameters for the random intercept model</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Parameter ( \gamma )</td>
</tr>
<tr>
<td>Intercept</td>
<td>( \gamma_{000} ) Grand mean controlling for all the effects</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td></td>
</tr>
<tr>
<td>Student level</td>
<td></td>
</tr>
<tr>
<td>( a_i )</td>
<td>( \gamma_{100} ) Main effect of ( a_i ) within class within school</td>
</tr>
<tr>
<td>Class level</td>
<td></td>
</tr>
<tr>
<td>( X_i )</td>
<td>( \gamma_{010} ) Main effect of ( X_i ) among classes within school</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>( \gamma_{020} ) Main effect of ( X_2 ) among classes within school</td>
</tr>
<tr>
<td>School level</td>
<td></td>
</tr>
<tr>
<td>( W_i )</td>
<td>( \gamma_{001} ) Main effect of ( W_i ) among schools</td>
</tr>
<tr>
<td>Random Effects (Variance Components)</td>
<td></td>
</tr>
<tr>
<td>Student level ( (e_{jk}) )</td>
<td>( \sigma^2 ) Variance component at student level</td>
</tr>
<tr>
<td>Class level ( (r_{0,k}) )</td>
<td>( \tau ) Variance component at class level</td>
</tr>
<tr>
<td>School level ( (u_{00k}) )</td>
<td>( \tau_\beta ) Variance component at school level</td>
</tr>
</tbody>
</table>

Experimental Design

**Sample size.** As in most statistical analyses, the larger the sample size, the more precise the estimates are. Bryk and Raudenbush (1992) noted that in HLM analyses the more complex the model specified, the greater amount of data needed for estimation. In the case of the High School and Beyond data, where approximately 45 students were nested in each of the 160 schools, they indicated that “three random coefficients plus a
random intercept is about as rich a model as the data can sustain" (p. 258) Drawing on the general regression rules, Bryk and Raudenbush (1992) recommended that the regression rule of thumb of 10 observations per predictor can also be applied in hierarchical linear models.

For the current study, there were several considerations on sample size determination. First, since the study was designed to examine missing data methods at level-2, the level-2 sample size was a desirable condition to vary in the study. However, in a 3-level hierarchical model changes in the level-2 sample size will result in changes in the total sample size or group size at level-3 (level-2 total size = level-3 total size x level-2 group size). For example, with balanced design, a level-2 sample size of 100 could have seven different combinations at level-3: 2 (50), 4 (25), 5 (20), 10 (10), 20 (5), 25 (4) and 50 (2). As discussed by Kim (1999) Bassiri (1988), and Van der Leeden & Busing (1994) more power was gained by increasing the number of groups as opposed to the number of individuals per group. Thus, different total size (e.g., number of units at level-3) and different group size at a certain level (e.g., number of level-2 units nested in per level-3 unit) do have differential impacts on parameter estimation.

Second, the determination of level-2 and level-3 group sizes and total sizes was made to reflect the typical educational setting. Eighteen studies using hierarchical linear models were reviewed. All of these studies utilized hierarchical structures—either 3 levels with students within classrooms within schools, or 2 levels with students nested in classes.

For level-3 total size (i.e., number of schools), the 10-to-1 rule of thumb was considered. As discussed by Bryk and Raudenbush (1992), [in a 2-level HLM] for
predicting any single level-2 outcome (e.g., the level-1 intercept term), 10 observations were needed per predictor, while for multiple level-2 outcome (e.g., level-2 intercepts and slopes), they did not offer clear guidelines. Accordingly, in a 3-level hierarchical model, for predicting a single level-3 outcome (i.e., level-2 intercept) with one predictor variable, the minimum number of level-3 units needed is 10, and for a more stable model the level-3 total size was determined at 20 (e.g., 20 schools).

With respect of level-2 group size (i.e., number of classes per school), there was no clear standard in the literature about what is the group size needed. Among the HLM studies with level-2 being classes or teachers nested in schools (Bosker et al., 1990; Kyriakides et al., 2000; Marks, 2000; Finn et al., 2001; Opdenakker, 2001; Lamb, 2002; Xue, 2002; Archibald, 2005; Goddard, 2007), the average level-2 group sizes ranged from 2 to 9 classes per school, with the median around 3 to 4, while in the multilevel studies with schools nested in countries (Franco et al., 2005), the level-2 group size could go up to 206. As the level-3 total size is set at 20, to have a total size for level-2 units as small as possible but also stable, the group size of 3 classes per school was chosen, which set the smallest level-2 sample size condition at 60 (i.e., 60 classes) and this sample size fairly reflected the small level-2 size found in the 3-level HLM literature which is 47 (Goddard, 2007) and 56 (Kyriakides et al., 2000).

As for medium and large level-2 total size, among the 3-level HLM studies reviewed, the level-2 total sizes ranged from 47 (Goddard, 2007) to 2690 (Xue, 2002) with median between 163 and 169 and top third beginning from 329 (Finn et al. 2001). Thus, the level-2 total sizes were set as 180 and 360 to reflect the medium and large level-2 size conditions.
Next, the level-1 group size (i.e., number of students per class) is considered. Regarding the number of students nested within classrooms, Muthén (1994) suggested a typical class size being 20; Raudenbush and Willms (1991) noted that the average class size ranges from 20 to 35. In a class size effect study with a 3-level design (Nye, Hedges, & Konstantopoulos, 2000), regular class size was defined as 18-21 students per class. In the HLM studies where students were nested in classrooms or in teachers at a certain level, the average classroom sizes (i.e., group size at class or teacher level) ranged from 5 (Xue, 2002) to 27 (Franco et al., 2005) with a median around 18 to 19 students per class. Given the above considerations on class size, the classroom size was set at 20.

As mentioned earlier, a change in level-2 total size would result in a change in either level-3 total size or level-2 group size, and they may have different impacts on parameter estimation (Kim, 1999; Bassiri, 1988; Van der Leeden & Busing, 1994). Therefore, there were two ways to design—A) to fix the level-2 group size with varying level-3 total size, or B) to fix the level-3 total size with varying level-2 group size. Table 2 presents the two ways of sample size design. With the small size condition for design A and B being the same, the sample size conditions in this study were five—small (20*3), medium_A (60*3), medium_B (20*9), large_A (120*3) and large_B (20*18).

In both design A and B, representing small, medium and large sample sizes there were 60, 180 and 360 observations at level-2, as well as 1200, 3600 and 7200 observations at level-1. Regarding level-3 observations there were two different designs each for the small, medium and large study conditions. Design A and B differed in the way of how level-2 observations were nested in level-3 units. Specifically, in design A the number of level-2 observations nested in a level-3 unit was fixed at 3, and thus the
The total number of level-3 units were 20, 60, and 120 for small, medium and large design.

Design B fixed the total number of level-3 units at 20 and thus varied the number of level-2 observations nested in a level-3 unit at 3, 9, and 18 for small, medium and large study designs.

Table 2.

**Design of sample size**

<table>
<thead>
<tr>
<th>Level-2 sample size</th>
<th>A. Level-2 group size fixed, but level-3 total size varies</th>
<th>B. Level-3 total size fixed, but level-2 group size varies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total size</td>
<td>Group size</td>
</tr>
<tr>
<td>Large</td>
<td>Level-1</td>
<td>7200 students</td>
</tr>
<tr>
<td></td>
<td>Level-2</td>
<td>360 classes</td>
</tr>
<tr>
<td></td>
<td>Level-3</td>
<td>120 schools</td>
</tr>
<tr>
<td>Medium</td>
<td>Level-1</td>
<td>3600 students</td>
</tr>
<tr>
<td></td>
<td>Level-2</td>
<td>180 classes</td>
</tr>
<tr>
<td></td>
<td>Level-3</td>
<td>60 schools</td>
</tr>
<tr>
<td>Small</td>
<td>Level-1</td>
<td>1200 students</td>
</tr>
<tr>
<td></td>
<td>Level-2</td>
<td>60 classes</td>
</tr>
<tr>
<td></td>
<td>Level-3</td>
<td>20 schools</td>
</tr>
</tbody>
</table>

*Proportion of missingness.* In the missing data studies, the percentages of missingness ranging from 5% (Prosser, 1991) to 50% (Collins, Schafer & Kam, 2001; Dalton & Schulz, 2006). The proportion of missing data can affect the efficiency of the MDT’s. Previous MDT studies suggested that in linear regression analyses, if the missing data are less than 10% (Donner, 1992) or even 15% (Raymond & Roberts, 1987), mean substitution and regression estimates may be effective even with MAR. In this study, the percentages of missingness were set at 10% and 40%, representing the low and high levels of missingness found in the literature. The missingness in the study was generated as MAR.
**Missing Data Treatment (MDT).** This study compared six methods of missing data treatment (MDT). They were listwise deletion (LD), mean substitution (MS), restrictive EM (REM), inclusive EM (IEM), restrictive MI (RMI) and inclusive MI (IMI). The restrictive approach used only level-2 variables to impute level-2 missing data, while the inclusive approach used both level-1 and level-2 variables in the imputation process. The procedures of the six MDTs detailed under data generation section.

**Summary of Experimental Design**

Fixed and random parameter estimates produced by six MDTs were compared to those obtained from the complete datasets. Table 3 presents the study design. Two factors—level-2 sample size and proportion of missing data were experimentally manipulated in the study. The level-2 sample size had 5 conditions—small (20 x 3), medium_A (60 x 3), medium_B (20 x 9), large_A (120 x 3) and large_B (20 x 18), and the proportion of missing data had 2 conditions—10% and 40%.

Table 3.

*Design of the study*

<table>
<thead>
<tr>
<th>Level-2 sample size</th>
<th>% of missing</th>
<th>Missing data treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
</tr>
<tr>
<td>Small (N=60)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Medium_A (N=60x3)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Medium_B (N=20x9)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Large_A (N=120x3)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Large_B (N=20x18)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

*1000 replications in each cell.*
Sample size and proportion of missingness served as between factors and the missing data treatment (i.e., six MDTs) was the within factor in this 2 between, 1 within experimental design. The conditions were fully crossed and 1000 replications of each dataset under every condition were generated.

Data Generation

Data were generated to simulate typical psychological and educational data found in the 3-level HLM literature. In consideration of the parameter specification, six studies were examined. All of those studies used “intercept-only” (or “random coefficient”) HLM models, had a hierarchical structure of students nested within classes within schools, and clearly reported variance components or ICCs (Scheerens, 1989; Kyriakides, 2000; Mark, 2000; Opdenakker, 2001; Lamb, 2002; Archibald, 2006) (see Table 4).

Table 4.

<table>
<thead>
<tr>
<th>Article</th>
<th>Variance Proportion (ICC)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Students ($\sigma^2$)</td>
<td>Class ($\tau_\pi$)</td>
<td>School ($\tau_p$)</td>
</tr>
<tr>
<td>Archibald (2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>.712</td>
<td>.059</td>
<td>.229</td>
</tr>
<tr>
<td>Math</td>
<td>.617</td>
<td>.108</td>
<td>.275</td>
</tr>
<tr>
<td>Kyriakides (2000)</td>
<td>.766</td>
<td>.152</td>
<td>.082</td>
</tr>
<tr>
<td>Math</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mark (2000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement (Elem)</td>
<td>.875</td>
<td>.062</td>
<td>.063</td>
</tr>
<tr>
<td>Engagement (Mid)</td>
<td>.906</td>
<td>.050</td>
<td>.044</td>
</tr>
<tr>
<td>Engagement (HS)</td>
<td>.958</td>
<td>.042</td>
<td>.000</td>
</tr>
<tr>
<td>Opdenakker (2001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>.666</td>
<td>.170</td>
<td>.164</td>
</tr>
<tr>
<td>Math</td>
<td>.720</td>
<td>.189</td>
<td>.091</td>
</tr>
<tr>
<td>Math</td>
<td>.754</td>
<td>.204</td>
<td>.042</td>
</tr>
<tr>
<td>Lamb (2002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math (USA)</td>
<td>.687</td>
<td>.230</td>
<td>.083</td>
</tr>
<tr>
<td>Bosker (1990)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>.740</td>
<td>.069</td>
<td>.019</td>
</tr>
<tr>
<td>Average</td>
<td>.778</td>
<td>.122</td>
<td>.100</td>
</tr>
</tbody>
</table>
In those six studies, the variance component at students level \((\sigma^2)\) ranges from .62 to .96 with a mean of .778; the intra-class correlation (ICC) at class level \((\rho_{\text{class}})\) or variance component \(\tau_\pi\) in conditional model ranges from .05 ~ .23 with a mean of .122; and the ICC at school level \((\rho_{\text{school}})\) or the variance component \(\tau_\beta\) in conditional model ranges from .04 ~ .27 with a mean of .1. Thus, the variance components of student, class and school level \((\sigma^2, \tau_\pi\text{ and } \tau_\beta)\) were set to be .778, .122 and .10. The square root of those variance component values— .882, .349 and .316 were the values for \(e_{ijk}\), \(r_{0jk}\) and \(u_{00k}\) respectively.

SAS (SAS Institute, 2004) RANNOR function was used to generate level-1 and level-3 predictor variables \((a_I\text{ and } W_I)\) from a normal distribution with a mean of 0 and standard deviation of 1. Level-2 predictor variables \((X_I\text{ and } X_2)\) were generated as bivariate normal (SAS Institute, 2007) and they were specified to have a correlation of approximately .3, reflecting a moderate correlation found in relevant Monte Carlo studies (Zhang, 2005). Data were generated in the nested HLM structure stipulated in Equations 3.1-3.3. Specifically, the school level predictor \((W_I)\), school level random effect \((u_{00k})\) and grand intercept \((\gamma_{000})\) were generated first, then the school level outcome \((\beta_{00k})\) was generated by Equation 3.3. Then the school level outcome \((\beta_{00k})\) was used to generate class level outcome \((\pi_{0jk})\) by adding class-level random effect \((r_{0jk})\) and predictor variables \((X_j\text{ and } X_2)\) (Equation 3.2). Then, class level outcome \((\pi_{0jk})\) was used to generate the student-level outcome variable \((Y_{ijk})\) (Equation 3.1). For the parameterization of fixed effects \((\gamma_{010}, \gamma_{020}, \gamma_{100}, \gamma_{200}, \gamma_{001})\), .3 was arbitrarily picked.
After data were generated, descriptive statistics and HLM parameters in the data were examined. Descriptive statistics as well as univariate normality results of all the variables generated under sample size Large_A condition (360 classes nested in 120 schools) are shown in Table 5. Table 6 presents the HLM parameters specified in data generation and the actual HLM parameters under the same sample size.

### Table 5.

**Descriptive statistics of the generated data**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Kolmogorov-Smirnov's $D$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>7200</td>
<td>.505</td>
<td>1.202</td>
<td>-.020</td>
<td>.030</td>
<td>.01</td>
<td>.1500</td>
</tr>
<tr>
<td>$a_1$</td>
<td>7200</td>
<td>-.014</td>
<td>1.002</td>
<td>.004</td>
<td>.047</td>
<td>.01</td>
<td>.1500</td>
</tr>
<tr>
<td>$X_1$</td>
<td>360</td>
<td>-.073</td>
<td>1.014</td>
<td>.077</td>
<td>.139</td>
<td>.02</td>
<td>.1500</td>
</tr>
<tr>
<td>$X_2$</td>
<td>360</td>
<td>-.039</td>
<td>.972</td>
<td>-.077</td>
<td>.042</td>
<td>.04</td>
<td>.1500</td>
</tr>
<tr>
<td>$W_1$</td>
<td>120</td>
<td>.030</td>
<td>.946</td>
<td>-.026</td>
<td>.448</td>
<td>.06</td>
<td>.1500</td>
</tr>
</tbody>
</table>

### Table 6.

**Parameters specified and actual parameter estimates generated in the data**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Specified</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\gamma_{000}$</td>
<td>.5</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Student level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>$\gamma_{10}$</td>
<td>.3</td>
</tr>
<tr>
<td><strong>Class level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_1$</td>
<td>$\gamma_{010}$</td>
<td>.3</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$\gamma_{020}$</td>
<td>.3</td>
</tr>
<tr>
<td><strong>School level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$W_1$</td>
<td>$\gamma_{001}$</td>
<td>.3</td>
</tr>
<tr>
<td>Random Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student level ($\epsilon_{0k}$)</td>
<td>$\sigma^2$</td>
<td>.778</td>
</tr>
<tr>
<td>Class level ($\nu_{0jk}$)</td>
<td>$\tau_{\pi}$</td>
<td>.123</td>
</tr>
<tr>
<td>School level ($\mu_{00k}$)</td>
<td>$\tau_{\beta}$</td>
<td>.100</td>
</tr>
</tbody>
</table>
**MAR procedure.** A variable, \( za \), was created at the class level (level-2) which had a .3 correlation with \( X_j \). Then cases were divided into two groups—the upper 50 percent and the lower 50 percent based on the values of \( za \). Using SAS SURVEYSELEC procedure, a sampling selection of probability proportional to size (PPS) without replacement was then conducted in the lower 50% group of \( za \). Once the 10% and 40% of the total \( za \) size was specified and the values on \( za \) were selected, the corresponding \( X_j \) data points were removed. In this case, the missingness procedure created MAR missingness on \( X_j \) that the missingness was unrelated to \( X_j \) itself but dependent on variable \( za \). Similar MAR procedures were used in other missing data studies (Collins et. al., 2001; Dalton & Schulz, 2006; Zhang, 2005).

**Missing Data Treatment (MDT).** The six missing data treatment methods were applied on level-2 variable \( X_j \), which had missing data points.

- **MDT1.** Listwise deletion (LD) — all cases were removed if values on \( X_i \) was missing.
- **MDT2.** Mean substitution (MS) — the mean of the remaining \( X_j \) values was computed to substitute all the missing values on \( X_j \).
- **MDT3.** Restrictive EM (REM) — SAS MI procedure with EM option was performed, using only level-2 variable \( X_2 \) to impute missing data on variable \( X_j \).
- **MDT4.** Inclusive EM (IEM) — first, level-1 variables \( Y \) and \( a_l \) were aggregated at level-2, and then SAS MI procedure with EM option was performed, using the aggregated level-1 variable \( Y \) and \( a_l \) as well as the level-2 variable \( X_2 \) to impute missing data on variable \( X_j \).
• MDT5. Restrictive MI (RMI) — SAS MI procedure with MCMC option was performed using only level-2 variable $X_2$ to impute missing data on variable $X_l$. Between-imputation iteration of 200 and imputation of 5 datasets were specified. Documentation within the SAS MI procedure suggests the relative efficiency of five imputed datasets exceeds 98% for the 10% missing category 90% for 50% missingness which would suggest that the 40% missingness category used in this study would exceed this amount (SAS Institute, 2004).

• MDT6. Inclusive MI (IMI) — on the aggregated dataset from MDT 4, SAS MI procedure with MCMC option was performed, using the aggregated level-1 variable $Y$ and $a_l$, as well as the level-2 variable $X_2$ to impute missing data on variable $X_l$. Between-imputation iteration of 200 and imputation of 5 datasets were specified.

Research Questions

The study addressed the following research questions:

1. When compared with complete data, do the parameter estimates of the fixed effects calculated from six MDTs (i.e., listwise deletion, mean substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield statistically significant difference in the amount of bias under different experimental conditions (sample size and proportion of missing data)?

2. When compared with complete data, do the parameter estimates of the random effects calculated from six MDTs (i.e., listwise deletion, mean substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield statistically
significant difference in the amount of bias under different experimental conditions (sample size and proportion of missing data)?

3. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of fixed effects that are accurate and precise when compared with parameter estimates from complete data?

4. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of random effects that are accurate and precise when compared to parameter estimates from complete data?

Data Analysis

This experiment study represented a 5 x 2 x 6 design, with two between-subject factors (sample size and percentage of missing) and one within-subject factor (MDT).

Analysis of Parameter Estimate Bias (PEB)

Dependent variables. Following Enders (2001), parameter estimate bias (PEB) was calculated. The raw bias in a parameter estimate was expressed relative to the true parameter calculated as the value of the difference between the parameter estimate in each experimental condition and the true parameter in the corresponding complete dataset divided by the true parameter in the complete dataset. In other Monte Carlo studies, this parameter estimate bias, employing the same calculation, was defined as "Relative deviate" (Posser, 1990; Darandari, 2004). The calculation formula was:

$$PEB = \frac{Estimate_{imputed} - Parameter_{complete}}{Parameter_{complete}}$$  [3.5]
With 1000 replications in each cell, the above formulas produced 1000 parameter estimate bias (PEB) values for each parameter in each design cell (see Table 3). With 8 parameters in the HLM model (five fixed effects and three random effects), there were total of 8 dependent variables in the data analysis—PEB indices for intercept ($\gamma_{000PEB}$), fixed effects ($\gamma_{010PEB}, \gamma_{001PEB}, \gamma_{020PEB}$ and $\gamma_{001PEB}$) and random effects ($\sigma^2_{PEB}, \tau_{\beta PEB}$ and $\tau_{\beta PEB}$).

**Analyses.** The primary analyses for research question 1 & 2 were eight separate 5 x 2 x 6 ANOVAs with two between-subject factors (sample size and percentage of missing) and one within-subject factor (MDT). Each of the eight PEB indices was the dependent variable in ANOVA analysis separately.

**Between effects in the ANOVA analysis.**
- Sample size—this main effect of sample size would indicate that if the biases in the particular parameter estimate (e.g., $\gamma_{001PEB}$) would differ under different level-2 sample size conditions regardless of the percentage of missing and MDT.
- Percentage of missing—this main effect of percentage of missing would indicate that if the biases in the particular parameter estimate would differ with different percentage of missing regardless of sample size and MDT.
- Sample size x percentage of missing—this interaction effect of sample size and percentage of missing would indicated that regardless of MDT, if the bias in the particular parameter estimate generated under different missing percentages (i.e., 10% and 40%) differ under different sample size conditions.

**Within effects in the ANOVA analysis.**
• MDT—this main effect of MDT would indicate whether the biases in the particular parameter estimate generated by the six MDTs are different while controlling for sample size and percentage of missing.

• MDT x sample size—this interaction effect of MDT and sample size would indicate if different MDTs would generate different bias in parameter estimates with different sample size regardless of percentage of missing.

• MDT x percentage of missing—this interaction effect of MDT and percentage of missing would indicate if different MDTs would generate different bias in parameter estimates with different levels of missing regardless of sample size.

• MDT x sample size x percentage of Missing—this three-way interaction would indicate that whether different MDTs would generate different bias in parameter estimates with different sample sizes and with different levels of missing.

*Generalized partial omega-squared.* Generalized partial omega-squared ($\omega^2$) were calculated (see Appendix B) according to Olejnik and Algina (2003). Omega-squared is an estimate of the variance in the dependent variable accounted for by the independent variable effects estimated for the population. The generalized partial $\omega^2$s, as coefficients of effect size, are interpreted as the proportion of variance in the dependent variable that can be accounted for by the independent variable holding other effects constant. It indicates the unique contribution of each independent variable in explaining the variation in the dependent variable. Thus, a large effect size value indicates that a factor or the interaction between the factors uniquely generated large bias in the estimates and had a large impact on the parameter estimates.
According to Olejnik and Algina (2003), the generalized partial $\omega^2$s are comparable across different research designs and different outcome measures, adjusted for differences in the number of manipulated factors (Olejnik & Algina, 2003), and they provide indices that are consistent with Cohen's (1988) guidelines for defining the magnitude of the effect size, that is, values of .01, .06 and .14 indicate small, medium and large effect sizes (Cohen, 1988, p. 283-287).

Post hoc analysis. Six 5 x 2 ANOVAs were conducted following each 5 x 2 x 6 ANOVA. The 5 x 2 ANOVAs was intended to examine the impact of sample size and proportion of missingness with each MDT. Effect size values were computed to indicate the influence of each of the between factors on the PEB generated by the MDTs.

Analysis of Accuracy and Precision of the Estimates

For research question 3 & 4, accuracy and precision of the estimates generated by each MDT were investigated through root-mean-square error (RMSE). RMSE was a measure of the overall accuracy and precision of the parameter estimates and was calculated as the square root of the average squared difference between the estimate and the true parameter (Collins et. al, 2001) in each design cell (see Table 3). In several Monte Carlo studies, this measure was employed to indicate the accuracy of the estimates (Rayman & Roberts, 1987; Roth & Switzer, 1995; Enders, 2001; Collins et. al, 2001). It was calculated as:

\[
RMSE = \sqrt{\frac{\sum (Estimate_{imputed} - Parameter_{complete})^2}{N}}
\]  \[3.6\]

($N$=number of replications in each cell)

It was desirable that an estimate was accurate and had minimum variance (precise).

Accuracy was defined by the amount of deviation of estimates from the true parameter.
Precision was defined by the amount of dispersion of estimates around the true parameter. Enders (2001) pointed out that “mean-square-error (MSE) measures the sum of a parameter's variance and squared bias” (p. 357). Thus, if parameter estimates are unbiased, MSE quantifies the sampling variance of an estimate which could be regarded as the precision aspect of the estimation. If the estimates are biased, MSE serves to quantify the overall bias of an estimate, combining accuracy and precision (i.e., variance) into a single numeric term. The square root of the mean-square error is more interpretable than in squared terms because the RMSE is scaled on the same units as the data, thus it is an index of expected value of the “error” in the estimates. Collins, Schafer and Kam (2001) suggest the RMSE measure “combines the concepts of bias and efficiency, because it can be shown that the mean-square error of an estimate is equal to its squared bias plus its variance” (p. 340). RMSE is presented for each of the eight parameter estimates.
CHAPTER IV

RESULTS

The present study was intended to compare the performance of six different missing data treatment (MDT) methods—listwise deletion (LD), mean substitution (MS), restrictive EM (REM), inclusive EM (IEM), restrictive MI (RMI) and inclusive MI (IMI) for generating unbiased parameter estimates with MAR missingness in a 3-level HLM model. Each parameter in the HLM model was evaluated using parameter estimate bias (PEB) and root-mean-square error (RMSE) measures. PEB, calculated as the proportion of the raw bias of the estimates to the true parameter, examined the relative bias of parameter estimates generated in each replication as compared to the true parameter; and RMSE, computed as the square root of the average squared difference between the estimates and the true parameter, examined the overall accuracy and precision of the parameter estimates in every cell. Coverage, the percentage of times that the true parameter value was included in 95% confidence interval of the estimates, and width, the average length of 95% confidence interval of the estimates, were also evaluated.

Parameter Estimate Bias (PEB)

With two between-subject factor (sample size and proportion of missingness) and one within-subject factor (MDT), eight balanced 5 x 2 x 6 ANOVAs were conducted separately to investigate five parameter estimate bias (PEB) of fixed effects ($\gamma$'s) and three parameter estimate bias (PEB) of random effects ($\sigma^2$, $\tau$, and $\tau_\beta$) in the HLM model. With 1000 replication, there were 1000 PEB values in each design cell. Following each 5 x 2 x 6 ANOVA, six 5 x 2 ANOVAs were conducted to examine the impact of sample size and proportion of missingness in every MDT. Generalized partial omega-
squared values (Olejnik and Algina, 2003) were calculated for each main and interaction
effects as effect size coefficients. A large effect size value indicates that a factor or the
interaction effect uniquely produced large bias in the parameter estimates.

The 5 x 2 x 6 ANOVA results for PEB of fixed and random effects are presented
in Table 7, and the 5 x 2 ANOVA results for each MDT are presented in Table 8. The
mean PEB values of fixed and random effects are reported in Table 9 and Table 10.

PEB of Fixed Effects

Research Question 1: When compared with complete data, do the parameter
estimates of the fixed effects calculated from six MDTs (i.e., listwise deletion, mean
substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield
statistically significant difference in the amount of bias under different experimental
conditions (sample size and proportion of missing data)?

PEB of intercept ( $\gamma_{000}$ ). In the 5 x 2 x 6 ANOVA with PEB of intercept as
dependent variable, all the between and within-subject effects (main and interaction
effects) were statistically significant (all $p's<.001$, Table 7), which indicated that the
factors and interactions had significant effects in producing biases in the estimates. Based
on the omega-squared values which attempted to quantify the effects, the effect size of
sample size and proportion of missingness and their interaction were large, and the effect
size of MDT and the interaction between MDT and proportion of missingness were also
large (all $\omega^2 > .14$).

As indicated by the effect size values (Table 7), MDT was the largest source of
variation in PEB of the intercept that 58% of the variance that was not accounted by other
effects could be accounted for by MDT alone (not including its interactions with other
Table 7.

ANOVA (5x2x6) results for parameter estimate bias (PEB)

<table>
<thead>
<tr>
<th>Source</th>
<th>Intercept</th>
<th>Level-1 variable</th>
<th>Fixed Effects</th>
<th>Level-2 variable, with missing data</th>
<th>Level-2 variable, without missing data</th>
<th>Level-3 variable, with missing data</th>
<th>Level-3 variable, without missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_{000}$</td>
<td>$Y_{100}$</td>
<td>$Y_{010}$</td>
<td>$Y_{001}$</td>
<td>$Y_{011}$</td>
<td>$Y_{010}$</td>
<td>$Y_{001}$</td>
</tr>
<tr>
<td></td>
<td>$F^*$</td>
<td>$\omega^*$</td>
<td>$F^*$</td>
<td>$\omega^*$</td>
<td>$F^*$</td>
<td>$\omega^*$</td>
<td>$F^*$</td>
</tr>
<tr>
<td>Between subject effects</td>
<td>N</td>
<td>4</td>
<td>2492</td>
<td>.390</td>
<td>3603</td>
<td>.202</td>
<td>7495</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>10659</td>
<td>.367</td>
<td>41</td>
<td>.001</td>
<td>15720</td>
<td>.538</td>
</tr>
<tr>
<td>NxM</td>
<td>4</td>
<td>1643</td>
<td>.263</td>
<td>1843</td>
<td>.115</td>
<td>4312</td>
<td>.561</td>
</tr>
<tr>
<td>Error</td>
<td>9990</td>
<td>(.001)</td>
<td>(&lt;.001)</td>
<td>(.019)</td>
<td>(.007)</td>
<td>(.003)</td>
<td>(&lt;.001)</td>
</tr>
<tr>
<td>Within subject effects</td>
<td>MDT</td>
<td>5</td>
<td>30486</td>
<td>.581</td>
<td>334</td>
<td>.027</td>
<td>116014</td>
</tr>
<tr>
<td>MDTxN</td>
<td>20</td>
<td>834</td>
<td>.132</td>
<td>3149</td>
<td>.057</td>
<td>3093</td>
<td>.243</td>
</tr>
<tr>
<td>MDTxM</td>
<td>5</td>
<td>15514</td>
<td>.414</td>
<td>164</td>
<td>.013</td>
<td>39960</td>
<td>.509</td>
</tr>
<tr>
<td>MDTxNxM</td>
<td>20</td>
<td>355</td>
<td>.061</td>
<td>1674</td>
<td>.355</td>
<td>1912</td>
<td>.166</td>
</tr>
<tr>
<td>Error</td>
<td>49950</td>
<td>(&lt;.001)</td>
<td>(&lt;.001)</td>
<td>(.001)</td>
<td>(&lt;.001)</td>
<td>(&lt;.001)</td>
<td>(&lt;.001)</td>
</tr>
</tbody>
</table>

Note: N—Sample size
M—Proportion of missingness
MDT—Missing data treatment

Values in parentheses represent mean square errors (MSE).

*For all F-tests, Huyn-Feldt adjusted $p<.001$.

** Generalized partial $\omega^2$ were calculated based on Olejnik & Algina (2003, p. 444-445, see Appendix B).

**Bolded** $\omega^2$ values indicate large effect size, $\omega^2>.14$. 
Table 8.

ANOVA (5x2) results for parameter estimate bias (PEB) by MDT

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Level-1 variable</th>
<th>Level-2 variable</th>
<th>Level-2 variable</th>
<th>Level-3 variable</th>
<th>Level-1 random effect</th>
<th>Level-2 random effect</th>
<th>Level-3 random effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{00}$</td>
<td>$\gamma_{01}$</td>
<td>$\gamma_{10}$</td>
<td>$\gamma_{11}$</td>
<td>$\gamma_{00} \times X_1$ (with missing data)</td>
<td>$\gamma_{00} \times W_1$</td>
<td>$\gamma_{00}$</td>
<td>$\tau_\pi$</td>
</tr>
<tr>
<td>F*</td>
<td>$\omega^2$</td>
<td>F*</td>
<td>$\omega^2$</td>
<td>F*</td>
<td>$\omega^2$</td>
<td>F*</td>
<td>$\omega^2$</td>
<td>F*</td>
</tr>
<tr>
<td>LD N (df=4)</td>
<td>92</td>
<td>.035</td>
<td>3251</td>
<td>.565</td>
<td>1367</td>
<td>.353</td>
<td>2977</td>
<td>.543</td>
</tr>
<tr>
<td>LD M (df=1)</td>
<td>4963</td>
<td>.332</td>
<td>138</td>
<td>.014</td>
<td>3897</td>
<td>.280</td>
<td>1023</td>
<td>.093</td>
</tr>
<tr>
<td>LD NxM(df=4)</td>
<td>59</td>
<td>.023</td>
<td>1717</td>
<td>.407</td>
<td>918</td>
<td>.268</td>
<td>1458</td>
<td>.368</td>
</tr>
<tr>
<td>MS N</td>
<td>2473</td>
<td>.497</td>
<td>835</td>
<td>.250</td>
<td>10582</td>
<td>.809</td>
<td>19247</td>
<td>.885</td>
</tr>
<tr>
<td>MS M</td>
<td>22237</td>
<td>.690</td>
<td>214</td>
<td>.021</td>
<td>15346</td>
<td>.605</td>
<td>49220</td>
<td>.831</td>
</tr>
<tr>
<td>MS NxM</td>
<td>2087</td>
<td>.455</td>
<td>221</td>
<td>.081</td>
<td>5551</td>
<td>.689</td>
<td>7697</td>
<td>.755</td>
</tr>
<tr>
<td>REM N</td>
<td>2582</td>
<td>.508</td>
<td>766</td>
<td>.234</td>
<td>6514</td>
<td>.723</td>
<td>9041</td>
<td>.783</td>
</tr>
<tr>
<td>REM M</td>
<td>24153</td>
<td>.707</td>
<td>371</td>
<td>.036</td>
<td>10502</td>
<td>.512</td>
<td>15481</td>
<td>.608</td>
</tr>
<tr>
<td>REM NxM</td>
<td>1549</td>
<td>.382</td>
<td>231</td>
<td>.084</td>
<td>4040</td>
<td>.618</td>
<td>5271</td>
<td>.678</td>
</tr>
<tr>
<td>IEM N</td>
<td>790</td>
<td>.240</td>
<td>1524</td>
<td>.378</td>
<td>6046</td>
<td>.797</td>
<td>4960</td>
<td>.665</td>
</tr>
<tr>
<td>IEM M</td>
<td>11865</td>
<td>.543</td>
<td>477</td>
<td>.045</td>
<td>1762</td>
<td>.150</td>
<td>2836</td>
<td>.221</td>
</tr>
<tr>
<td>IEM NxM</td>
<td>584</td>
<td>.189</td>
<td>816</td>
<td>.245</td>
<td>4311</td>
<td>.633</td>
<td>3520</td>
<td>.585</td>
</tr>
<tr>
<td>RMI N</td>
<td>7660</td>
<td>.754</td>
<td>4948</td>
<td>.663</td>
<td>5666</td>
<td>.694</td>
<td>23957</td>
<td>.906</td>
</tr>
<tr>
<td>RMI M</td>
<td>21096</td>
<td>.678</td>
<td>626</td>
<td>.059</td>
<td>154645</td>
<td>.939</td>
<td>78441</td>
<td>.887</td>
</tr>
<tr>
<td>RMI NxM</td>
<td>2782</td>
<td>.527</td>
<td>1408</td>
<td>.359</td>
<td>3077</td>
<td>.552</td>
<td>12556</td>
<td>.834</td>
</tr>
<tr>
<td>IMI N</td>
<td>5368</td>
<td>.682</td>
<td>647</td>
<td>.206</td>
<td>7123</td>
<td>.740</td>
<td>13435</td>
<td>.843</td>
</tr>
<tr>
<td>IMI M</td>
<td>8541</td>
<td>.461</td>
<td>582</td>
<td>.055</td>
<td>25942</td>
<td>.722</td>
<td>16047</td>
<td>.616</td>
</tr>
<tr>
<td>IMI NxM</td>
<td>2197</td>
<td>.468</td>
<td>380</td>
<td>.132</td>
<td>2864</td>
<td>.534</td>
<td>5719</td>
<td>.696</td>
</tr>
</tbody>
</table>

Note: N—Sample size. M—Proportion of missingness. All $d_{f_{error}}=9990$, all mean square errors (MSE) <.01.

* For all F-tests, $p<.001$. ** Generalized partial $\omega^2$ were calculated based on Olejnik & Algina (2003).

Bolded $\omega^2$ values indicate large effect size, $\omega^2>$.14.
Table 9.

Mean values of parameter estimate bias (PEB) for fixed effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample size</th>
<th>Missing 10%</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
<td>MS</td>
<td>REM</td>
<td>IEM</td>
<td>RMI</td>
</tr>
<tr>
<td>( \gamma_{00 \omega} )</td>
<td>1</td>
<td>0.005</td>
<td>0.015</td>
<td>0.007</td>
<td>0.008</td>
<td>0.012</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>2</td>
<td>0.010</td>
<td>0.012</td>
<td>0.011</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.004</td>
<td>0.020</td>
<td>0.019</td>
<td>0.012</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.007</td>
<td>0.011</td>
<td>0.016</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.008</td>
<td>0.016</td>
<td>0.017</td>
<td>0.007</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma_{10 \omega} )</td>
<td>1</td>
<td>0.007</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>(Level-1</td>
<td>2</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>variable ( a1 )</td>
<td>3</td>
<td>0.004</td>
<td>0.023</td>
<td>0.011</td>
<td>0.055</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.026</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma_{20 \omega} )</td>
<td>1</td>
<td>0.005</td>
<td>0.131</td>
<td>0.056</td>
<td>0.026</td>
<td>0.098</td>
</tr>
<tr>
<td>(Level-2</td>
<td>2</td>
<td>0.014</td>
<td>0.037</td>
<td>0.016</td>
<td>0.032</td>
<td>0.032</td>
</tr>
<tr>
<td>variable ( X1, \text{with missing} )</td>
<td>3</td>
<td>0.013</td>
<td>0.032</td>
<td>0.008</td>
<td>0.026</td>
<td>0.020</td>
</tr>
<tr>
<td>missing ( data )</td>
<td>4</td>
<td>0.020</td>
<td>0.038</td>
<td>0.015</td>
<td>0.042</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.018</td>
<td>0.030</td>
<td>0.015</td>
<td>0.036</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma_{01 \omega} )</td>
<td>1</td>
<td>0.010</td>
<td>0.027</td>
<td>0.016</td>
<td>0.025</td>
<td>0.032</td>
</tr>
<tr>
<td>(Level-3</td>
<td>2</td>
<td>0.097</td>
<td>0.012</td>
<td>0.010</td>
<td>0.004</td>
<td>0.044</td>
</tr>
<tr>
<td>variable ( W1 )</td>
<td>3</td>
<td>0.017</td>
<td>0.015</td>
<td>0.014</td>
<td>0.034</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.005</td>
<td>0.006</td>
<td>0.025</td>
<td>0.017</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.016</td>
<td>0.004</td>
<td>0.002</td>
<td>0.032</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Shaded numbers indicate the mean PEB value ≤ 1% deviant from the true parameter.

2a. Missing data treatment
   LD—listwise deletion
   MS—mean substitution
   REM—restrictive EM
   IEM—inclusive EM

2b. Sample size
   1—60 classes [20 schools (3 classes/school)]
   2—180 classes [60 schools (3 classes/school)]
   3—180 classes [20 schools (9 classes/school)]
   4—360 classes [120 schools (3 classes/school)]
   5—360 classes [20 schools (18 classes/school)]
Table 10.

Mean values of parameter estimate bias (PEB) for random effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Size</th>
<th>Missing 10%</th>
<th>Missing 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
<td>MS</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>1</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>(Level-1 variance component)</td>
<td>2</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>( \tau_\pi )</td>
<td>1</td>
<td>-0.180</td>
<td>0.593</td>
</tr>
<tr>
<td>(Level-2 variance component)</td>
<td>2</td>
<td>-0.001</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.027</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.016</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-0.008</td>
<td>0.074</td>
</tr>
<tr>
<td>( \tau_\beta )</td>
<td>1</td>
<td>0.053</td>
<td>-0.144</td>
</tr>
<tr>
<td>(Level-3 variance component)</td>
<td>2</td>
<td>0.028</td>
<td>-0.022</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.033</td>
<td>-0.033</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.032</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.004</td>
<td>0.017</td>
</tr>
</tbody>
</table>

*Shaded numbers indicate the mean PEB value ≤ 1% deviant from the true parameter.*

a. Missing data treatment

- LD — listwise deletion
- MS — mean substitution
- REM — restrictive EM
- IEM — inclusive EM
- RMI — restrictive MI
- IMI — inclusive MI

b. Sample size

- 1 — 60 classes [20 schools (3 classes/school)]
- 2 — 180 classes [60 schools (3 classes/school)]
- 3 — 180 classes [20 schools (9 classes/school)]
- 4 — 360 classes [120 schools (3 classes/school)]
- 5 — 360 classes [20 schools (18 classes/school)]
factors), $\omega^2_{\text{MDT}} = .581$; and the interaction between MDT and proportion of missingness was accounted for 41% of variance that was not explained by other effects. The large effect sizes of sample size and proportion of missingness and their interaction revealed that the estimation of the intercept was also impacted by these factors.

Six 5 x 2 ANOVAs were also conducted separately on each MDT to examine the impact of sample size and proportion of missingness on each MDT method. The effect size values (Table 8) showed that LD was the least impacted by the two between-factors. Proportion of missingness was the largest source of variation for LD, MS, REM and IEM, and sample size was the most influential factor for RMI and IMI.

The mean PEB values (Table 9) revealed that every MDT produced less bias in the intercept when there was 10% missingness compared with 40%. With 40% missingness, every MDT generated less bias under small sample size compared with other sample sizes.

PEB of $a_i$ effect (level-1 predictor) ($\gamma_{100}$). In the 5 x 2 x 6 ANOVA on PEB of level-1 fixed effect, all the between and within-subject effects (main and interaction effects) were statistically significant (all $p$'s<.001). However, the omega-squared values showed that only three effect sizes were large: the interaction between MDT and sample size, the three-way interaction between MDT, sample size and proportion of missingness, and the main effect of sample size ($\omega^2$s>.14, Table 7). In another words, any effect involving sample size could be accounted for a large percentage of variance in PEB of $\gamma_{100}$, and sample size greatly impacted the estimation of the level-1 fixed parameter. MDT and proportion of missingness exhibited small effect sizes on PEB of $\gamma_{100}$. 
In the six 5 x 2 ANOVAs on each MDT, the effect size values (Table 8) showed that REM was least impacted by the two between-factors (sample size and proportion of missingness). Since all MDTs except LD retained the original level-1 sample size, the effect of sample size appeared to be the largest source of variation, and proportion of missingness did not exhibit a large impact on the MDTs.

The reason for the small effect sizes of proportion of missingness is that all the MDT, except LD, retained the original level-1 sample size in the HLM model with both missingness conditions, thus not much bias was introduced into the estimation of level-1 fixed effect by this factor. LD, by conducting deletion of the cases with missing data, reduced the original sample size. For example, in the small sample size, LD reduced the level-1 total size from 1200 students to 1080 students. As shown in the mean PEB table (Table 9), every MDT, except LD, generated trivial or no bias (less than 1%) under both levels of missingness.

PEB of $X_1$ effect (level-2 predictor with missing data) ($\gamma_{010}$). All the main effects and interaction effects were statistically significant in the 5 x 2 x 6 ANOVA on PEB of $\gamma_{010}$ (all $p$’s<.001, Table 7). With missing data on this variable, it was expected that a large amount of bias would be generated in parameter estimation of this fixed effect. The effect size coefficients indicated that sample size, proportion of missingness and MDT, each factor alone, can account for a large percentage of variation in PEB of $\gamma_{010}$, $\omega^2_{\text{SIZE}}=.689$, $\omega^2_{\text{MISS}}=.538$ and $\omega^2_{\text{MDT}}=.751$. All the two-way and three-way interactions also had large effect sizes on producing biased estimates ($\omega^2_{\text{SIZE} \times \text{MISS}}=.561$, $\omega^2_{\text{MDT} \times \text{SIZE}}=.243$, $\omega^2_{\text{MDT} \times \text{MISS}}=.509$ and $\omega^2_{\text{MDT} \times \text{SIZE} \times \text{MISS}}=.166$).
In the six 5 x 2 ANOVAs on each MDT, sample size, proportion of missingness and their interaction showed statistical significance for every MDT (all p’s<.001, Table 8). The effect size values were generally huge ($\omega^2 = .150$–.939) for all the MDTs. In comparison, LD appeared to be least impacted by the two factors ($\omega^2 = .268$–.353). It is also noticeable that for LD, MS, REM, IEM and IMI, sample size exhibited a stronger impact than proportion of missingness, while proportion of missingness had a larger effect size than sample size for RMI.

The mean PEB values (Table 9) indicated that the mean biases produced by LD and REM were much smaller than those produced by other MDTs, and they generated small biases (less than 1% deviation from the true parameter) across sample sizes with 10% missingness. When there was 40% missing data, every MDT generated the largest bias under small sample size. In comparison with other MDTs, IEM, RMI and IMI produced huge bias under every condition (14%–72% deviation).

*PEB of $X_2$ effect (level-2 predictor without missing data) ($\gamma_{020}$).* The results of 5 x 2 x 6 ANOVA on PEB of $\gamma_{020}$ revealed that all the main effects and interaction effects had statistically significant effects in generating biases in the estimates (all p’s<.001, Table 7). Since the missing data occurred on a variable at the same level of $X_2$, the estimation of the $X_2$ effect appeared to be greatly impacted by the design factors. The omega-squared values indicated that sample size, proportion of missingness and MDT, each factor alone, strongly impacted on the estimate bias ($\omega^2_{\text{size}} = .753$, $\omega^2_{\text{miss}} = .601$, and $\omega^2_{\text{MDT}} = .518$).

In comparison of the effect sizes of the factors in PEB of $X_1$ (variable with missing data) effect and PEB of $X_2$ (variable without missing data) effect, MDT
expectedly had a greater impact on $\gamma_{010}$ ($X_1$ effect) estimation, $\omega^2_{MDT, X_1} = .751$, $\omega^2_{MDT, X_2} = .518$, which indicated that MDT had a greater effect on the variable with missing data. Sample size and proportion of missingness appeared to have greater impact on the estimation of $\gamma_{020}$ ($X_2$ effect), $\omega^2_{SIZE, X_1} = .698$, $\omega^2_{SIZE, X_2} = .753$, $\omega^2_{MISS, X_1} = .538$, $\omega^2_{MISS, X_2} = .601$, which indicated that these two factors had a greater impact on the parameter estimation of the variable without missing data but was at the same level where missing data occurred.

In the 5 x 2 ANOVAs, the two factors and their interaction exhibited large effect sizes for every MDT ($\omega^2 = .221$ to .906, Table 8) except that for LD, proportion of missingness could only explain 9% of variation in PEB of $\gamma_{020}$ uniquely. The mean PEB values (Table 9) revealed that every MDT except LD generated the largest mean biases when sample size was small with both 10% and 40% missing data, and LD generated the smallest bias under small sample size. Under other sample sizes, REM and IEM produced less biases in parameter estimates compared with other MDTs.

**PEB of $W_1$ effect (level-3 predictor) ($\gamma_{001}$).** The full factorial ANOVA results on PEB of $\gamma_{001}$ indicated that all the main and interaction effects were statistically significant (all $p'$s < .001, Table 7). The study factors and their interaction also exhibited large effect sizes ($\omega^2$s > .14) on PEB of $\gamma_{001}$. The interaction between MDT and sample size was accounted for the largest variation (41%) in PEB of $\gamma_{001}$, holding other effects constant, $\omega^2_{MDT \times SIZE} = .410$.

In the 5 x 2 ANOVAs, the effect sizes for sample size, proportion of missingness and their interaction were large ($\omega^2 = .153$ to .751, Table 8) for all the MDTs except that the
proportion of missingness had a medium effect on LD. For most MDTs, sample size was more influential than proportion of missingness in biased estimation. However, IEM was under greater impact of proportion of missingness. As shown by the mean PEB values (Table 9), with 10% missingness, REM and MS generated least biased estimates when level-2 total sample size was medium or large (level-2 total size ≥ 180). With 40% missingness, REM also generated smallest biases among all the MDTs under medium and large sample sizes. In another words, REM consistently produced the least biased \( \hat{\gamma}_{001} \) estimates when sample size was not small.

**PEB of Random Effects**

Research Question 2. When compared with complete data, do the parameter estimates of the random effects calculated from six MDTs (i.e., listwise deletion, mean substitution, restrictive EM, inclusive EM, restrictive MI and inclusive MI) yield statistically significant difference in the amount of bias under different experimental conditions (sample size and proportion of missing data)?

**PEB of level-1 random effect (\( \sigma^2 \)).** All the main and interaction effects were statistically significant in the 5 x 2 x 6 ANOVA with PEB of \( \sigma^2 \) as the dependent variable (all \( p \)'s<.001, Table 7). As indicated by the \( \omega^2 \) values, all the main effects of the factors did not show large effect size (\( \omega^2 s<.14 \)). The interaction between MDT and sample size was the largest source of variation, \( \omega^2_{MDT \times SIZE}=.387 \), and the three way interaction of sample size, proportion of missingness and MDT was another large source of variation, \( \omega^2_{MDT \times SIZE \times MISS}=.241 \).

In the six 5 x 2 ANOVAs, sample size and the interaction between sample size and proportion of missingness exhibited large effect sizes with LD (\( \omega^2 s>.14 \), Table 8).
The effect of sample size was also large for the two inclusive methods— IEM and IMI ($\omega^2$s>.14). Since the imputation methods (i.e., MS, REM, IEM, RMI, and IMI) retained the original level-1 sample size, proportion of missingness showed only a small impact on those methods. LD which reduced the original level-1 sample size, was expected to have a large impact of proportion of missingness, but in fact it was only moderately impacted by proportion of missingness ($\omega^2$s=.068).

Investigation of mean PEB values (Table 10) indicated that every MDT, except LD under 40% missingness, generated none or trivial bias (< 1% deviation from the true parameter) in the estimates of level-1 random effect. The examination of the mean PEB values for LD explained the moderate impact of missingness on LD. With small proportion of missingness (10%) the average of estimates generated by LD deviated .1% to .3% from the true parameter across sample sizes, and with large proportion of missingness (40%) LD generated estimates that deviated .4% to 2.1% under different sample sizes.

PEB of level-2 random effect ($\tau_\pi$). ANOVA results revealed that all the main and interaction effects had statistically significant effects on level-2 random effect ($\tau_\pi$) estimation (all $p$'s<.001, Table 7). With missing data at this level, all the factors and their interactions presented very large effect sizes ($\omega^2$s ranges from .465 to .899). Sample size exhibited the strongest impact on the biases of $\tau_\pi$ ($\omega^2_{\text{size}}$=.899)—it was accounted for 90% of the variation in PEB of $\tau_\pi$ that could not be accounted for by the other effects. Proportion of missingness ($\omega^2_{\text{miss}}$=.721) and MDT ($\omega^2_{\text{MDT}}$=.723) also exhibited large effect sizes on the bias of $\tau_\pi$. 
In the $5 \times 2$ ANOVAs (Table 8), LD appeared to be least impacted by the factors and their interactions in the estimation of $\tau_\alpha$. Compared with the proportion of missingness and interaction effects, sample size had the largest effect size for every MDT on the biases of $\tau_\alpha$. It was also noticeable that the proportion of missingness only moderately affected estimates produced by LD ($\omega^2_{MISS-LD}=.088$) and it tremendously impacted estimates generated by all the imputation MDTs ($\omega^2$, ranges from $.597$ to $.922$).

The mean PEB values of level-2 random effects (Table 10) were generally huge. Every MDT produced huge biases when the sample size was small, especially with 40% missing data. The large effect of sample size was mostly contributed by the biases generated under small sample size by every MDT. Under small sample size LD generated smallest bias compared with other MDTs. Under medium and large sample size conditions, IEM produced estimates that were least deviant from the true parameter than those produced by other MDTs.

**PEB of level-3 random effect ($\tau_\beta$).** For PEB of $\tau_\beta$ (level-3 random effect), all the main and interaction effects were statistically significant (all $p$'s < .001, Table 7). The $\omega^2$ values indicated that effect sizes of all the factors and their interactions were large ($\omega^2$s > .14), and sample size uniquely contributed the largest source of variance ($\omega^2_{SIZE}=.657$).

In the $5 \times 2$ ANOVAs, the results (Table 8) of effect sizes revealed a similar pattern of PEB of $\tau_\alpha$ (level-2 random effect). LD was least impacted by the factors and their interactions ($\omega^2=.054-.155$). For all other MDTs, sample size ($\omega^2 = .614 -.900$) had the largest effect size over proportion of missingness and their interactions.
Examination of mean PEB of $\tau_\beta$ (Table 10) indicated that every MDT tended to produce larger biases when sample size was small. With both low (10%) and high (40%) level of missingness LD generated less deviant estimates than other MDTs. RMI and IMI generated more biased estimates than other MDTs. MS and REM, displayed similar patterns with both levels of missingness. They underestimated $\tau_\beta$ with large biases when sample size was small (deviant about -11% under 10% missingness and -50% under 40% missingness) or medium (deviant about 2%~3% under 10% missingness and 10% under 40% missingness) but overestimate $\tau_\beta$ with better performance when sample size was large (deviant about 1.5% under 10% missingness and 7% under 40% missingness).

Summary of Parameter Estimate Bias

- The biases in parameter estimates generated by the six MDTs were greatly impacted by level-2 sample size and proportion of missingness. Different MDTs generated significantly different amounts of bias on level-2 and level-3 parameter estimate across the study conditions.

- With a high proportion of missingness, every MDT tended to generate more deviant parameter estimates and thus greater estimation bias.

- Under the small sample size condition, LD consistently generated the least biased estimates, while the imputation methods (MS, REM, IEM, RMI and IMI) consistently generated more biased estimates.

- The imputation MDTs generated less biased estimates as level-2 sample size increased.

- Since all the imputation methods retained the original level-1 sample size, they generated no or trivial bias on level-1 fixed and random parameter estimates.
However, they produced slightly downward-biased intercept estimates. LD reduced level-1 sample size by excluding cases with missing data and thus overestimated level-1 intercept and fixed effect parameter estimates and underestimated level-1 random effect with more bias.

- With missing data occurring at level-2, all the MDTs generated more biased estimates for level-2 fixed and random effects. Every MDT generated more bias in level-2 random effect estimates than in level-2 fixed effect estimates. Under the small sample size condition, LD produced the least biased estimates (but still deviant) for both fixed and random effect estimates while other MDTs produced more biased estimates, with the biases strikingly large at 40% missingness. Under medium and large sample sizes, REM and LD generated the least biased estimates for fixed effects for variables with and without missing data; IEM generated the least biased estimates of level-2 random effect.

- For level-3 fixed effect estimation, REM and MS produced the least biased estimates. For level-3 random effect estimation, the imputation methods tended to underestimate the estimates under small and medium sample sizes but to overestimate under large sample sizes. LD, REM and MS generated less deviant level-3 random effect estimates.

Accuracy and Precision of Parameter Estimates

Accuracy and precision were evaluated with root-mean-square error (RMSE) (Rayman & Roberts, 1987; Roth & Switzer, 1995; Enders, 2001; Collins et. al, 2001). As introduced in Chapter III, RMSE, a measure of overall accuracy and precision, was calculated as the square root of the average squared difference between the estimate and
the true parameter. RMSE was computed for each design cell (Table 3, Chapter III) over the 1000 replications, thus there was one RMSE value per design cell. A small value of RMSE indicates that the parameter estimates generated under that particular condition were less biased (accuracy) and the 1000 parameter estimates were more clustered around the true parameter (precision). The RMSE index can be interpreted as the size of the “error” in the estimates (Enders, 2001) produced by a certain MDT under a particular experimental condition. RMSE for the estimates of five fixed effects are reported in Table 11 and RMSE for the estimates of three random effects are presented in Table 12.

Accuracy and Precision of the Estimates of Fixed Effects

Research Question 3. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of fixed effects that are accurate and precise when compared with parameter estimates from complete data?

Intercept ($\gamma_{00}$). When the proportion of missingness was low (10%), the RMSE indices indicated that the parameter estimates generated by every MDT were generally accurate and precise (RMSE=.005~.012) except when sample size was small (RMSE=.015~.021). Among all the MDTs, IMI produced more accurate and precise parameter estimates under small sample size (RMSE=.015), and IEM performed better under other sample size conditions (RMSE=.005~.009). For all the MDTs, the RMSE values decrease as the sample size increases, this indicates that MDTs produced more accurate and precise estimates as sample size increase.
Table 11.

**Root-mean-square error (RMSE) for fixed effects estimates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample size&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Missing 10%</th>
<th>Missing 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
<td>MS</td>
</tr>
<tr>
<td>( \gamma_{000} )</td>
<td>1</td>
<td>.019</td>
<td>.021</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>2</td>
<td>.012</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.009</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.009</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.009</td>
<td>.010</td>
</tr>
<tr>
<td>( \gamma_{100} )</td>
<td>1</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>(Level-1 variable)</td>
<td>2</td>
<td>.005</td>
<td>.005</td>
</tr>
<tr>
<td>( X_1 )</td>
<td>3</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.003</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.003</td>
<td>.003</td>
</tr>
<tr>
<td>( \gamma_{010} )</td>
<td>1</td>
<td>.009</td>
<td>.009</td>
</tr>
<tr>
<td>(Level-2 variable)</td>
<td>2</td>
<td>.011</td>
<td>.014</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>3</td>
<td>.011</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.009</td>
<td>.010</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.007</td>
<td>.007</td>
</tr>
<tr>
<td>( \gamma_{020} )</td>
<td>1</td>
<td>.011</td>
<td>.038</td>
</tr>
<tr>
<td>(Level-2 variable)</td>
<td>2</td>
<td>.014</td>
<td>.015</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>3</td>
<td>.008</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.011</td>
<td>.012</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.009</td>
<td>.011</td>
</tr>
<tr>
<td>( \gamma_{001} )</td>
<td>1</td>
<td>.013</td>
<td>.015</td>
</tr>
<tr>
<td>(Level-3 variable)</td>
<td>2</td>
<td>.009</td>
<td>.009</td>
</tr>
<tr>
<td>( W_1 )</td>
<td>3</td>
<td>.008</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.008</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.010</td>
<td>.006</td>
</tr>
</tbody>
</table>

<sup>a</sup> Shaded numbers indicate the smallest value across the rows under different levels of missingness.

<sup>b</sup> Sample size:
- LD—listwise deletion
- MS—mean substitution
- REM—restrictive EM
- IEM—inclusive EM
- RMI—restrictive MI
- MI—inclusive MI

1—60 classes [20 schools (3 classes/school)]
2—180 classes [60 schools (3 classes/school)]
3—180 classes [20 schools (9 classes/school)]
4—360 classes [120 schools (3 classes/school)]
5—360 classes [20 schools (18 classes/school)]
Table 12.

Root-mean-square error (RMSE) for random effects estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Size</th>
<th>Missing 10%</th>
<th>Missing data treatment</th>
<th>Missing 40%</th>
<th>Missing data treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
<td>MS</td>
<td>REM</td>
<td>IEM</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1</td>
<td>.009</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>(Level-1 variance component $\sigma^2$)</td>
<td>2</td>
<td>.006</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.005</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>5</td>
<td>.004</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>$\tau_\pi$</td>
<td>1</td>
<td>.015</td>
<td>.047</td>
<td>.038</td>
<td>.025</td>
</tr>
<tr>
<td>(Level-2 variance component $\tau_\pi$)</td>
<td>2</td>
<td>.007</td>
<td>.009</td>
<td>.008</td>
<td>.005</td>
</tr>
<tr>
<td>4</td>
<td>.006</td>
<td>.013</td>
<td>.013</td>
<td>.006</td>
<td>.019</td>
</tr>
<tr>
<td>5</td>
<td>.005</td>
<td>.011</td>
<td>.011</td>
<td>.003</td>
<td>.020</td>
</tr>
<tr>
<td>$\tau_\beta$</td>
<td>1</td>
<td>.018</td>
<td>.028</td>
<td>.024</td>
<td>.022</td>
</tr>
<tr>
<td>(Level-3 variance component $\tau_\beta$)</td>
<td>2</td>
<td>.009</td>
<td>.006</td>
<td>.005</td>
<td>.006</td>
</tr>
<tr>
<td>3</td>
<td>.006</td>
<td>.005</td>
<td>.005</td>
<td>.007</td>
<td>.006</td>
</tr>
<tr>
<td>4</td>
<td>.007</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.005</td>
</tr>
<tr>
<td>5</td>
<td>.006</td>
<td>.005</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
</tr>
</tbody>
</table>

* Shaded numbers indicate the smallest value across the rows under different levels of missingness.

a: Missing data treatment  
   LD—listwise deletion  
   MS—mean substitution  
   REM—restrictive EM  
   IEM—inclusive EM  
   RMI—restrictive MI  
   IMI—inclusive MI

b: Sample size  
1—60 classes [20 schools (3 classes/school)]  
2—180 classes [60 schools (3 classes/school)]  
3—180 classes [90 schools (3 classes/school)]  
4—360 classes [120 schools (3 classes/school)]  
5—360 classes [180 schools (3 classes/school)]  
6—360 classes [20 schools (18 classes/school)]
When the proportion of missingness was high (40%), LD had the largest RMSE (.051) when sample size was small, but it generated more accurate and precise estimates when there were 20 school with 9 classes per school (sample size 3—medium level-2 total size [180 classes], design B). This performance pattern was reversed for other MDTs that they had the smallest RMSE under small sample size but the largest RMSE under sample size 3. Generally, LD performed well under sample size 3 (medium level-2 total size, design B) and IMI performed better than other MDTs under other sample size conditions.

*Level-1 fixed effect of* $a_1$ ($\gamma_{100}$). All the imputation methods (MS, REM, IEM, RMI and IMI) retained the original sample size, thus these methods generated no or trivial bias (RMSE=.000~.001) in the parameter estimates. Since LD reduced the level-1 sample size, it generated biased estimates (RMSE=.003~.024) under both low and high level of missingness.

*Level-2 fixed effect of* $X_1$ (*variable with missing data*) ($\gamma_{010}$). When the proportion of missingness was low (10%), LD, REM, MS, and IEM performed much better than RMI and IMI under all the sample size conditions. Under the small sample size, LD (RMSE=.023) and IEM (RMSE=.021) generated most unbiased estimates. Under other sample sizes, LD, REM, MS, and IEM (RMSE=.006~.014) all performed well in producing estimates with accuracy and precision. RMI produced the most biased estimates as indicated by the largest RMSE values (RMSE=.026~.066) across the sample sizes.

Under high level (40%) of missingness and with missing data occurring on this variable, all MDTs produced somewhat biased estimates. LD generated the least biased
estimates (RMSE=.113) when sample size was small. LD and REM (RMSE=.013-.025) performed better than other MDTs in generating more accurate and precise estimates under other sample sizes. Again, RMI (RMSE=.099-.300) generated the most biased estimates under all the sample size conditions.

With both levels of missingness, it is noticeable that REM performed better than IEM except for small sample size, and IMI performed better than RMI under all the sample sizes. This finding indicated that for the parameter estimation of the variable with missing data, the restrictive approach outperformed the inclusive approach when EM imputation was used, however, the inclusive approach performed better than the restrictive approach with the MI method.

Level-2 fixed effect of $X_2$ (variable without missing data) ($\gamma_{020}$). The RMSE values indicated a similar pattern of estimation performance by the six MDTs under low and high levels of missingness. LD generated the least biased estimates when sample size was small. IEM had the smallest RMSE values under medium sample sizes with both design A and B, and it also performed the best under large sample size design A (level-2 total size of 360 classes, 120 schools with 3 classes per school). IMI and REM performed better than other MDTs under large sample size design B when there were 120 schools with 3 classes per school.

The comparison of restrictive and inclusive approach revealed that the inclusive approaches appeared to outperform restrictive approach with both EM and MI imputation methods except that REM outperformed IEM under large sample size design A.

Level-3 fixed effect of $W_1$ ($\gamma_{001}$). LD, with both levels of missingness, generated the least biased estimates when sample size was small. REM and MS produced estimates
with better accuracy and precision than other MDTs (RMSE=.004~.008) across both levels of missingness and different sample sizes except when sample size was small.

Regardless of sample size and proportion of missingness, the restrictive EM performed better than the inclusive EM as indicated by the smaller RMSE values for REM (RMSE=.004~.014) and larger RMSE values for IEM (RMSE=.005~.050). RMI and IMI had very similar RMSE values with 10% missingness, but RMI appeared to have slightly smaller RMSE values with 40% missingness.

Accuracy and Precision of the Estimates of Random Effects

Research Question 3. Given different experimental conditions (sample size and proportion of missing and/or their combinations), which of the six MDTs yields parameter estimates of random effects that are accurate and precise when compared with parameter estimates from complete data?

Level-1 random effect ($\sigma^2$). Similar to the parameter estimation performance of level-1 fixed effects, MS, REM, IEM, RMI and IMI restored the original sample size, thus generated no or trivial biases (all RMSEs=.000) under both low and high levels of missingness. It was expected that LD, which reduced the original sample size, produced biased estimates (RMSE=.004~.019).

Level-2 random effect ($\tau_\pi$). All the MDTs generated less accurate and less precise estimates with 40% missingness compared with 10% missingness. With both levels of missingness, all the MDTs performed better under medium and large sample sizes than under small sample size. Under small sample size with both levels of missingness, LD, among all the MDTs, produced the most accurate and precise estimates (RMSE=.015 for 10% missingness and RMSE=.042 for 40% missingness). It was noticed that with 40%
missingness under small sample size, the estimates by LD carried large error (RMSE=.042), but the RMSE values of other MDTs were more than three times of that value (RMSE=.133~.154).

Under other sample sizes with both levels of missingness, IEM performed the best in producing estimates with accuracy and precision (RMSE=.003~.014) and was followed by LD (RMSE=.005~.021) in terms of the estimation performance. With 10% missingness, REM, IMI and MS had RMSE values that were not greatly larger than those for IEM and LD, however, with 40% missing, those three methods generated RMSE values twice or even three times larger than that of IEM.

Since IEM had smaller RMSE values than REM and IMI had smaller RMSE values than RMI under all the sample size and missingness conditions, the inclusive approach appeared to outperform the restrictive approach for the estimation of level-2 random effect.

Level-3 random effect ($\tau_\beta$). Again, with both levels of missingness, all the MDTs generated the most inaccurate and imprecise estimates when sample size was small. Under small sample size with both low and high levels of missingness, LD, once more, performed better than other MDTs (RMSE=.018 for 10% missingess and RMSE=.036 for 40% missingness). Under other sample sizes with both levels of missingness, REM and MS (RMSE=.004~.016) performed better than other MDTs. Although REM and RMI generated smaller RMSE values than IEM and IMI, the differences in RMSE values were generally small (.001~.011) with both levels of missingness, which indicated that the estimates generated by the restrictive and inclusive strategies did not differ greatly on accuracy and precision.
Further Investigation on Parameter Estimates

RMSE, the square root of mean-square-error (MSE), was calculated and evaluated as a measure of overall accuracy and precision for MDT performance in every design cell. Since MSE can also be expressed as the squared bias (bias is the deviation between estimates and the true parameter) plus the variance of the estimates (i.e., \( \text{MSE} = \text{Bias}^2 + \text{Variance} \)) (Enders, 2001; Collins et al., 2001), and MSE indicates the quality of the estimates in terms of bias (accuracy) and variation (precision). Thus, a more accurate estimate with large variance could result in a same RMSE value as a less accurate estimate with less variance. Therefore, mean PEB values with 95% confidence interval (CI of the 1000 PEB values in each cell using the standard deviation of PEB values) for every cell were plotted to assist the evaluation of RMSE values.

The biases of the estimates greatly varied under different levels of missingness, thus figures were plotted separately with 10% and 40% missingness on different scales (Y-axis) for PEB values. Figure 1 and Figure 2 present mean PEB values with 95% CI of fixed effect estimates for 10% and 40% missingness, and Figure 3 and Figure 4 display mean PEB values with 95% CI of random effect estimates under 10% and 40% missingness. In the figures, accuracy (or bias) was indicated by the deviation from the reference line, and the reference line was set at zero which indicated no bias in the estimates. Precision (or variation) was indicated by the length between the upper and lower error bars—a shorter length indicated better precision of the estimates.

Additionally, following Collins et al. (2001), coverage and width were examined to evaluate the performance of the six MDTs. Coverage was calculated as the proportion of times that the true parameter value was included in the range of the 95% confidence
interval of the estimates (CI for each of the 1000 estimates using the standard error of the estimate). Width was calculated as the average length of the 95% confidence interval of the estimates (CI for each of the 1000 estimates using the standard error of the estimate). Coverage below 90% was considered problematic which indicated the parameter estimation produced deviant estimates and/or standard error estimates. Good coverage and short length of width indicated excellent accuracy and precision. If two methods produced similar coverage, the one method yielded shorter width would be preferred. Coverage and width are presented in Table 13 and Table 14.

Parameter estimates of fixed effects. The performance patterns of all the MDTs across sample size were very similar under both levels of missingness. The difference was that the MDTs generated less biases in the estimates with 10% missingness (in Figure 1, PEB was scaled from -.12 to .12) than with 40% missingness (in Figure 2, PEB was scaled from -.50 to .50).

Figure 1 displays that under 10% missingness estimates were generally accurate and precise for intercept and a\textdagger\ main effect ($\gamma_{100}$, level-1 predictor), but LD tended to overestimate and MS, REM and IEM tended to underestimate the intercept across all the sample sizes. RMI and IMI tended to overestimate the intercept under small sample size but underestimate under other sample sizes. LD generated intercept estimates with small bias but less precision (i.e., smaller bias but greater variance) compared with other MDTs. For example, in the first row of Figure 1 (intercept), the comparison between LD and REM under sample size 2 (first panel and third panel in the first row) indicated that the absolute mean PEB values were similar ($\text{PEB}_{LD}=-.010$, $\text{PEB}_{REM}=-.011$), however, the estimates of LD had a larger confidence interval than that of REM, thus the LD estimates
Figure 1. PEB of parameter estimates (fixed effects) with 95% CI under 10% missingness

Missing data treatment
LD—listwise deletion
MS—mean substitution
REM—restrictive EM
IEM—inclusive EM
RMI—restrictive MI
IMI—inclusive MI

Sample size
1—60 classes [20 schools (3 classes/school)]
2—180 classes [60 schools (3 classes/school)]
3—180 classes [20 schools (9 classes/school)]
4—360 classes [120 schools (3 classes/school)]
5—360 classes [20 schools (18 classes/school)]
Figure 2. PEB of parameter estimates (fixed effects) with 95% CI under 40% missingness

Missing data treatment
LD—listwise deletion
MS—mean substitution
REM—restrictive EM
IEM—inclusive EM
RMI—restrictive MI
MI—inclusive MI

Sample size
1—60 classes [20 schools (3 classes/school)]
2—180 classes [60 schools (3 classes/school)]
3—180 classes [20 schools (9 classes/school)]
4—360 classes [120 schools (3 classes/school)]
5—360 classes [20 schools (18 classes/school)]
Figure 3. PEB of parameter estimates (random effects) with 95% CI under 10% missingness

Figure 4. PEB of parameter estimates (random effects) with 95% CI under 40% missingness

Missing data treatment:
LD—listwise deletion
MS—mean substitution
REM—restrictive EM
IEM—inclusive EM
RMI—restrictive MI
IMI—inclusive MI

Sample size
1—60 classes [20 schools (3 classes/school)]
2—180 classes [60 schools (3 classes/school)]
3—180 classes [20 schools (9 classes/school)]
4—360 classes [120 schools (3 classes/school)]
5—360 classes [20 schools (18 classes/school)]
Table 13.

Coverage and width of fixed effects estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Size</th>
<th>Coverage</th>
<th>Width</th>
<th>Missing 10% Coverage</th>
<th>Width</th>
<th>Missing 40% Coverage</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_{00} ) (Intercept)</td>
<td>1</td>
<td>Coverage</td>
<td>100</td>
<td>.38</td>
<td>100</td>
<td>.16</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Width</td>
<td>2</td>
<td>100</td>
<td>.23</td>
<td>100</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Coverage</td>
<td>100</td>
<td>.28</td>
<td>100</td>
<td>.15</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Width</td>
<td>3</td>
<td>100</td>
<td>.42</td>
<td>100</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Coverage</td>
<td>100</td>
<td>.11</td>
<td>100</td>
<td>.06</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Width</td>
<td>4</td>
<td>100</td>
<td>.06</td>
<td>100</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Coverage</td>
<td>100</td>
<td>.06</td>
<td>100</td>
<td>.06</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Width</td>
<td>5</td>
<td>100</td>
<td>.04</td>
<td>100</td>
<td>.04</td>
</tr>
</tbody>
</table>

Note: Sample size 1—60 classes [20 schools (3 classes/school)], 2—180 classes [60 schools (3 classes/school)], 3—180 classes [20 schools (9 classes/school)], 4—360 classes [120 schools (3 classes/school)], 5—360 classes [20 schools (18 classes/school)]. LD—Listwise Deletion, MS—Mean Substitution, REM—Restricted EM, IEM—Inclusive EM, RMI—Restricted MI, IMI—Inclusive MI. Bolded values: coverage < 90%. Shaded values indicate the best coverage or the shortest width across the rows under different levels of missingness.
Table 13—Continued

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Size</th>
<th>Missing 10%</th>
<th>Missing 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LD</td>
<td>MS</td>
</tr>
<tr>
<td>$\gamma_{00}$ (Level-2 predictor variable $X_1$, with missing data)</td>
<td>1 Coverage Width</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2 Coverage Width</td>
<td>.23</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>3 Coverage Width</td>
<td>.16</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>5 Coverage Width</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>$\gamma_{02}$ (Level-2 predictor variable $X_2$, without missing data)</td>
<td>1 Coverage Width</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2 Coverage Width</td>
<td>.21</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>3 Coverage Width</td>
<td>.16</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>5 Coverage Width</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>$\gamma_{001}$ (Level-3 predictor variable $W_3$)</td>
<td>1 Coverage Width</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>4 Coverage Width</td>
<td>.27</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>5 Coverage Width</td>
<td>.16</td>
<td>.16</td>
</tr>
</tbody>
</table>

Note: Sample size 1—60 classes [20 schools (3 classes/school)], 2—180 classes [60 schools (3 classes/school)], 3—180 classes [20 schools (9 classes/school)], 4—360 classes [120 schools (3 classes/school)], 5—360 classes [20 schools (18 classes/school)]. LD—Listwise Deletion, MS—Mean Substitution, REM—Restrictive EM, IEM—Inclusive EM, RMI—Restrictive MI, IMI—Inclusive MI. Bolded values: coverage < 90%. Shaded values indicate the best coverage or the shortest width across the rows under different levels of missingness.
Table 14.
Coverage and width of random effects estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Size</th>
<th>Missing 10%</th>
<th>Missing 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LD</td>
<td>MS</td>
<td>REM</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Coverage Width</td>
<td>.08</td>
<td>.08</td>
<td>.07</td>
</tr>
<tr>
<td>3 Coverage Width</td>
<td>.08</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>4 Coverage Width</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>5 Coverage Width</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>( \tau_\pi )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Coverage Width</td>
<td>.08</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>3 Coverage Width</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>4 Coverage Width</td>
<td>.07</td>
<td>.07</td>
<td>.07</td>
</tr>
<tr>
<td>5 Coverage Width</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>( \tau_\beta )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Coverage Width</td>
<td>.25</td>
<td>.23</td>
<td>.23</td>
</tr>
<tr>
<td>4 Coverage Width</td>
<td>.09</td>
<td>.09</td>
<td>.09</td>
</tr>
<tr>
<td>5 Coverage Width</td>
<td>.27</td>
<td>.27</td>
<td>.27</td>
</tr>
</tbody>
</table>

Note: Sample size 1—60 classes [20 schools (3 classes/school)], 2—180 classes [60 schools (3 classes/school)], 3—180 classes [20 schools (9 classes/school)], 4—360 classes [120 schools (3 classes/school)], 5—360 classes [20 schools (18 classes/school)]. LD—Listwise Deletion, MS—Mean Substitution, REM—Restrictive EM, IEM—Inclusive EM, RMI—Restrictive MI, IMI—Inclusive MI. **Bolded** values: coverage < 90%. □ Shaded values indicate the best coverage or the shortest width across the rows under different levels of missingness.
carried a larger “error” than REM estimates ($\text{RMSE}_{\text{LD}}=.012$, $\text{RMSE}_{\text{REM}}=.007$). With respect to coverage and width of the intercept estimates, every MDT yielded 100% coverage rate (Table 13). IEM appeared to generate the shortest width and LD generated the largest width across sample sizes.

As for the estimation of $aI$ fixed effect, by retaining the original level-1 sample size, all the imputation methods (MS, REM, IEM, RMI and IMI) fully retained the level-1 data, thus they generated $\gamma_{100}$ estimates with no or trivial bias. Although LD reduced the level-1 sample size, it still generated $\gamma_{100}$ estimates with satisfactory accuracy and precision. On the performance of coverage and width, every MDT performed equally well under 10% missingness, but LD had a smaller coverage rate and the largest width under 40% missingness among all the methods (Table 13).

The estimates for level-2 fixed effects $X_1 (\gamma_{010})$ and $X_2 (\gamma_{020})$ exhibited very different performance patterns among the MDTs. For the fixed effect of $X_1$ (level-2 variable with missing data), all the MDTs greatly underestimated $\gamma_{010}$ under small sample size, only LD and IEM performed well on both accuracy and precision. Under other sample sizes, LD and REM, performed better than other MDTs with a slight underestimation, RMI and IMI showed a large underestimation while IEM tended to overestimate $\gamma_{010}$ with some level of bias. All the MDTs performed well on the coverage and width with 10% missingness, but none of them produced satisfactory coverage rate with 40% missingness under small sample size (all <90%). RMI generated the worst coverage rate that it showed zero coverage under several sample sizes.

For the fixed effect of $X_2$ (level-2 variable without missing data), all the MDTs except LD greatly overestimate $\gamma_{020}$ under small sample size, and LD generated the most
accurate and precise estimates with best coverage rate and shortest width. Under other sample sizes, RMI and IMI generated the most biased estimates. IEM and REM outperformed other MDTs in generating estimates with the best accuracy, precision, coverage and width.

For the estimation of level-3 fixed effect $W_I (\gamma_{001})$, every MDT produced 100% coverage rate, and every MDT except LD produced comparable width. MS, REM and RMI presented a similar performance pattern—underestimation under sample size 1 and 3 and overestimation under other sample sizes. The two inclusive methods IEM and IMI consistently underestimated $\gamma_{001}$ across all the sample sizes with slightly more bias among all the MDTs, but they generated 100% coverage with the shortest width. LD performed poorly and tended to underestimate $\gamma_{001}$ under sample size 2 and 4, and it tended to overestimate $\gamma_{001}$ under other sample sizes with better performance. Moreover, LD generated estimates with the largest width which indicated that LD produced the most biased standard errors estimates.

Parameter estimates of random effects. Similar to their performance on the fixed effects estimation, all the MDTs generated estimates with much larger bias and variance under the high level (40%) of missingness (Figure 4, PEB was scaled from -.60 to 1.80) than under the low level (10%) of missingness (Figure 3, PEB was scaled from -.15 to .45), and they exhibited similar performance patterns across sample sizes with different levels of missingness. It should be noted that the reason that PEB was scaled up to 1.80 under 40% missingness (Figure 4) was due to the huge estimation bias under small sample size, and the estimation bias of MDTs under other sample size was within the range of -.60 to .60 (i.e., 60% of deviation from the true parameter).
For level-1 random effect ($\sigma^2$) estimation, all the imputation methods (MS, REM, IEM, RMI and IMI) reserved the original sample size, thus produced no or trivial deviation in the estimates, and they performed equally well on accuracy, precision, coverage and width. Although LD reduced the sample size by excluding the cases with missing data, it produced $\sigma^2$ estimates with satisfactory accuracy and precision even under 40% missingness but with less coverage and larger width (Table 14).

All the MDTs presented the most deviant performance pattern for the estimation of level-2 random effect ($\tau_\pi$) estimates. LD had a quite different performance pattern from other MDTs while the other MDTs (MS, REM, IEM, RMI and IMI) tended to display a similar performance pattern. MS, REM, RMI and IMI consistently overestimated $\tau_\pi$. IEM overestimated $\tau_\pi$ under the small and medium sample sizes but it underestimated $\tau_\pi$ under the large sample size, and it produced the most accurate and precise estimates even under 40% missingness except under the small sample size. All the imputation methods strikingly overestimated $\tau_\pi$ under the small sample size, especially with 40% missingness. With high level of missingness (40%), none of the methods generated satisfactory coverage rates under the small sample size and only IEM and LD produced good coverage rates (>90%) under other sample sizes. LD generated relatively accurate estimates but with larger variance under most sample sizes. In summary, LD produced the best estimates under small sample size and IEM performed the best in producing both accurate and precise $\tau_\pi$ estimates under other sample sizes.

For the estimation of level-3 random effect ($\tau_\beta$), LD tended to overestimate $\tau_\beta$ under all sample sizes. The two inclusive methods (IEM and IMI) consistently
underestimated $\tau_\beta$ under all sample sizes. MS, REM and RMI tended to underestimate $\tau_\beta$ under small and medium sample sizes, but overestimated $\tau_\beta$ under large sample sizes.

With respect to the accuracy and precision of the estimates, IEM generated estimates with some level of bias but with the best coverage and width (100% coverage and the shortest width). LD generated the most accurate but less precise estimates under all sample sizes. Further examination also indicated that LD could not estimate $\tau_\beta$ due to inadequate model fit in 9 out of 1000 replications under small sample size with 40% missingness. It was found previously that LD had a poor performance (the largest RMSE) on level-3 fixed effect ($\gamma_{001}$) estimation under sample size 2 and 4 with 40% missingness. One reason for LD’s poor performance on estimation of $\tau_\beta$ (level-3 random effect) and $\gamma_{001}$ (level-3 fixed effect) with 40% of missingness was the small group size at level-2-sample size 1, 2 and 4 had a level-2 group size of 3 classes per school (design A). With small level-2 group size (i.e., 3 classes per school) LD was more likely to reduce level-3 total size (total number of schools) if all 3 classes nested in a particular school were deleted and thus poor estimation performance due to insufficient sample size at level-3.

**Summary of Accuracy and Precision of the Estimates**

*Fixed effects estimation:*

- Every MDT generated marginally deviant intercept estimates at level-1. All the imputation methods produced no or trivial bias for the level-1 fixed effect estimate. Thus, they performed well (accurate and precise parameter estimation) for level-1 fixed effects. LD performed satisfactorily on the accuracy of the
estimates but produced estimates with the least precision and the longest width of confidence interval.

- With low level of missingness, LD, MS, REM and IEM produced very accurate and precise level-2 estimates for fixed effects of variables with and without missing data. With high level of missingness, REM and LD also performed well on accurate and precise estimation of level-2 fixed effects. The two MI methods generated more biased estimates with both levels of missingness, and IMI performed better than RMI on level-2 fixed effect estimation. It should be noted that under small sample size with 40% missingness, none of the methods produced satisfactory coverage rate.

- With low level of missingness, every MDT generated level-3 fixed effect estimates with good accuracy and precision. With high level of missingness, REM outperformed other MDTs in producing accurate and precise estimates.

*Random effects estimation:*

- All the imputation methods performed well for level-1 random effect estimation by retaining the original level-1 sample size. LD reduced the sample size but still produced estimates with good accuracy and precision.

- Under small sample size, all the MDTs produced inaccurate and imprecise estimates for the level-2 random effect, and they failed to produce satisfactory coverage rate with 40% missingness. Under other sample sizes, IEM generated the most accurate and precise estimates and it also produced 100% coverage rate with the shortest width. Other MDTs produced extremely poor coverage rate with
high level of missingness. RMI generated estimates with least accuracy and precision, and it had zero coverage rate with 40% missingness.

- Under small sample size, LD produced the most accurate and precise level-3 random effect estimates at both levels of missingness. Under other sample sizes, every MDT produced estimates with satisfactory accuracy and precision, but LD consistently generated estimates with the least precision.
CHAPTER V

DISCUSSION

This chapter summarizes the results and findings of the previous chapter about the MDTs performance. This study was designed to compare the performance of six MDTs (listwise deletion, mean imputation, restrictive EM, inclusive EM, restrictive MI and inclusive MI) on parameter estimation in a 3-level HLM model with MAR missing data on one level-2 predictor variable. The comparisons were focused on the estimation bias of fixed effects and random effects produced by each MDT. Two factors—level-2 sample size and proportion of missing data were experimentally manipulated and investigated in the study.

The first section of this chapter discusses the effect of sample size on parameter estimation in this 3-level HLM model. The second section presents the discussion about the performance of each MDT and the comparisons between restrictive and inclusive strategies. The third section summarizes the findings and presents recommendations for research practice. The final section addresses the limitations and suggestions for future studies.

The Effect of Sample Size

The Effect of Level-2 Total Sample Size

The investigation of parameter estimate bias (PEB) revealed that parameter estimates of both fixed and random effects were greatly impacted by sample size. Analysis of variance with repeated measures for parameter estimates indicated that generally sample size had a large effect size on the performance of all the MDTs. The effect of level-2 total sample size was mainly reflected in the performance differences
under sample size condition 1, 2 and 4 where the level-2 total sizes were 60, 180 and 360 classes with fixed group size of 3 classes per school (design A). As a result of increasing level-2 total size, level-3 total size were also increasing as 20, 60 and 120 schools for sample size condition 1, 2 and 4. Therefore, the effect of level-2 total size was also carried over onto the parameter estimation at level-3.

Since level-1 total size and group size were fixed for all sample size conditions and all the imputation methods (MS, REM, IEM, RMI and IMI) retained the original level-1 sample size, the performances of imputation methods on level-1 estimation were not greatly affected by the change of level-2 size. However, LD was impacted by the change of level-2 total size due to the sample size reduction on level-1 parameter estimation. The analysis revealed that the effect of level-2 total size had quite small impact on LD at level-1. LD generated level-1 fixed and random effect estimates with satisfactory accuracy and precision, but it tended to slightly overestimate the parameters.

With missing data occurring at level-2, the impact of small level-2 total size was substantial. Under small sample size (60 level-2 units in this study), all the imputation MDTs (MS, REM, IEM, RMI and IMI) produced the most biased and the most imprecise estimates for all the level-2 and level-3 parameters with both low and high levels of missingness. Moreover, all of the MDTs failed to generate satisfactory coverage under small sample size with 40% missingness. Given the small size, the poor performance of the imputation methods was largely due to lack of information which negatively affected the efficiency of the imputation process. All the imputation methods substantially overestimate level-2 random effect and greatly underestimate level-3 random effect under small sample size. LD was least impacted by the small level-2 sample size, but it
produced quite biased fixed effect estimates for the predictor with missing data and biased level-2 and level-3 random effect estimates. Results also indicated that with small sample size and high level of missingness (40%), none of the MDT was effective in level-2 parameter estimation.

When the level-2 total size increased, the performance of all MDTs improved tremendously. Both level-2 fixed and random estimates generated under larger level-2 size (180 and 360 level-2 units) by all the MDTs were significantly less biased and with better precision. The improvement on the accuracy, precision, coverage and width were also reflected in the level-3 fixed and random effect estimates. This effect of level-2 and level-3 total size was consistent with findings reported in missing data literature that increase in sample size would increase the performance of imputation methods (e.g., mean substitution, regression imputation, EM imputation, and multiple imputation) (Roth & Switzer, 1995; Witta, 1992; Gibson & Olejnik, 2003; Zhang, 2005).

The Effect of Level-2 Group Size

In the HLM literature, it was suggested that the performance of parameter estimation is mainly a function of number of groups (total sample size at the higher level) rather than size of groups (number of lower-level units nested in a higher-level unit) (Bassiri, 1988; Snijders & Bosker, 1999; Darandari, 2004). Studies (Kim, 1990; Moke, 1995; Darandari, 2004) reported that in a 2-level HLM the random effect estimates were affected more by the number of groups (sample size at level-2) than by the number of observations per group (level-1 units per level-2 unit).

The current study found a similar effect on the performance of MDTs in the context of 3-level HLM. Both level-3 fixed and random effect estimates generated by all
the MDTs except LD were more influenced by the increase of level-3 sample size (level-3 total sample size) than by the number of level-2 units per level-3 unit (level-2 group size). Table 15 presented the comparison between the effect of total size and group size as suggested in the literature.

Table 15.

*Comparison between the effects of total size and group size for level-3 estimates*

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Level-3 total size</th>
<th>Level-2 group size</th>
<th>Sample Size Change</th>
<th>RMSE comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2 Sample size 20, 60</td>
<td>3</td>
<td>Increase in level-3 total size</td>
<td>Improvement (size 1 &amp; 2) &gt; Improvement (size 1 &amp; 3)</td>
<td></td>
</tr>
<tr>
<td>1 &amp; 3 Sample size 20</td>
<td>3, 9</td>
<td>Increase in level-2 group size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 &amp; 4 Sample size 20, 120</td>
<td>3</td>
<td>Increase in level-3 total size</td>
<td>Improvement (size 1 &amp; 4) &gt; Improvement (size 1 &amp; 5)</td>
<td></td>
</tr>
<tr>
<td>1 &amp; 5 Sample size 20</td>
<td>3, 18</td>
<td>Increase in level-2 group size</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The comparisons suggested that for both level-3 fixed and random effects estimation, increase in level-3 total size led to a greater improvement on the performance of MDTs than increase in level-2 group size did, which was indicated by a larger RMSE difference between sample size 1 and 2 (also size 1 & 4) and a smaller RMSE difference between sample size 1 and 3 (also size 1 & 5) (Table 15). Moreover, the group size effect presented a similar pattern for level-2 random effect estimation by the MDTs, that is, MDTs' (not including LD) performances on level-2 random effect estimation were more affected by number of level-3 total size than by level-2 group size.
Performance of Missing Data Treatment (MDT)

Listwise Deletion

Listwise deletion, among all the MDTs in the study, produced the most accurate parameter estimates across different levels of missingness for all the HLM parameters under small sample size. The similar findings were reported in Gibson and Olejnik (2003) that with MCAR missingness on a level-2 predictor in a 2-level HLM, listwise deletion produced both fixed and random effect estimates that did not differ significantly from the parameter estimates produced from complete data under small sample size (level-2 total size=30). Listwise deletion, which simply excluded cases with missing data, did not introduce additional variation into the data and the change of the variance-covariance structure of the data was a result of the reduction in sample size. Thus with missingness at level-2 it tended to underestimate parameters of both fixed and random effect at level-2. Generally, in this 3-level HLM analysis listwise deletion appeared to be a viable choice when sample size is small.

Although listwise deletion generated the most accurate estimates and seems to be a good choice when sample size is small, it should be utilized with caution that a) the accuracy of the estimates was relative accuracy in comparison with other MDTs and in fact the bias was still large especially with 40% missingness (24% deviation for the variable with missing data, 89% deviation for level-2 random effect); b) it did not generate estimates with good precision and for some parameters it generated estimates with the largest variance (imprecise estimates); c) with 40% missing data it failed to estimate level-3 random effect 9 out of 1000 times due to insufficient data as a result of sample size reduction; d) sample composition or representativeness is changed due to the
deletion of cases with MAR missingness and therefore the validity of making statistical inferences based on the reduced sample to the population might be challenged.

Mean Substitution and Restrictive EM Imputation

Mean substitution and restrictive EM imputation presented very similar performance patterns on both fixed and random effects parameter estimation in this 3-level HLM. Restrictive EM imputation consistently produced better estimates than mean substitution across sample size and proportion of missingness. Under medium or large sample size (level-2 total size ≥ 180) restrictive EM imputation consistently produced most accurate or satisfactory estimates with low level of missingness (10%), and it performed satisfactorily with high level (40%) of missingness on the estimation of most parameters except that it generated quite biased level-2 random effect estimate.

These findings to some extent agreed with Dalton and Schultz (2006). They found that with MAR missingness on a level-2 predictor variable, restrictive EM method fully recovers the parameters of both fixed and random effects in a 2-level hierarchical model with 10%, 30% and even 50% of missing data. A possible explanation for the performance difference might be their use of MAR missingness variable (the cause of missingness) in the EM imputation process. They employed the variable which caused the MAR missingness in their imputation, which means that they included 100% of the cause of missingness and the imputation process was able to use the full information to re-produce the missing data. As Collins, Schafer and Kam (2001) pointed out that when missingness exceeds 25% excluding the cause of missingness in the EM or MI imputation would result in biased estimates. In the current study, the variable used in restrictive EM method was only moderately correlated (.31) with the cause of
missingness, thus with high proportion of missing data restrictive EM method produced biased estimates. It is likely that increase in the level of correlation between the variable/s utilized in imputation and the cause of missingness would increase the performance of the restrictive EM method.

In practice, it is unlikely to include 100% of the cause of missingness in the imputation process because the reason for missing data is usually unknown or can only be speculated. Researchers could consider including a large set of variables in the EM imputation which can reduce the chance of excluding important auxiliary variables and improve the performance of EM in imputing missing data.

**Inclusive EM Imputation**

The inclusive EM method presented a somewhat different performance pattern as restrictive EM method did on level-2 parameter estimation. It overestimated the parameter for level-2 predictor $X_I (\gamma_{00})$ on which the missing data occurred while other MDTs all underestimated $\gamma_{00}$ and restrictive EM performed well on $\gamma_{00}$ estimation. However, the inclusive EM method outperformed all other MDTs on the estimation of level-2 random effect even with 40% of missingness. Generally speaking, the inclusive EM method performed satisfactorily on both fixed and random effect estimation with 10% missingness but it generated somewhat biased fixed effect estimates with 40% missingness. It greatly outperformed all other MDTs on level-2 random effect estimation under both low and high levels of missingness.

**Restrictive MI and Inclusive MI**

The MI methods did not perform well on the parameter estimation in the HLM analysis, and the two MI methods performed especially poorly on the estimation at level-
2 where the missing data occurred. The reasons might be the insufficient sample size and the moderate correlation (.31) between the cause of missingness and the variable/s used in MI process (auxiliary variable).

Graham and Schafer (1999) studied the performance of MI under small sample size (N=50) and concluded that the problem of MI with small sample size lies in the small size itself, not the MI procedure and MI cannot compensate for having too little data. In Schafer and Olsen (1998), Schafer (1999a) and Allison (2000), they all mentioned that MI works effectively with sufficiently large sample size or effective sample size (number of non-missing cases). In fact, the two MI methods did reveal a performance improvement under large sample size in this study. However, there are no clear guidelines about how large the sample size can be defined as sufficient because the number of sample size needed is also related to the proportion of missingness and the correlation ($r_{yz}$) between the cause of missingness and the auxiliary variable. For example, it is shown by Graham (2008) that with original sample size of 500, 25% of missing data with $r_{yz} = .2$, 33% of missingness with $r_{yz} = .6$ and 50% of missingness with $r_{yz} = .83$ can achieve the same level of efficiency. Thus, the MI methods would have performed better if the correlation between the cause of missingness and the number of auxiliary variables in MI procedure were higher and with larger sample size.

**Restrictive and Inclusive Methods**

As cautioned by Schafer and Olsen (1998) that the variables used to impute the data should contain all the variables to be used in analysis. However, if level-1 variables (both outcome variable and predictor variable) were used in the inclusive method (IEM and IMI) in their original values, that would have resulted in a unique imputed value for
the level-2 variable having missing values for each level-1 observation (student) and the multivariate normality would have been violated in the imputation process. Thus, in the current study level-1 variables were aggregated to level-2 groups (units) and the aggregated means were used in the imputation.

This study revealed that the inclusive and restrictive methods performed differently on parameter estimation at level-2 (where missing data occurred), and the two strategies influenced the performance of EM imputation and MI differently.

EM imputation worked better with restrictive strategy on fixed effect estimation while it worked better with inclusive method on random effect estimation. The imputed values generated by restrictive EM method tended to be around the regression line (with the variable used in imputation) which was mostly parallel with the regression line had there been no missing data but biased downward. The inclusive EM methods, with the information from the aggregated level-1 variables, tended to generate imputed values that tried to restore the variation pattern in the data but with some deviation. Therefore, the restrictive EM performed better on fixed effect estimation and the inclusive EM performed better on random effect estimation.

Multiple Imputation (MI) worked better with inclusive strategy than restrictive strategy on both fixed and random effect estimation at level-2. Although neither of the MI methods performed effectively in comparison with other MDTs at level-2 estimation, inclusive MI outperformed restrictive MI on all the parameter estimation across sample sizes and proportion of missingness. This finding indicated that including more variables in imputation process would lead to a more effective performance of MI.
Collins et al (2001) compared the restrictive and inclusive strategy on Maximum-likelihood (ML) and MI performance and concluded that for both ML and MI imputation, "far from being harmful, the inclusion of variables is at worst neutral, and at best extremely beneficial" (p. 348). Different from Collins et al (2001) on the performance of restrictive on inclusive EM (ML) methods, Dalton and Schultz (2006) reported that restrictive EM method outperformed inclusive EM method in both fixed and random parameter estimation in a 2-level HLM model with MAR on level-2.

This study in part agreed with Dalton and Schultz (2006) that restrictive EM method outperformed inclusive EM method on the estimation of fixed effect in HLM model, but disagreed with them that in this study inclusive EM method outperformed restrictive EM method on random effects estimation.

The findings of this study also partly agreed with Collins et al (2001) about the better performance of inclusive method over restrictive method for EM and MI imputation. This study found that inclusive EM method produced more biased fixed effect estimates than restrictive EM did, but it performed a lot better than restrictive method for random effect estimation.

Summary of Results and Conclusions

Missing Data Treatments (MDT) have been explored in the context of single-level data analysis and multilevel analysis with two-levels. This study compared six MDTs in a 3-level hierarchical model with MAR missingness. Listwise deletion, the default procedure in many statistical packages, was a viable choice when sample size is small. However, it tended to produce estimates with least precision and it should be used cautiously when the proportion of missingness is large (40%). Researchers who decide to
employ listwise deletion should also be aware that listwise deletion is valid under MCAR but not under MAR where the true randomness of missing data is violated, thus extra consideration should be given when researchers make statistical interpretation or conclusion based on the reduced sample. Under special situations like a small sample size with large proportion of missingness, researchers should seriously consider the appropriateness of conducting HLM analysis since all of the MDTs in this study failed to produce estimates with satisfactory coverage rate.

The restrictive EM imputation method was found to be effective in fixed effect parameter estimation in the 3-level HLM analysis but it produced biased estimates for the random effects in terms of accuracy and precision and it was not effective in terms of coverage and width. Mean substitution presented a similar performance pattern as restrictive EM, but it produced even more biased estimates and it also requires MCAR assumption. Thus mean substitution is not recommended to deal with MAR missingness in 3-level HLM analysis.

The inclusive EM method was very effective in random effect estimation. It outperformed all other MDTs on random effect estimation and it produced fixed effect estimates with good accuracy, coverage and width under 10% missingness, but it was problematic on level-2 fixed effect estimation (the variable with missing data) under 40% missingness.

The MI methods did not perform well in producing estimates with good accuracy and precision. However, with larger sample sizes and higher correlations between the cause of missingness and auxiliary variable or variables, the performance of inclusive MI procedure could improve tremendously.
Limitations of the Study

This study employed a 3-level random intercept (intercept-only) model with one level-1, two level-2 and one level-3 variables. The true parameters for fixed and random effects were chosen to reflect typical values found in the 3-level HLM studies, so the values for fixed effect parameters and the random effect parameters (intra-class correlations, or proportion of variance at each level) were held constant. Other models or other values for the true parameters might lead to different findings.

The 3-level HLM model adopted in the study did not include the cross-level interaction effect. In the 3-level HLM model in this study, the intercepts at each level varied while the slopes were not allowed to vary. The exclusion of the cross-level interaction was based on the consideration that in practice, researchers need theoretical suggestions or strong statistical evidence to include a cross-level effect in a multilevel model. Moreover, in the literature, there is variety in the model specifications that studied cross-level effect in a 3-level context (e.g., interaction between level-1 and level-2 or interaction between level-2 and level-3) and it is difficult to research on the typical range of effect. Adopting a model with intercept-and-slope-as-outcome and including a cross-level effect would be useful in understanding and comparing the performance of MDTs under different HLM models.

All the sample size conditions in the present study were balanced designs, which might not be the case with real data. This study investigated limited levels of sample size, did not cover all the situations in real practice. Parameter estimates produced by the MDTs might be more biased under unbalanced design (Raudenbush and Bryk, 2002), but
this bias will decrease as the total number of level-2 or level-3 units increase or as the imbalance of sizes decreases (Darandari, 2004).

In this study, univariate normality was assumed for all the variables (independent and dependent variables), and multivariate normality was established in EM and MI procedures. Zhang (2005) reported that the nonnormality conditions, severe or not, did not significantly affect EM or MI imputation performances; however, with violation of multivariate normality EM or MI methods might generate slightly different imputed values.

Future Research

Future studies can be conducted to further explore the performance of MDTs in 3-level HLM analysis. For example, new studies could adopt more complex 3-level models that include cross-level interaction effects and/or more predictor variables. Studies could include more variation in sample sizes, magnitude of fixed and random effect, more complex missingness pattern (i.e., missingness at more than one variables) and more variability in or combinations of the types of missingness (i.e., MCAR and/or MAR and/or MNAR, or missingness among the levels). Other research can be considered to study the impact of the correlation among the variable with missing data and the auxiliary variables on the performance of EM and MI imputation methods, or to investigate the performance of imputation methods when level-2 missingness is imputed within level-3 units (i.e., impute level-2 missing data by level-3 group) or to explore other ways in selecting auxiliary variables into EM or MI procedure in the context of multilevel models.
REFERENCES


Bock (Eds.), *Multilevel Analysis of Educational Data* (pp. 159-204). San Diego, CA: Academic Press.


Appendix A

SAS Program for Monte Carlo Simulation
proc printto log='D:\MyFiles\mclog2.log';run;
proc printto print='D:\MyFiles\mclst2.lst' new; run;

%macro simu(o);

proc datasets kill;run;
quit;

%do o=1 %to &o;
%do ss=1 %to 5;
%if &ss=1 %then %do;%let mo=sm;%let sch=20;%let class=3;
%let stu=20;%end;
%if &ss=2 %then %do;%let mo=meda;%let sch=60; %let class=3;
%let stu=20;%end;
%if &ss=3 %then %do;%let mo=medb;%let sch=20; %let class=9;
%let stu=20;%end;
%if &ss=4 %then %do;%let mo=rgb;%let sch=20; %let class=18;
%let stu=20;%end;
%if &ss=5 %then %do;%let mo=rga;%let sch=120; %let class=3;
%let stu=20;%end;

data hlm ;
do sch=1 to &sch;
   u=.316*rannor(&sch);   *school level, i.e. level 3 model;
w1=rannor(0);           *school level random effect;
b0=.5+.3*w1+u;           *school level intercept;
do class=1 to &class;
   r=.349*rannor(&class); *classroom random effect;
x1=rannor(0);           *classroom predictor1;
x2=.3*x1+sqrt(1-.3*.3)*rannor(0);  *classroom predictor2 (.3 correlation with predictor1, bivariate normal);
za=.3*x1+sqrt(1-.3*.3)*rannor(0);  *missingness operator on X1 (.3 correlation with x1, bivariate normal); 
pi0=b0+.3*x1+.3*x2+r;  *classroom intercept;
do stu=1 to &stu;
   e=.882*rannor(&stu);  *student level, i.e., level 3 model;
a1=rannor(0);          *student random effect;
Y=pi0+.3*a1+e;          *student predictor;
output;
end;
end;
end;
keep sch class stu Y a1 x1 x2 w1 za;

data hlm ;
set hlm ;
t=t1;  *constant for later merging;
id=_N_; *SUBJ id;
classid=&class*(sch-1)+class; *creates a unique classroom ID;
run;

*************** HLM on complete dataset ***************;
ods output Mixed.CovParms=ran; *random effect estimates;
ods output Mixed.SolutionF=fix; *fixed effects estimates;
proc mixed data=hlm covtest noclprint noitprint noinfo;
class class sch;
model y=al x1 x2 wl /solution ddfm=bw;
random intercept/sub=class(sch) ;
random intercept/sub=sch ;
run;

data ran;
set ran;
repl=&o ;
mis="m10";
run;

data fix;
set fix;
repl=&o ;
mis="m10";
proc append base=ran_&mo data=ran force;
proc append base=fix_&mo data=fix force;

*********** Loop that creates missingness ***********;
%do m=1 %to 2;
  *10% missingness;
  %if &m=1 %then %do;%let mis=m10;%let spn=%eval((&sch*&class)/10);%end;
  *40% missingness;
  %if &m=2 %then %do;%let mis=m40;%let spn=%eval((&sch*&class)*4/10);%end;
*********** Missing 10% or 40% ******************
proc means data=hlm noprint;
class classid;
  var za x1 x2;output out=b_comp mean(za)=za mean(x1)=x1
  mean(x2)=x2 mean(zb)=zb mean(al)=al mean(y)=yy mean(wl)=wl;
data b_comp;
set b_comp;
t=1;
if classid= . then delete;
proc means data=b_comp noprint;
  var za;output out=b1_comp p50=p50_z;
  *restricting upper limit for MAR;
data b2_comp;
set b1_comp ;t=1;
data b3_comp;
merge b_comp b2_comp;
by t; drop _type_ _freq_; data ml; set b3_comp; xx1=x1; if za>p50_z then z0=0; else z0=1; run;

*Select 10% or 40% of level-2 cases for missing in the lower half of za; proc surveyselect data=ml sampsize=&spn method=pps out=m3; size z0;
data m4;set m3;xx1=.;v=0;keep v classid xx1;
proc sort;by classid;
proc sort data=ml;by classid;
data m5;merge ml m4;by classid;
   * m5 is the level-2 dataset with missing data;
data m10;merge hlm m5 (keep=classid v xx1);
   by classid;if v=. then v=1;run;
   * m10 is the HLM dataset with missing data on level-2 variable xx1;

*************** Listwise deletion--HLM  ***************

proc mixed data=m10 covtest noclprint noitprint noinfo;
class class sch;
model y=a1 xx1 x2 w1/solution ddfm=bw;
random intercept /sub=class(sch);
random intercept /sub=sch;
ods output solutionf=fixld;
ods output CovParms=ranld;
data ranld;set ranld;
   repl=&o;mis="&mis";
data fixld;set fixld;
   repl=&o;mis="&mis";
proc append base=ranld_&mo data=ranld force;
proc append base=fixld_&mo data=fixld force;

************************ Mean substitution ****************************
proc means data=m5  noprint;
   var xx1;
   where xx1>.;
   output out=ms1 mean(x1)=xx1bar;
data ms2;
   set ms1; t=1;
data ms3;
   merge ms2 m10; by t;
data ms;
   set ms3;
   if xx1=. then xx1=xx1bar;run;

***** Mean Substitution-HLM *****
proc mixed data=ms covtest noclprint noitprint noinfo;
class class sch;
model y=a1 xx1 x2 w1/solution ddfm=bw;
random intercept /sub=class(sch);
random intercept /sub=sch;
ods output solutionf=fixms;
ods output CovParms=ranms;run;
data ranms;set ranms;repl=&o;mis="&mis";
data fixms;set fixms;repl=&o;mis="&mis";
proc append base=ranms_&mo data=ranms force;
proc append base=fixms_&mo data=fixms force;

*********** Restrictive EM-imputation ***********
proc mi data=m5 seed=68923 nointprint;	em out=remimp;
var xxl x2;run;
data remimpl;
set remimp;
keep classid xxl v;
proc sort;by classid;

proc sort data=hlm;by classid;
data rem;
merge hlm remimpl;
by classid;
run;

***** REM-imputed-HLM *****
proc mixed data=rem covtest nointprint noitprint noinfo;
class class sch;
model y=al xxl x2 w1/solution ddfm=bw;
random intercept /sub=class(sch);
random intercept /sub=sch;
ods output solutionf=fixrem;
ods output CovParms=ranrem;run;
data ranrem;set ranrem;repl=&o;mis="&mis";
data fixrem;set fixrem;repl=&o;mis="&mis";
proc append base=ranrem_&mo data=ranrem force;
proc append base=fixrem_&mo data=fixrem force;

************ Inclusive EM-imputation ************
proc mi data=m5 seed=68923 nointprint;	em out=iemimp;
var xxl x2 yy al;run;
data iemimpl;
set iemimp;
keep classid xxl v;
proc sort;by classid;
proc sort data=hlm;by classid;
data iem;
merge hlm iemimpl;
by classid;
run;
****** IEM-imputed-HLM ******;
proc mixed data=iem covtest noclprint noitprint noinfo;
class class sch;
model y=al xx1 x2 wl/solution ddfm=bw;
random intercept /sub=class(sch);
random intercept /sub=sch;
ods output solutionF=fixiem;
ods output CovParms=raniem;
run;
data raniem;set raniem; repl=&o; mis="&mis";
data fixiem;set fixiem; repl=&o; mis="&mis";
proc append base=raniem_&mo data=raniem force;
proc append base=fixiem_&mo data=fixiem force;

*************** Restrictive MI-imputation ****************************;
*create dataset for merge;
proc mi data=mlO seed=68923 out=impl simple nimpute=5 noprint;
mcmc chain=single initial=em;
var xx1 x2; run;
data rmil (drop=xx1); set impl; classid=&class*(sch-1)+class;

*****************************;
*Impute on level-2 dataset;
proc mi data=m5 seed=68923 out=imp simple nimpute=5 noprint;
mcmc chain=single initial=em niter=200;
var xx1 x2; run;
data rmi2 ; set imp ;

data rmi;
merge rmil rmi2(keep=_imputation_ classid xx1); by _imputation_ classid; run;

****** RMI-impute-HLM ******;
proc mixed data=rmi covtest noclprint noitprint noinfo;
class class sch;
model y=al xx1 x2 wl/solution ddfm=bw; by _imputation_; 
random intercept /sub=class(sch);
random intercept /sub=sch;
ods output solutionF=outfix_mi;
ods output CovParms=outran_mi;
proc sort data=outfix_mi;
by effect;
proc means noprnt;
var estimate StdErr;
by effect;
output out=firxi mean=estimate stderr;
data outran_mi;
set outran_mi;
if Subject="" then Subject="residual";
proc sort data=outran_mi;
by subject;
proc means noprnt;
by subject;
var estimate StdErr;
output out=ranrmi mean=estimate stderr;run;

data ranrmi;set ranrmi;repl=&o;mis="&mis";
data fixrmi;set fixrmi;repl=&o;mis="&mis";
proc append base=ranrmi_&mo data=ranrmi force;
proc append base=fixrmi_&mo data=fixrmi force;

*************** IMI-imputation on level-2 data ***********************;
proc mi data=m5 seed=68923 out=imil simple nimpute=5 noprint;
mcmc chain=single initial=em niter=200;
var yy al xx1 x2;

data imi2;
set imi1;
rename xx1=xxlim;
data imi2;
set imi2;
keep _imputation_ classid xxlim;
data imi;
merge rmi imi2;
by _imputation_ classid;
run;

***** IMI-impute-HLM *****;
proc mixed data=imi covtest noclprint noitprint noinfo;
class class sch;
model y=al xxlim x2 w1/solution ddfm=bw;by _imputation_; 
random intercept /sub=class(sch);
random intercept /sub=class(sch);
ods output solutionf=outfix_imi;
ods output CovParms=outran_imi;
proc sort data=outfix_imi;
by effect;
proc means noprint;
by effect;
var estimate StdErr;
output out=fiximi mean=estimate stderr;

data outran_imi;
set outran_imi;
if Subject="" then Subject="residual";
proc sort data=outran_imi;
by subject;
proc means noprint;
by subject;
var estimate StdErr;
output out=ranimi mean=estimate stderr;

data ranimi;set ranimi;repl=&o;mis="&mis";
data fiximi;set fiximi;repl=&o;mis="&mis";
proc append base=ranrmi_&mo data=ranrmi force;
proc append base=fixrmi_&mo data=fixrmi force;

***********************************************************************;
data ran1_&mo;
  set ran_&mo;
  if Subject=" " then Subject="residual";
data ran2_&mo;
  set ran_&mo;
  if Subject=" " then Subject="residual";
  mis="m40";
proc append base=ran1_&mo data=ran2_&mo force;
proc sort data=ran1_&mo; by repl subject mis;

data ranld1_&mo;
  set ranld_&mo;
  if Subject=" " then Subject="residual";
  rename estimate=est_ld stderr=se_ld;
  proc sort; by repl subject mis;
data ranms1_&mo;
  set ranms_&mo;
  if Subject=" " then Subject="residual";
  rename estimate=est_ms stderr=se_ms;
  proc sort; by repl subject mis;

data ranreml_&mo;
  set ranrem_&mo;
  if Subject=" " then Subject="residual";
  rename estimate=est_rem stderr=se_rem;
  proc sort; by repl subject mis;

data ranieml_&mo;
  set raniem_&mo;
  if Subject=" " then Subject="residual";
  rename estimate=est_iem stderr=se_iem;
  proc sort; by repl subject mis;

data ranimid_&mo;
  set ranimi_&mo;
  rename estimate=est_imi stderr=se_imi;
  proc sort; by repl subject mis;

data ranrmi_&mo;
  set ranrmi_&mo;
  rename estimate=est_rmi stderr=se_rmi;
  proc sort; by repl subject mis;

data ranall_&mo (drop= _type_ _freq_ covparm) ;
  merge ran1_&mo ranld1_&mo ranms1_&mo ranreml_&mo ranieml_&mo ranimid_&mo ranrmi_&mo ;
  by repl subject mis;
  smpn=&ss;
  run;

******************************************************************************


data fix1_&mo;
  set fix_&mo;

data fix2_&mo;
  set fix_&mo;
  mis="m40";
proc append base=fix1_&mo data=fix2_&mo force;
proc sort data=fix1_&mo; by repl effect mis;

data fixld1_&mo;
  set fixld_&mo;
  rename estimate=est_ld stderr=se_ld;
  if effect="xxl" then effect="x1";
proc sort; by repl effect mis;
data fixmsl_&mo ;
   set fixms_&mo ;
   rename estimate=est_ms stderr=se_ms;
   if effect="xxl" then effect="xl";
   proc sort; by repl effect mis;
data fixreml_&mo ;
   set fixrem_&mo ;
   rename estimate=est_rem stderr=se_rem;
   if effect="xxl" then effect="xl";
   proc sort; by repl effect mis;
data fixieml_&mo ;
   set fixiem_&mo ;
   rename estimate=est_iem stderr=se_iem;
   if effect="xxl" then effect="xl";
   proc sort; by repl effect mis;
data fiximil_&mo ;
   set fiximi_&mo ;
   rename estimate=est_imi stderr=se_imi;
   if effect="xxlim" then effect="xl";
   proc sort; by repl effect mis;
data fixrml_&mo ;
   set fixrmi_&mo ;
   rename estimate=est_rmi stderr=se_rmi;
   if effect="xxl" then effect="xl";
   proc sort; by repl effect mis;
data fixall_&mo (drop=_type_ _freq_) ;
   merge fixl_&mo fixldl_&mo fixmsl_&mo fixreml_&mo fixieml_&mo fiximil_&mo fixrml_&mo ;
   by repl effect mis;
   smpn=&ss;
   run;
%end;
%end;
%end;
proc append base=allran data=ranall_sm force;run;
proc append base=allran data=ranall_med data=ranall_lrg data=ranall_sm data=ranall_lrga data=ranall_lrgb force;run;
%end;
%simu(o=1000);
Appendix B

Generalized Omega Squared Calculation
Formulas for Generalized Omega Squared Calculation Based on Olejnik and Algina (2003, p 444-445)

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\omega^2$ (2 Between* 1 Within Design)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between</strong></td>
<td></td>
</tr>
<tr>
<td>N (A)</td>
<td>$\frac{SS_A - df_A \times MS_{bet}}{SS_A + (N- df_A) \times MS_{bet} + N(K-1)MS_{within}}$</td>
</tr>
<tr>
<td>M (B)</td>
<td>$\frac{SS_C - df_C \times MS_{bet}}{SS_C + (N- df_B) \times MS_{bet} + N(K-1)MS_{within}}$</td>
</tr>
<tr>
<td>NxM (AB)</td>
<td>$\frac{SS_{AB} - df_{AB} \times MS_{bet}}{SS_{AB} + (N- df_{AB}) \times MS_{bet} + N(K-1)MS_{within}}$</td>
</tr>
<tr>
<td><strong>Within</strong></td>
<td></td>
</tr>
<tr>
<td>MDT (C)</td>
<td>$\frac{SS_C - df_C \times MS_{within}}{SS_C + N \times MS_{bet} + (N-1)(K-1)MS_{within}}$</td>
</tr>
<tr>
<td>MDTxN (CA)</td>
<td>$\frac{SS_{CA} - df_{CA} \times MS_{within}}{SS_{CA} + N \times MS_{bet} + (N- df_A)(K-1)MS_{within}}$</td>
</tr>
<tr>
<td>MDTxM (CB)</td>
<td>$\frac{SS_{CB} - df_{CB} \times MS_{within}}{SS_{CB} + N \times MS_{bet} + (N- df_B)(K-1)MS_{within}}$</td>
</tr>
<tr>
<td>MDTNxM (CAB)</td>
<td>$\frac{SS_{CAB} - df_{CAB} \times MS_{within}}{SS_{CAB} + N \times MS_{bet} + (N- df_{AB})(K-1)MS_{within}}$</td>
</tr>
</tbody>
</table>

N—total number of observations, K—number of levels in C (MDT).

*Both between factors are manipulated factors.