A MONTE CARLO INVESTIGATION OF ROBUSTNESS TO NONNORMAL INCOMPLETE DATA OF MULTILEVEL MODELING

A Dissertation

by

DUAN ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Educational Psychology

A MONTE CARLO INVESTIGATION OF ROBUSTNESS TO NONNORMAL INCOMPLETE DATA OF MULTILEVEL MODELING

A Dissertation

by

DUAN ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Victor L. Willson
Committee Members,	James F. McNamara
	David J. Martin
	F. Michael Speed
Head of Department,	Michael Benz

August 2005

Major Subject: Educational Psychology

ABSTRACT

A Monte Carlo Investigation of Robustness to Nonnormal Incomplete Data of Multilevel Modeling. (August 2005)

Duan Zhang, B.S., University of International Business and Economics, China;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Victor L. Willson

Due to its increasing popularity, hierarchical linear modeling (HLM) has been used along with structural equation modeling (SEM) to analyze data with nested structure. In spite of the extensive research on commonly encountered problems such as violation of normality and missing data treatment within the framework of SEM, these areas have been much less explored in HLM. The present study compared HLM and multilevel SEM through a Monte Carlo study from the perspectives of the influence of nonnormality and performance of multiple imputation based on the expectation-maximization (EM) algorithm under various combinations of sample sizes at two levels. The statistical power, parameter estimates, standard errors, and estimation bias for the main effects and cross-level interaction in a two-level model were compared across the four design factors: analysis method, normality condition, missing data proportion, and sample size. HLM and multilevel SEM appeared to have similar power detecting the main effect, while HLM had better power for the cross-level interaction. Neither seemed to be sensitive to violation of the normality assumption. A higher proportion of missing data resulted in larger standard errors and estimation bias. Sample sizes at both the individual and cluster levels played a role in the statistical power for parameter estimates. The two-way interactions for the four factors were generally nonzero. Overall, both HLM and multilevel SEM were quite robust to violation of normality. SEM appears more useful in more complex path models while HLM is superior in detecting main effects. Multiple imputation based on the EM algorithm performed well in producing stable parameter estimates for up to 30% missing data. Sample size design should take into account the level at which the research is most focused.

ACKNOWLEDGEMENTS

This dissertation would not be possible without the help of several people. First of all, I would like to express my deep gratitude and intellectual indebtedness to the chair of my advisory committee, Dr. Victor L. Willson. I feel so fortunate to have had him leading me through all my graduate school years. His patient guidance, insightful comments, reflections, and feedback on my research were indispensable ingredients in my successful completion of the dissertation. His energizing research spirit and unprecedented dedication to students has been my very first incentive and inspiration to pursue a career in academia.

In the same manner, I would like to extend my heartfelt thanks to Dr. Jan N. Hughes who assisted me doing empirical research with quantitative methods and introduced me to the world of child development. Like a mother and friend, she showed me how to balance one’s career and personal life as a successful female professor.

I also would like to thank the members of my advisory committee - Dr. James F. McNamara, Dr. David J. Martin, and Dr. F. Michael Speed, for their time, constructive feedback and encouragement.

Last but not least, a special appreciation is directed to my mother who is my greatest source of motivation. She has always been there listening to my dreams, encouraging me to pursue them and sharing my joys unfolding them. My special gratitude also goes to my dear husband, Jiansheng Lei, for his love for me, his endurance with me, his true belief in me, and for his hard work all these years to support me and our family.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

In many areas of behavioral sciences the research involves hierarchical data. In education researchers may examine how teacher practice influences students' peer relationships within classrooms. In sociology people are nested within family household, then within communities. Their attitudes toward online shopping might be affected by the community they live in. Psychologists, on the other hand, would be more often interested in studying how persons' psychological functioning or behaviors change over time. When subjects are measured repeatedly, the observations are nested within subjects, and the growth trajectories may interact with subjects' characteristics such as ethnicity or socioeconomic status.

All the data in the above examples have more than one unit of analysis since variables measured from all the levels should be taken into account. They share one common trait: depending on the clustering structure, the observations collected within the same cluster tend to aggregate together more than those from different ones. Such data violate one of the three primary assumptions for Ordinary Least Squares Estimation (OLS): observation independence (the other two are normality and homogeneity of variance). Thus, it is not appropriate to handle these data with statistical procedures

This dissertation follows the style of *Journal of Educational and Behavioral Statistics*.

within an OLS framework such as multiple regression. Julian (2001) found that when the intra-class correlation exceeded a minimal level, ignoring such hierarchical structures and treating the data as single-level will result in biased fit indices and estimation problems in parameters and their standard errors. Multilevel analysis techniques using hierarchical linear modeling (HLM) and structural equation modeling (SEM) (Raudenbush & Bryk, 2002; Hox, 2002) have been developed to address these challenges by taking the hierarchical structure into account in estimation to yield parameter estimates with appropriate standard errors.

HLM has been widely applied in educational and behavioral sciences as the currently-preferred method for multilevel data. This has been based on assumptions of improved estimation of standard errors of individual effects and appropriate partitioning of variance-covariance components under a random sampling model. This technique is capable of fitting random effect models, contextual effect models and cross-level interaction models (Hox & Kreft, 1994). Being in the general framework of multivariate statistics, HLM maintains the traditional assumptions of linear model analysis for linearity and normality, but allows violation those of homoscedasticity and more importantly, observation independence (Raudenbush & Bryk, 2002).

Being a general and powerful multivariate framework, structural equation modeling (SEM) embraces most traditional procedures such as path analysis, factor analysis, discriminant analysis, and canonical correlation models (Hox, 2002). It allows researchers to test the plausibility of their theoretical models and to define and adjust these models for subsequent investigation. In addition, many articles have demonstrated that SEM can also be used to deal with multilevel data (Goldstein & McDonald, 1988, Muthén, 1989; Muthén & Satorra, 1989; Longford & Muthén, 1992). Current SEM software such as LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), and AMOS (Arbucle, 1999), M-Plus (Muthén & Muthén, 1998-2004) can be adapted to implement multilevel SEM.

Bollen (1989) and Hox (2002) have pointed out that HLM can be theoretically specified as a special case of the more general structural equation model in that it is linear model whose covariance structure can be written in finite form. That being said, however, in practice many HLM designs have very complex representations in SEM form. Hox (2002) noted that for certain designs, such as latent growth curve analysis assuming large N, both approaches produce identical results. Despite the extensive previous research multilevel analysis, there has been very limited research comparing HLM and multilevel SEM in terms of their ability of handling multilevel data.

Since multivariate normality is a standard assumption for multivariate analysis, it also holds for HLM and SEM. Researchers should test their data to see if they meet the assumption prior to data analysis. However, real world research data can never be expected to meet all the statistical assumptions, especially normality. Ignoring nonnormality and treating the data as normally distributed will often result in inaccurate parameter estimates and misleading data interpretation. Thus, it is important to examine the robustness of HLM and SEM results treating nonnormal data under different sample size conditions.

In addition to nonnormality, missing data is another commonly encountered problem in data analysis, especially with large sample sizes where the possibility of non-response and/or subject attrition gets higher. Rubin (1976) developed a typology for missing data, which classified three distinct types of missing data: the data of one variable is considered to be missing at random (MAR) if the missingness of this variable only depends on the observed data of other variables in the dataset but not on the values of the variable itself. On the other hand, missing not at random is defined so that the missingness depends on not only on the observed data, but more importantly, on the missing data as well. A special case of MAR is missing completely at random (MCAR) which requires that the missingness is independent of the observed or missing data in the dataset. Traditional solutions such as mean substitution and listwise deletion have been found to be highly flawed and inefficient, as they reduce the data variability and alter the original covariance structure to a great extent. Little and Rubin (1987) showed that theoretically speaking, only data missing at random (MAR) including missing completely at random (MCAR) could be mathematically estimated, which is not always the case for real world data.

Currently, the preferred approach in missing data area is imputation using the expectation-maximization (EM) algorithm (Little & Rubin, 1990; Rubin, 1991; Schafer, 1997) within the ML framework. This method estimates a plausible value for each of the missing cells to mimic the original data covariance matrix as closely as possible. However, the underlying assumption of multivariate normality may limit its applicability (Graham, Hofer, & McKinnon, 1996). Gold, Bentler, & Kim (2003) demonstrated that a modified pairwise deletion method (asymptotically distribution-free available-case method) performs as well as the EM algorithm for estimating standard errors with large sample size, and yielded mixed results with respect to parameter estimates.

Sample size is a crucial factor in multilevel analysis due to statistical power. Because variances at all the levels in the data will be analyzed simultaneously, the multilevel analysis generally requires larger sample sizes at higher level than other multivariate procedures. A rule of thumb suggested in the literature (eg, Hox, 2002; Roberts, 2003) is at least 100 at the second or cluster level while the size at the first level is less important and may be small. Sample size investigation under SEM (eg., Arminger & Rothe, 1993; Fan & Wang, 1998; Finch, West, & MacKinnon, 1997) yields similar results in that under nonnormal conditions larger second level sample sizes will produce more accurate parameter estimates than small samples.

Significance of the Study

As discussed previously, multilevel analysis using HLM has drawn increasing attention in educational and behavioral sciences. In spite of its prominence, unlike SEM, there has not been much research addressing the robustness of this modeling method to data violating statistical assumptions such as nonnormal and incomplete data under different sample sizes. Furthermore, few of these SEM studies have been done in a multilevel framework. As researchers are finding more and more areas common to both HLM and multilevel SEM, it will be informative for guidelines to be provided for empirical researchers in their explorations with respect to how their models may perform with real world data, as what has been done in general SEM.

The author has investigated the influence of sample sizes at different levels and the correlation between the first level beta weights of the main effect and the group means on the statistical power of three models dealing with multilevel data: HLM, deviation SEM, and a hybrid approach of HLM and SEM with a simple two-level model (Willson & Zhang, 2003; Zhang & Willson, 2004). The results showed that HLM has poor power with small sample sizes compared to the other two approaches, but the power of all three models increased with higher correlations between the first level beta weights of the main effect and the group means. This difference vanished with group size larger than 35. The performances of the deviation SEM and the hybrid approach of HLM and SEM were very similar to each other. The current proposed study will extend the above studies, investigating the difference between these two multilevel modeling methods from the perspectives of nonnormality and incomplete data.

Purpose and Research Questions

The purpose of the study is to compare HLM and multilevel SEM in terms of their statistical power and accuracy of parameter estimates with nonnormal incomplete data under various sample sizes. Due to the number of factors examined, balanced sample size will be maintained. The specific research questions to be addressed in this study include:

1. How do HLM and multilevel SEM differ in terms of statistical power under nonnormal data with or without MAR missing data?
2. Do the parameter estimates differ by the design factors (analysis methods, nonnormality conditions, sample sizes, and proportions of missing data)?

3. Are the parameter estimates biased from the two methods under nonnormal MAR incomplete data treated by the EM algorithm multiple imputation with respect to the population parameters? Which method generates more estimate bias?
4. Does the percentage of incomplete data affect the functioning of the two methods under the EM algorithm treatment for missing data?
5. Given nonnormal incomplete data, which method is more robust under small sample sizes?
6. With the two-level model, which level's sample size has a more important role in ensuring adequate statistical power and accurate parameter estimates?

Limitations of the Study

Given the research design, the current study has several limitations. As will be discussed in detail in Chapter Three, the present study investigated only limited combinations of nonnormal and missing data percentage conditions, which might not represent the most frequently encountered situations in practice. Only a low to moderate intraclass correlation level (.15 to .30) was studied, which left opportunities for future research data with different nesting effect. In addition, the four sample size designs examined are all balanced and fairly large, thus adequate for multilevel analysis. This produces results not very applicable to small sample cases. However, it should be noted that small sample sizes may be problematic due to statistical power concerns. Nonnormality with missing data will only make that situation much worse, which is why such cases were omitted in the current study. Finally, the two-level model used in the Monte Carlo study is quite simple, whereas multilevel SEM could be used for much more complex models. So the results should be taken cautiously when looking at more complex SEM models.

CHAPTER II

REVIEW OF THE LITERATURE

This literature review provides information about studies related to multilevel modeling of nonnormal incomplete data. For that purpose, this chapter is divided into four distinct sections: a brief and general introduction of multilevel modeling using hierarchical linear modeling (HLM) and structural equation modeling (SEM); a review of methodological and empirical research studies investigating violation of univariate or multivariate normality; a brief history of missing data treatment and review of recent methodological research in the field; and a comprehensive review of research studies dealing with nonnormal and/or missing data in a multilevel context. All four sections are intended to give the readers an overview of the research problem and lead them through the findings of previous literature that significantly influenced the hypotheses of the current study.

Multilevel Modeling Using HLM and SEM

Multilevel data, also called hierarchical data, refers to data that inherently contain some hierarchical or nesting structure. Such data are prevalent in education and/or social science research where they are usually collected from more than one level of research unit, such as students clustered within classrooms or schools. As a result, there are several units of analysis in such data, and units at all the levels must be taken into account in data analysis. Traditional statistical procedures typically disaggregate higher-level variables to individual level and treat them as the individual variables, or vice versa (aggregating lower level variables to higher level). By doing this some important information is omitted from data analysis, such as conditional independence of response or within-cluster information. Goldstein (1995) pointed out that applying conventional regression models to multilevel data will results in potentially biased estimates, smaller standard errors, and inflated Type I error rates. Julian (2001) also found that ignoring the nesting relationship between the units of analysis at different levels and treating them as independently observed will lead to violations of statistical assumptions such as homoscadaticity in some designs and may potentially produce significantly biased parameter estimates.

The unique challenges presented by multilevel data include but are not limited to: within-cluster dependencies, homogeneity and with-cluster covariation, and sources of variation within and across clusters predicted from sampling theory. Certain statistical techniques have been developed to address these challenges, among which HLM and SEM are the most widely accepted and applied methods for multilevel data. They can be employed to answer similar research questions while each of them has individual areas of specialization that the other does not address adequately.

Hierarchical linear models (HLM), also known as random coefficient models, have been discussed for more than three decades in social sciences. HLM has been developed almost simultaneously under different names in different fields: it was called “multilevel linear models” (eg, Goldstein, 1987; Mason et al., 1983) in sociology; “mixed effects or random effects model” (eg, Laird & Ware, 1982) in biometrics; “random coefficient regression models” (eg, Rosenberg, 1973) in econometrics; and "covariance components models" (eg, Dempster, Rubin, & Tsutakawa, 1981) in statistics. The title "hierarchical linear models" was created firstly by Lindley and Smith (1972) to encompass all of the above. Its development encountered serious difficulty due to the limitation of computational capacity. The breakthrough appeared when the EM algorithm using maximum likelihood estimation was developed by Dempster, Laird, and Rubin (1977), which provided a "conceptual feasible and broadly applicable approach to covariance component estimation" (Raudenbush & Bryk, 2002). HLM applications began in the early 80s but have really flourished during the past ten or fifteen years.

Hierarchical linear models can be applied to a wide variety of settings involving hierarchical structure. These settings include but are not limited to contextually nested data, temporally or longitudinally nested data (repeated measures or latent growth curve), cross-sectional data and even cross-classified hierarchical data. HLM has apparent advantages over previously developed statistical procedures for multilevel structure. Estimation of variance is greatly improved within individual units by including estimates from other units or levels. One of HLM's most important features is its capacity to study cross-level effects, which usually focus on the contextual effect of a higher-level variable on lower level effects. It is particularly attractive because in social science research the primary effect of interest is not always statistically significant even though it may interact with certain contextual characteristics (in longitudinal data the subjects' personal characteristics). Researchers are sometimes more interested in the interaction effect when they intend to develop treatment differentially effective for specific groups. In addition, HLM can partition variance at different levels, thus giving appropriate parameter estimates and standard errors with separate within-subject and between-subject covariance matrices.

Structural equation modeling (SEM) being in the general framework of multivariate statistics is another powerful and comprehensive modeling technique that encompasses many traditional statistical procedures from simple regression and path analysis to complex discriminant analysis and canonical correlation function (Bollen, 1989). SEM is primarily aimed at studying the relationships among sets of variables, which can be either observed or unobserved (latent). It is used as a confirmatory more than exploratory modeling method, and thus allows researchers to test their hypothesized models and modify them subsequently according to their theory and sample-based evidence.

A typical structural equation model containing both observed and latent variables can be separated into a measurement model and a structural model so that measurement errors are easily incorporated into data analyses. The model specification is fairly flexible in that the researcher is entitled to decide which paths to fix or free as well as specific variance values. SEM also provides information about the degree to which the hypothesized model fits the observed data by comparing the hypothesized and observed variance-covariance matrices. As a confirmatory technique, SEM requires a substantive theory underlying the hypothesized model and a representative sample for data analysis. When the model fit is not satisfactory, theoretical justifications are needed to revise the model in addition to the mere statistical modification indices (Mueller, 1997).

During the past ten years, application of SEM has undergone an expansion in almost every area of social and behavioral research due to the rapid improvement in major SEM software such as LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), and AMOS (Arbucle, 1999), M-Plus (Muthén & Muthén, 1998-2004).

In addition to the general simple or complex path analysis, SEM can be used to model longitudinal and hierarchical data. Specifically, latent growth modeling, a generalized model to investigate change and development (Meredith & Tisak, 1990), targets longitudinal or time change data. For hierarchical data, on the other hand, SEM differentiates between the within-subject matrices. SEM fits a separate within and between-subject models to jointly estimate the parameters (Muthén & Satorra, 1989). According to Bollen (1989) and Hox (2002), HLM and SEM have a close connection in a sense that HLM can be theoretically specified as a special case of the more general structural equation model as a linear model whose covariance structure can be written in finite form. Despite the massive amount of research that has been conducted on each of them, there is a need in the literature to compare HLM and SEM in terms of their capacities to address the same research questions concerning multilevel data.

Sample size has always been an issue in experiment design and data analysis due to its influence on statistical power. This is particularly an issue in multilevel analysis using HLM or SEM because of the computational complexity involved in these analyses, greater than for ordinary multivariate statistical procedures. Large sample size is an appealing feature of a study in that analyses typically have greater power to detect most effects, and the results are more readily generalized to a population. Some researchers investigating the influence of nonnormality also have found that in SEM large sample size tends to offset some bias in parameter estimates and standard errors.

Previous literature has suggested some rules of thumb for conducting multilevel analysis (e.g., Hox, 2002; Roberts, 2003) to require at least 100 groups in the sample, while the group size was less important. However, in the simulation studies (Willson & Zhang, 2003; Zhang & Willson, 2004) the author has conducted, the first level sample size did matter. In a two-level cross-interaction model, it was found that the power did not increase further when the group size exceeded 35, when the number of groups was fixed at 120. Further research is needed to understand these inconsistent results.

Violation of Normality Assumption in Data Analysis

Normality is a common assumption for data analysis with continuous variables. The data are expected to follow a normal distribution so that certain procedures applied to them result in mathematically tractable parameter estimates and statistical indices. Normal theory maximum likelihood (ML) and Generalized Least Squares (GLS) are two estimation methods that are most widely used in HLM and SEM. The assumption of multivariate normality must be met for these techniques to yield interpretable computational results. However, real world data are almost never expected to meet all statistical assumptions, especially normality. As a somewhat common practice, approaches like HLM and SEM have been applied to nonnormal data (Bentler, 1994; Bentler & Dudgeon, 1996; Micceri, 1989, Brown, 1990).

Normality can be assessed in both univariate and multivariate frameworks. Univariate normality is explicitly defined or measured by the skewness and kurtosis of a variable's distribution. Compared with this definition, multivariate normality is more difficult to assess. Nevertheless, in multivariate statistics like SEM, the multivariate normality assumption will not hold if any of the analysis variables violates univariate normality (Bollen, 1989).

In terms of univariate normality, in a strict sense, a variable is considered to be normally distributed if both the skewness and kurtosis of its distribution are zero. In practice, the influence of nonnormality on results depends on the degree of nonnormality. There has not been any consensus about the specific range of these two normality indices to be regarded as nonnormal but ignorable in data analysis. Muthén and Kaplan (1985) argued that any nonnormal conditions with the absolute values of skewness and kurtosis less than 1 would not cause significant distortion in computational results. Some Monte Carlo studies (e.g., Hu, Bentler, & Kano, 1992) studied simulated extremely nonnormal data with kurtosis as high as 20 even such situations are not likely to be common in actual data. Certain transformations can reduce the degree of skewness and kurtosis, although skewness is easier to correct than kurtosis. There has been agreement that data with both nonzero skewness and kurtosis tend to produce more bias in results than those having only one violation (Chou, Bentler, & Satorra, 1991; Muthén & Kaplan, 1985).

Since SEM and HLM are commonly applied to nonnormal data, extensive efforts in research have been devoted to evaluate estimation bias in parameters and standard errors and to develop estimation procedures and test statistics robust to nonnormality. Some studies (e.g., Muthén, 1989; Gold, Bentler, & Kim, 2003; Curran, West, & Finch

1996) examined the influence of nonnormality on parameter estimates, standard errors and fit indices in the framework of SEM using Monte Carlo method, while including a few empirical studies. No such work has been done in HLM as the author can find.

The general approach in SEM simulation was to design data matrices in which the variables exhibited various degrees of skewness and/or kurtosis. This design represents real world situations where not all the variables of interest have identical distributions. Other methods simulating nonnormal data include initiating data generation with chi-square and exponential distributions (Arminger & Rothe, 1993), and specifying multivariate skewness and kurtosis (Muthén & Kaplan, 1985, Finch, 1993). The above studies consistently found that nonnormality had more influence on standard errors than on parameter estimates when ML estimation was used, especially for large sample size. Moderate nonnormality (the absolute values of skewness and/or kurtosis between 1.0 and 2.3) did not bias the standard errors of ML and Generalized Least Squares (GLS) parameter estimates (Fan & Wang, 1998).

Using real nonnormal data and varying sample sizes to fit a SEM model with both manifest and latent variables, Wang, Fan, and Willson (1996) also found that ML and GLS generated consistent and almost identical estimators, and standard errors tended to be underestimated, but the problem was not serious in large samples. They suggested that the chi-square test statistics were acceptable given nonnormal data but appropriate sample sizes. Other fit indices other than chi-square tests have not been studies adequately for nonnormal data. So Bentler and Bonnett (1980) and Bollen (1989) suggested than chi-square test statistics should always be reported no matter which other fit indices were selected to use in the studies, even though they did not intend to focus on chi-square statistics as the primary or sole criterion for model fit.

Methodological Research on Missing Data Treatment

Missing data has become an important issue in empirical studies. Lack of responses occurs due to many and various causes, especially in projects with large sample sizes and in longitudinal research studies where the probability of subject attrition increases. Missing data is a nuisance for data analysis because it may bias parameter estimates and standard errors and inflate Type I and II error rates thus reduce statistical power of the analyses (Brockmeier, Kromrey, & Hogarty, 2003). Concern over possible distorted results of data analysis with missing data has led to extensive research.

The traditional approaches to handle missing data included listwise deletion, pairwise deletion, and mean substitution, widely employed by empirical researchers (Tanguma, 2000). Listwise deletion, being the most commonly used method and default in popular software packages such as SPSS and SAS, deals with missing data in an intuitive way by dropping all cases with missing values. It reduces data variability and statistical power by excluding cases. Pairwise deletion, on the other hand, uses more information than listwise deletion by including cases that provide complete data for certain analyses. Concern for statistical power is slightly alleviated, but results are not comparable across different analyses based on different samples. Also, it can result in non-positive definite covariance matrices and resultant convergence of estimation under OLS. Mean substitution was regarded as conservative but neat solution for missing data.

Its major problem lies in reduction in data variability and biased results because of unreliable inference on missing values (Kromrey & Hines, 1994).

Historically, the framework of missing data inference was developed by Rubin (1976) and is still in use by researchers today. A missing data mechanism terminology is known widely but poorly understood. According to Rubin (1976), missing data can be classified into three distinct categories: the values of one variable are considered to be missing at random (MAR) if its own values do not contribute to the phenomenon of missingness. Otherwise, the situation is missing not at random (MNAR). Under MAR, causes of missingness for certain variable involve values of other variables within the dataset whether or not they are complete or include missingness. One special case of MAR is missing completely at random (MCAR) in which the missingness distribution in the sample is fundamentally random as it does not relate to any variable in the dataset.

Currently the preferred method of missing data treatment is multiple imputation (MI) using the expectation-maximization (EM) algorithm (Little & Rubin, 1990; Rubin, 1991; Schafer, 1997) within the maximum likelihood (ML) framework. The EM algorithm was invented by Dempster, Laird, & Rubin (1977) and it allowed the computation of ML estimates for missing cells. The basic idea of multiple imputation developed by Rubin (1987) was to treat missing data as random variables and replaced the missing cells with more than one simulated value. The goal was to create simulated values for the missing cells so that the complete data would replicate the original variance-covariance matrix.

MI is superior to mean imputation or regression imputation by retaining the advantage of conditional distribution but improving estimation by taking into account the missingness uncertainty. Complete datasets generated by MI are analyzed with familiar complete-data methods and software. By producing one set of plausible complete data versions and analyzing them consistently, the MI procedure obtains parameter estimates and standard errors from the combined datasets, which reflected “missing data uncertainty as well as finite sample variation” (Schafer & Graham, 2002). Given the computational complexity, MI can easily be conducted in freeware programs like NORM (Schafer, 1999) or well developed commercial software like SAS PROC MI and PROC MIANALYZE (SAS Institute, 2002).

Even though MI with EM algorithm has proven to be effective and efficient in dealing with missing data (e.g., Bunting, Adamson, & Mulhall, 2002), it requires the assumptions of multivariate normality and MAR to be met, which hinder its further application. However, these two assumptions are seldom tested. This holds for normality because it is usually overlooked as in data analyses while results can be quite robust. In terms of MAR, very few procedures were proposed to test it (e.g., Cohen & Cohen, 1983; Tabachinick & Fidell, 1996). In empirical studies, it was rarely tested like normality. Applied researchers are pressed to have some guidance for nonrandomly missing data.

Research on Incomplete Data under Special Distributions

Being such a commonly encountered nuisance in behavioral science research, missing data has been extensively explored in regard to its distortion of data analysis results and optimal treatment under different situations. There has been some general agreement on missing data treatment. Deletion methods including both listwise and pairwise tend to perform well when estimating regression weights with large sample size and low percentage of missing data (Kim & Curry, 1977; Roth & Switzer, 1995), while imputation procedures (both simple and multiple) will produce better parameter estimates with comparably smaller sample size and higher percentages of missing data (Basilevsky et al., 1985; Raymond & Roberts, 1987; Roth & Switzer, 1995).

Due to its prevalence, missing data may occur under some special conditions such as nonnormality or hierarchical data that further complicate data analysis and may invalidate the general rules for missing data procedures. Even though much less research has been conducted on these conditions, a few studies (reviewed by Gold & Bentler, 2000) explored treatments of nonnormal incomplete data, and one study (Gibson & Olejnik, 2003) examined the influence of missing data in a hierarchical data context. Both produced insightful results and provided some guidance for applied researchers.

Graham et al (1996) studied the variances and covariances of simulated univariately incomplete nonnormal data in a SEM model comparing expectation-maximization (EM) and available-case listwise deletion methods. The skewness of their data ranged from -.84 to 3.16 and kurtosis from -.04 to 11.97. EM appeared to be superior to available-case listwise deletion method for estimating root mean square residuals and the average standard errors. Nonnormlity was proved to result in overestimation of the covariance matrix and reduce estimation efficiency compared with the results of normal data. Gold and Bentler (2000) extended the above study by comparing four different missing data treatments including structured-model EM and saturated-model EM with nonnormal MCAR data in a large SEM model (a four-factor ten-variable path model). EM methods performed better than the other two regardless of the percentage of missing data. Both MAR and MCAR data were simulated in Enders (2001) and estimated by ML and traditional approaches like listwise, pairwise deletion, and mean substitution. Under MCAR, very little bias was found in all methods except for mean substitution. ML showed improved efficiency with higher proportions of missing data. It also produced the least parameter bias under MAR condition. Different missing data methods exhibited little effect on standard errors thus resulting in fairly reliable confidence interval under all procedures.

As HLM has become more popular and accessible to applied researchers in various fields, the topic of missing data with hierarchical data structure has not been extensively explored as much. The only recent study found here was conducted by Gibson and Olejnik (2003). This simulation study compared five missing data techniques in the context of missing data at the second level of a two-level hierarchical linear model. The factors being examined were number of level-2 variables, level-2 sample size, level-1 intercept-slope correlation, and percentage of missing data. Listwise deletion and EM generated satisfactory random effect estimates with level-2 sample size at 30, which was their smaller sample condition with a missing data proportion as high as 40%.

Based on the above literature review, it appeared that violation of normality and missing data treatment have been topics extensively investigated in single-level data analysis, especially in SEM. The sample size issue, due to its importance in determining the power of multilevel analysis, also attracted considerable amount of attention in the research community, mainly at the sample size of first level. However, there has not been any study, as to the author's awareness, that comprehensively studied the overall effects of the three factors, violation of normality, missing data treatment, and sample size design, in the context of HLM. The current study hopes to provide some guidelines for empirical researchers on these aspects. Additionally, since HLM and SEM can be employed to address similar research questions, a study comparing them would be informative to future researchers who will need to decide which one to follow. This is another practical goal of the present investigation.

CHAPTER III

METHODOLOGY

The current research was a Monte Carlo study designed to simulate the data for multilevel analyses in the presence of missing data treated by multiple imputation using expectation-maximization-based (EM) maximum likelihood (ML) estimates. A simple two-level model with cross-level interaction was simulated and analyzed by HLM and multilevel SEM. The design of the data matrix was intended to mirror a real world project in which students were nested in classrooms. Special issues were explored, such as (1) parameter estimate robustness to nonnormal data; (2) their robustness to incomplete data treated by EM algorithm; (3) their sensitivity to sample sizes at different levels for statistical power and (4) possible interaction effect among three factors of nonnormality, missing data, and sample size on the performance of HLM and multilevel SEM. The purpose was to provide some guidelines for empirical researchers doing hypothesis testing with data in similar conditions.

Data Generation

Design of the Simulated Model

A data matrix containing three variables *x1 - x3* was generated using SAS IML (SAS Institute, Inc. 2002). For easy interpretation, all three variables had a mean of 0 and standard deviation of 1. The parameters for the correlation coefficients among all three were set to be equal at .3. This value was chosen to represent a moderate effect size that might be reasonably detected in behavioral science research. Since nonnormality situations were investigated in the current study that may cause the sample correlations to deviate from the parameter value, an independent analysis of the three-variable raw data matrix generation was conducted. This separate simulation indicated that the correlation parameter value of .3 was achieved for all levels of data nonnormality, with a deviation always less than 8% for sample size of 100, 000. The generated complete raw data matrix was then sorted by the value of x2.

The two-level clustered data structure was constructed from the sorted data matrix. One individual predictor x1 and one cluster-level predictor, which was the “classroom” mean of x2, namely *x2bar*, together with the cross-level interaction between them *x1*x2bar* were utilized to predict the individual level outcome *x3*. After being sorted by x2 the scores were divided into classrooms for a given classroom size. The classroom means of x2 were then produced from the individual vluess of x2 in the classrooms to serve as the classroom level or level-2 predictor.

Since the dataset was sorted by x2 previously and then divided into classrooms, the classroom means of x2 were automatically sorted as well by classroom id. The goal was to create a given degree of cross-level interaction. Even though the data were sorted by x2, the individual distributions of x1 and x3 within different classrooms still overlapped since their correlations with x2 were only about .3.

Design of the Monte Carlo Study

The Monte Carlo study involved three independent factors that combined to produce 3 x 4 x 2 = 24 cells in a balanced design based on the three data features and hierarchical model structure. The independent factors were

1. Nonnormality conditions.

Data normality conditions were systematically explored for the first level independent variable pooled across all the observations and clusters. Three configurations were specified: normality, moderate nonnormality and severe nonnormality. These configurations were achieved by different combinations of skewness and kurtosis following the Fleishman's power transformation method (Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S., 2003). The criteria discussed for skewness and kurtosis were based on absolute values. Only positive values were simulated in data generation since it is the magnitude that makes a difference, not the direction, although it is realized that the correlation between variables of opposite skewness values would deviate much more from the parameter correlation value. Normality conditions required both skewness and kurtosis equal to 0. The functioning of HLM and multilevel SEM with the normal data served as the benchmark model with respect to their performance under the other two nonnormal conditions. Moderate nonnormality was defined for skewness of .5 and kurtosis of 1.0. For skewness of 1.0 and the kurtosis of 3.75, the data were defined to be severely nonnormal.

2. Percentage of incomplete data.

The incomplete data conditions were generated under the missing at random (MAR) assumption (Little & Rubin, 1989). Data were then imputed using the EM algorithm (Schaffer, 1997). Two levels of percentage of incomplete data were specified as suggested by Gold, Bentler, and Kim (2003): .15, and .30. The author considered these two levels to be typical for the low and moderate levels of incomplete data. They suggested that multilevel technique would be inappropriate if the percentage of incompleteness exceeded .40 in the first level variable. It should be noted that both levels of missingess were created based on the same original complete data matrix to avoid potential influence of random variation.

3. Sample size.

Two patterns of sample sizes were investigated in this study: one fixed the cluster size and varied the number of clusters while the other fixed the number of clusters and varied the cluster size. The reason was to evaluate which level sample size was more crucial in determining the models' statistical power. Hox (2002) and Roberts (2003) suggested the second-level sample size, which is the number of clusters, was more important than the clusters' sizes. In previous simulations by the author (Zhang & Willson, 2004), with an adequate number of clusters, the cluster size can also make a difference and the larger first level size did not always improve power. Cluster size of 30 to 35 was found to be adequate to produce stable variance-covariance matrices within clusters. The current study was expected to add to the literature by clarifying the influences of sample sizes at different levels on statistical power. Under each pattern, two situations were examined respectively, which resulted in four sample size conditions: 50 clusters with 10 and 50 observations in each cluster and 30 observations in each cluster with 10 and 50 clusters, resulting in 500, 2,500, 300, and 1,500 data points. Given these sample conditions, the intraclass correlations (ICC) for the dependent variable x3 among different sample sizes were randomly varied from .15 to .3 with generally higher ICC in sample sizes with more second-level units.

In summary, the three factors created 12 (3 x 4) original complete data matrices and another 24 (3 x 4 x 2) with incomplete data. The MAR data condition was manipulated with PROC SURVEYSELECT. After the complete data were generated, two independent procedures of PROC SURVEY were conducted to select a sample from the complete dataset in which x1 and x2 were individually set to be missing with certain criteria. MAR missingness was produced by deleting x1 and x3 for all the cases above the 25th percentile of x2 with the probability of 8% and 15% to yield an expected 15% and 30% percent of missing data. PROC SURVEYSELECT randomly selected certain data points and deleted without replacement from the complete data with the probability of 8% for x1 and x3 separately thus yielding approximately 15% missingness and similarly probability of 15% to yield 30% missingness. This missing data sample was then merged back into the original complete dataset to create the MAR condition.

Data Analysis

One hundred replications were conducted for each data matrix, with a total of 36 cases in simulation. After the data with missing points were generated, they were first analyzed by PROC MI with the EM algorithm, which produced five imputed datasets. These datasets were then analyzed by HLM in PROC MIXED (Singer, 1998) and multilevel SEM in PROC CALIS. There were three regression coefficients (x1, x2bar, and x1*x2bar) estimated in the HLM while the classroom mean estimate (x2bar) was omitted in the multilevel SEM due to the within-classroom standardization procedure in the first step. For easy comparison and interpretation, only the main and cross-level interaction effects were considered. Since both the missingness proportions were generated based on the same complete data and all data were then analyzed in the same way, all the data generation and analyses were accomplished with only 12 (three nonnormality conditions by four sample size designs) sets of syntaxes. The syntax is found in Appendix 1 for moderately nonnormal condition with a sample size of 10 clusters with 30 subjects within each cluster.

The model representation of HLM is specified as the following:

$$x3_{ij} = ß_{0j} + ß_{1j} * x1\_c_{ij} + e_{ij}$$

$$ß_{0j} = ?_{00} + ?_{01} * x2bar_{j} + u_{0j}$$

$$ß_{1j} = ?_{10} + ?_{11} * X2bar_{j} + u_{1j}$$

For multilevel SEM, both the dependent and indepent variables (x1 and x3) were standardized within each classroom to get their standard scores. Then the following model was specified:

$$zx3_{ij} = ?_{1}*zx1_{ij} + eij$$

$$?_{1} = a_{1} + ?_{1}*x2bar_{j} + ?_{1j}$$

The means and standard deviations for all the regression coefficients estimates and their standard errors from all five imputed datasets in each modeling method were saved. The complete data samples were also analyzed with the same method to serve as the known parameter values for bias evaluation. Three outcomes are examined: parameter estimates including standard errors, parameter estimation bias including standard error bias, and empirical statistical power. Analysis of variance was employed to evaluate the effects of design factors (nonnormality condition, analysis method, missingness, and sample size) on the parameter estimates and related bias.

Research Question #1 addresses the issue of statistical power difference between HLM and multilevel SEM under nonnormal data with or without missing. If missing data were present, they would be treated by multiple imputation procedure with the EM algorithm. This question was answered by computing and comparing the power from the two methods across various designs. The t-statistics, standard errors, and probabilities associated with t-statistics resulted for all the regression coefficients from analysis of the complete and imputed data were output to a data file. Counts of significance for p=. 05 were created and compared.

Research Question #2 addresses the issue of potential differences in parameter estimates by the design factors (analysis methods, nonnormality conditions, sample sizes, and proportions of missing data). For complete data analysis of variance was run to examine the main effect of the three design factors except for the proportions of misisngness. Furthermore, analysis of variance with repeated measures was conducted for the imputed data on the all four factors with missing data proportion being the within-subject factor with three levels (0, 15% and 30%) and the other three being the between-subject factors. The effect of proportions of missing data (Research Question #4) was also investigated through this procedure.

Research Question #3 addresses the issue of estimation bias resulted from running the EM algorithm multiple imputation on the incomplete data. The values from analyzing the complete data were taken as the population parameters. Following Gold, Bentler, and Kim (2003), the bias of parameter estimates was evaluated by the root mean square of the difference (RMSD) between the mean parameter estimate from imputed datasets and the corresponding known parameter value divided by the absolute value of the known parameter. The estimates from analysis of the complete data were the known parameter values. This bias index was computed for both the regression coefficients and standard error estimates and averaged across 100 iterations. Smaller bias implied a more robust modeling method. The same approach was also applied to the standard errors of parameter estimates. The biases were compared across the three design factors in ANOVA. Post-hoc comparisons were conducted to determine which design combinations produced less bias.

Conclusions were based on the results of the above three research questions about potentially different functioning of the two methods under varying nonnormality. The goal was to find a method that works well under severely nonnormal data. The performance of the analysis methods was evaluated by comparing statistical power and estimation bias under missingness with multiple imputation. The answers to Research Question #5 and #6 were based on similar inferences.

Summary of General Procedures for Data Analysis

In sum, for designing situations involving varying normality degrees, percentages of missingness, and sample sizes, the following steps were taken:

1. Complete data were analyzed by HLM and multilevel SEM with the regression coefficients and standard errors output to a data file.
2. Systematic missing data were generated from the complete data to meet MAR assumption with different percentages of missingness.

3. The data with missingness were analyzed by multiple imputation through the EM algorithm in PROC MI resulting five imputed datasets.
4. The imputed data were analyzed by HLM and multilevel SEM with the regression coefficients and standard errors output to a data file. The means across the five imputed datasets were taken as the estimates for the imputed data.
5. Empirical statistical power was computed for all the regression coefficients and compared.
6. Bias was evaluated by RMSD for parameters and standard errors.
7. A number of separate sets of ANOVA were conducted to examine the differences for parameter estimates, standard error estimates, and their estimation bias.

CHAPTER IV

RESULTS

The present study was intended to compare HLM and multilevel SEM under nonnormal data with or without missing data under multiple imputation across different sample sizes. Two types of analysis of variance (ANOVA) used to evaluate the results of simulated data: simple factorial ANOVA was employed for results of complete data while ANOVA with repeated measures was used for missing data. To simplify presentation and interpretation of the results among the four design factors (three for complete data evaluation), only their main effects and two-way interactions were reported even though the full factorial model was studied. These effects were believed to be most relevant to the purpose of the current study. Due to the model difference between HLM and SEM in multilevel analyses, in the simulated two-level model only the main and interaction effects were discussed, as the unit mean effect was not present in the multilevel SEM model.

Empirical Statistical Power of HLM and Multilevel SEM

The empirical power for the main and interaction was computed by counting the number of significant findings during the 100 iterations. The empirical statistical power for both the methods in various designs is presented in Table 1.

Table 1

Empirical Statistical Power for the Simulated Two-level Model

	Complete Data				Imputed Data with 15% missing				Imputed Data with 30% missing			
	HLM		*SEM*		*HLM*		*SEM*		*HLM*		*SEM*	
Normal	Main	Interaction	Main	Interaction	Main	Interaction	Main	Interaction	Main	Interaction	Main	Interaction
N=300 (30 x 10)	0.84	0.07	0.85	0.06	0.80	0.03	0.80	0.01	0.85	0.00	0.85	0.01
N=500 (10 x 50)	0.58	0.03	0.53	0.06	0.67	0.01	0.65	0.02	0.77	0.01	0.73	0.00
N=1500 (30 x 50)	1.00	0.05	1.00	0.02	1.00	0.05	1.00	0.00	1.00	0.04	1.00	0.01
N=2500 (50 x 50)	1.00	0.06	1.00	0.06	1.00	0.06	1.00	0.09	1.00	0.06	1.00	0.04
Moderately nonnormal												
N=300 (30 x 10)	1.00	0.06	1.00	0.00	1.00	0.02	1.00	0.00	1.00	1.00	1.00	0.00
N=500 (10 x 50)	0.60	0.07	0.58	0.04	0.61	0.05	0.60	0.03	0.68	0.03	0.64	0.00
N=1500 (30 x 50)	0.68	0.06	0.69	0.07	0.65	0.05	0.62	0.06	0.67	0.04	0.66	0.03
N=2500 (50 x 50)	1.00	0.21	1.00	0.02	1.00	0.20	1.00	0.04	1.00	0.19	1.00	0.04
Severely nonnormal												
N=300 (30 x 10)	0.90	0.12	0.90	0.00	0.80	0.03	0.81	0.02	0.90	0.00	0.91	0.00
N=500 (10 x 50)	1.00	0.16	1.00	0.02	1.00	0.09	1.00	0.03	1.00	0.05	1.00	0.00
N=1500 (30 x 50)	0.83	0.11	0.84	0.09	0.82	0.10	0.81	0.05	0.80	0.05	0.79	0.02
N=2500 (50 x 50)	1.00	0.39	1.00	0.03	1.00	0.37	1.00	0.04	1.00	0.36	1.00	0.04

Power for Complete Data

For complete data, HLM exhibited the same statistical power as multilevel SEM for the main effect across different sample sizes and across nonnormality conditions. The normal and moderately nonnormal data for sample size of 500 (50 units with 10 subjects per unit) yielded much lower power (about .60) for the main effect than the other three sample designs (above .80). This difference diminished under severely nonnormal data. In terms of the cross-level interaction, the power of HLM and multilevel SEM under normal and moderately nonnormal data were quite similar. The only exception was the moderately nonnormal data with the largest sample (N = 2,500 at 50 units with 50 subjects per unit) where HLM worked better (about .20) than SEM with the same sample size (.02) and better than both methods for the other three samples (below .10). The power was even higher (about .40) under severely nonnormal data, for which HLM consistently performed better than multilevel SEM. The latter basically had no power to detect the interaction effect. Among the four sample sizes, the largest (N= 2500) produced greater power than the other three whose power was similar to each other. This is particularly distinct on the interaction effect under nonnormal data.

Power for Imputed Data

The power based on imputed data was quite comparable to that for complete data in terms of both the main and cross-level interaction for both HLM and multilevel SEM. Similar to the power for complete data, imputed data condition resulted in excellent power to detect the main effect (above .80) but very poor power for the cross-level interaction effect (below .10). However, for normal data at the sample size of 500, the imputed data matrices produced higher power (about 20%) for the main effect than the original complete data despite of the proportion of missing data. Furthermore, under severely nonnormality the imputed data with 15% missingness at sample size of 300 yielded lower power (20%) than either the complete or incomplete data with 30% missingness. The power of the latter two data matrices was similar on both the main and interaction effects.

A full factorial model of analysis of variance with repeated measure for imputation was conducted to investigate the effects of four design factors on the statistical power of the main effect and interaction. The ANOVA tables are reported in Table 2 and 3. For the main effect of x1, among the within-subject effects the main effect of missing percentage and its two-way interaction with sample size and three-way interaction with sample size by normality condition were statistically significant beyond the .001 level. As for the between-subject effects, only the normality condition and sample size together with their interaction were significant (p < .001). The plots of estimated marginal means indicated that the sample size of 500 (50 clusters with 10 subjects per cluster) exhibited lower power than the other three sample sizes. Under severely nonnormal data the power of the main effect was abnormally high regardless of the missingness proportions. When testing for power for the cross-level interaction effect, all the effects in the model were significant (p < .001) except for the three-way interaction among missingness proportion, normality condition and analysis method.

Table 2

Analysis of Variance for Statistical Power of Main Effect

Source	df	F	partial ?^{2}	p
	Between subjects			
Normality (N)	2	268.604***	0.184	<.001
Sample (S)	3	900.238***	0.532	<.001
Analysis (A)	1	0.012	0	0.914
N x S	6	266.148***	0.402	<.001
N x A	2	0.04	0	0.961
S x A	3	0.008	0	0.999
N x S x A	6	0.032	0	1
between-subject error	2376	(.101)		
	Within subjects			
Missingness (M)	2	12.224***	0.005	<.001
M x N	4	2.122	0.002	0.075
M x S	6	6.803***	0.009	<.001
M x A	2	0.019	0	0.981
M x N x S	12	3.754***	0.009	<.001
M x N x A	4	0.196	0	0.94
M x S x A	6	0.258	0	0.956
M x N x S x A	12	0.074	0	1
M within-group error	4752	(.005)		

Note. The outcome has a grand mean of .856 and standard deviation of .324. Values enclosed in parentheses represent mean square errors.

*p < .05. **p < .01. ***p < .001.

Table 3

Analysis of Variance for Statistical Power of Interaction Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	3.901*	0.003	0.02
Sample (S)	3	8.259***	0.01	<.001
Analysis (A)	1	15.038***	0.006	<.001
N x S	6	14.218***	0.035	<.001
N x A	2	16.254***	0.013	<.001
S x A	3	7.959***	0.01	<.001
N x S x A	6	10.291***	0.025	<.001
between-subject error	2376	(.182)		
		Within subjects		
Missingness (M)	2	312.61***	0.116	<.001
M x N	4	17.764***	0.015	<.001
M x S	6	215.098***	0.214	<.001
M x A	2	36.07***	0.015	<.001
M x N x S	12	6.562***	0.016	<.001
M x N x A	4	0.73	0.001	0.571
M x S x A	6	15.845***	0.02	<.001
M x N x S x A	12	3.052***	0.008	<.001
M within-group error	4752	(.02)		

Note. The outcome has a grand mean of .505 and standard deviation of .292. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Parameter Estimates Differences

Two sets of analysis of variance were conducted with the parameter estimates and standard errors as the dependent variables. One was for the complete data and the other for the imputed data on the four design factors with the missingness proportion being the repeated measured within-subject factor (three for complete data with proportions of missingness absent).

ANOVA on Complete Data

With two analysis methods, three normality conditions, and four sample sizes, a three-factor 2 x 3 x 4 balanced ANOVA was used to examine the effects of these design factors on the parameter estimates and standard errors. All the main and two-way interaction effects appeared to be highly significant ($p < .001$). To follow up these results, the LSD means post-hoc comparison procedure was applied to the standard error estimates with regard to the main effect of individual factors. The standard errors differed by normality conditions, but the patterns were mixed for the main or cross-level interaction effects. This was not the case for sample size. Unlike the normality condition, larger samples produced smaller standard errors very consistently for both the effects. As for the analysis method, HLM had larger standard errors for the main effect and smaller estimates for the interaction effect than multilevel SEM.

ANOVA with Repeated Measures on Imputed Data

With the four design factors including the proportion of missingness, a balanced ANOVA with repeated measures on imputed data was explored for the parameter estimates and standard errors. The ANOVA tables of these analyses are reported in

Table 4 to Table 7. All the four design factors had significant effects for the regression coefficient of main effect of x1 in the simulated two-level model with the analysis method being the least differentiating factor (p < .05 while for the other three p < .001). All the interactions were significant at or beyond the .001 level except for two, which were the sample by method one and the one among normality, proportion of missingness, and sample size. For the cross-level interaction between x1*x2bar, all the effects were significant beyond .001 except for the main effect of normality condition (F = .452, p = .637) and the three-way interaction of normality condition by proportion of missingness by analysis method (F = 2.236, p = .069).

In terms of the standard error of the main effect of x1, all the effects were highly significant except for the two-way interaction between missingness proportion and normality condition (F = .377, p = .755). As for the standard error of the cross-level interaction, all the effects were highly significant (p < .001) except for the two-way interaction between analysis method by proportion of missingness (F = .893, p = .374). The post-hoc multiple comparison of standard error estimates yielded mixed results for normality conditions, but the standard errors were always smaller in larger samples for both effects. As in the complete data case, HLM produced larger standard errors on the main effect and smaller ones for the interaction effect than multilevel SEM. The higher proportion of missing data (30%) was found to create larger standard errors than the lower proportion (15%).

Table 4

Analysis of Variance for Main Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	78.594***	0.062	<.001
Sample (S)	3	755.086***	0.488	<.001
Analysis (A)	1	4.158*	0.002	0.042
N x S	6	716.782***	0.644	<.001
N x A	2	8.288***	0.007	<.001
S x A	3	1.103	0.001	0.347
N x S x A	6	2.522*	0.006	0.02
between-subject error	2376	(0.043)		
		Within subjects		
Missingness (M)	2	144..46***	0.057	<.001
M x N	4	14.829***	0.012	<.001
M x S	6	90.064***	0.102	<.001
M x A	2	11.635***	0.005	<.001
M x N x S	12	5.943***	0.015	<.001
M x N x A	4	0.822	0.001	0.511
M x S x A	6	10.682***	0.013	<.001
M x N x S x A	12	3.048***	0.008	<.001
M within-group error	4752	(0.001)		

Note. The outcome has a grand mean of .262 and standard deviation of .237. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Table 5

Analysis of Variance for Cross-level Interaction Effect

Source	df	F	partial ?^{2}	p
	Between subjects			
Normality (N)	2	0.452	0	0.637
Sample (S)	3	7.738***	0.01	<.001
Analysis (A)	1	22.429***	0.009	<.001
N x S	6	7.745***	0.019	<.001
N x A	2	10.762***	0.009	<.001
S x A	3	18.756***	0.023	<.001
N x S x A	6	4.493***	0.011	<.001
between-subject error	2376	(.002)		
	Within subjects			
Missingness (M)	2	327.784***	0.121	<.001
M x N	4	19.861***	0.016	<.001
M x S	6	253.664***	0.243	<.001
M x A	2	43.576***	0.018	<.001
M x N x S	12	8.031***	0.02	<.001
M x N x A	4	2.236	0.002	0.069
M x S x A	6	25.284***	0.031	<.001
M x N x S x A	12	4.173***	0.01	<.001
M within-group error	4752	(< .001)		

Note. The outcome has a grand mean of .001 and standard deviation of .034. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Table 6

Analysis of Variance for Standard Error of Main Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	86.2***	0.068	<.001
Sample (S)	3	8246.136***	0.912	<.001
Analysis (A)	1	130.641***	0.052	<.001
N x S	6	343.324***	0.464	<.001
N x A	2	53.561***	0.043	<.001
S x A	3	34.924***	0.042	<.001
N x S x A	6	76.544*	0.162	0.02
between-subject error	2376	(<.001)		
		Within subjects		
Missingness (M)	2	151.046***	0.06	<.001
M x N	4	0.377	0	0.755
M x S	6	63.23***	0.074	<.001
M x A	2	67.63***	0.028	<.001
M x N x S	12	3.7***	0.009	<.001
M x N x A	4	7.874***	0.007	<.001
M x S x A	6	26.238***	0.032	<.001
M x N x S x A	12	4.998***	0.012	<.001
M within-group error	4752	(<.001)		

Note. The outcome has a grand mean of .035 and standard deviation of .015. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Table 7

Analysis of Variance for Standard Error of Cross-level Interaction Effect

Source	df	F	partial ?^{2}	p
Between subjects				
Normality (N)	2	77.231***	0.061	<.001
Sample (S)	3	9527.481***	0.923	<.001
Analysis (A)	1	391.486***	0.141	<.001
N x S	6	166.391***	0.296	<.001
N x A	2	12.66***	0.011	<.001
S x A	3	30.328***	0.037	<.001
N x S x A	6	14.621***	0.036	<.001
between-subject error	2376	(< .001)		
Within subjects				
Missingness (M)	2	106***	0.043	<.001
M x N	4	52.089***	0.042	<.001
M x S	6	43.075***	0.052	<.001
M x A	2	0.893	0	0.374
M x N x S	12	75.13***	0.159	<.001
M x N x A	4	43.438***	0.035	<.001
M x S x A	6	14.691***	0.018	<.001
M x N x S x A	12	48.065***	0.108	<.001
M within-group error	4752	(< .001)		

Note. The outcome has a grand mean of .033 and standard deviation of .014. Values enclosed in parentheses represent mean square errors.

*p < .05. **p < .01. ***p < .001.

Estimation Bias Due to Multiple Imputation

The bias of parameter estimates was evaluated with the RMSD measure developed by Gold, Bentler, and Kim (2003) as introduced in Chapter Three. This index was computed for both the regression coefficients and standard error estimates across 100 iterations. Smaller bias implied a more robust modeling method. The means and standard deviations of these RMSD for the main effect, cross-level interaction effect, standard errors of main effect, and standard errors of interaction effect were individually reported in Table 8 to Table 11. Four full factorial ANOVA models with repeated measures were conducted to investigate the effects of four factors on RMSD of the two parameter estimates and their standard errors. The two RMSD indices from imputation of data with 15% and 30% missingness were the dependent variables. The ANOVA tables for these analyses are presented in Table 12 to Table 15.

RMSD of Parameter Estimates

With respect to the bias in main effect of x1, among the four design factors the only factor that did not exhibit a significant difference was the analysis method. The proportion of missing data and sample size were highly significant ($p < .001$) while the normality condition was significant beyond the .01 level. As for the bias in cross-level interaction, the strongest effects still appeared for sample size and missingness proportion ($p < .001$ for both) while the analysis method and normality condition did not achieve any statistically significant differences. All the interactions among missingness proportion, analysis method, and sample size were significant at or beyond .01 level.

Table 8

Means and Standard Deviations of RMSD in Parameter Estimates: Main Effect

Missingness	*HLM*				*SEM*			
Proportion	0.150		0.300		0.150		0.300	
	M	SD	M	SD	M	SD	M	SD
Normal								
N=300 (30 x 10)	0.524	0.332	0.592	0.443	0.539	0.458	0.627	0.546
N=500 (10 x 50)	0.761	1.000	0.977	1.458	0.702	0.620	0.895	0.818
N=1500 (30 x 50)	0.147	0.073	0.194	0.096	0.150	0.068	0.190	0.090
N=2500 (50 x 50)	0.099	0.049	0.118	0.056	0.102	0.046	0.116	0.053
Moderately nonnormal								
N=300 (30 x 10)	0.243	0.125	0.325	0.140	0.230	0.120	0.319	0.141
N=500 (10 x 50)	1.182	4.969	0.994	1.402	0.720	0.948	0.935	0.975
N=1500 (30 x 50)	0.473	0.765	0.551	0.720	0.405	0.388	0.517	0.654
N=2500 (50 x 50)	0.118	0.053	0.136	0.055	0.119	0.049	0.134	0.054
Severely nonnormal								
N=300 (30 x 10)	0.466	0.263	0.556	0.322	0.494	0.391	0.743	2.019
N=500 (10 x 50)	0.231	0.109	0.287	0.133	0.218	0.104	0.284	0.145
N=1500 (30 x 50)	0.336	0.223	0.418	0.266	0.359	0.281	0.450	0.329
N=2500 (50 x 50)	0.140	0.060	0.174	0.074	0.141	0.057	0.164	0.073

Table 9

Means and Standard Deviations of RMSD in Parameter Estimates: Cross-level Interaction Effect

Missingness	*HLM*				*SEM*			
Proportion	0.150		0.300		0.150		0.300	
	M	SD	M	SD	M	SD	M	SD
Normal								
N=300 (30 x 10)	1.150	0.874	1.211	0.784	1.190	1.053	1.658	1.353
N=500 (10 x 50)	1.043	1.106	1.131	1.023	1.000	0.880	1.085	0.968
N=1500 (30 x 50)	0.782	0.861	0.979	1.148	0.682	0.482	0.767	0.858
N=2500 (50 x 50)	0.620	0.515	0.689	0.541	0.603	0.421	0.678	0.466
Moderately nonnormal								
N=300 (30 x 10)	1.238	1.702	1.195	0.965	1.341	1.379	1.969	2.376
N=500 (10 x 50)	1.295	2.932	1.473	3.991	0.858	0.659	1.068	0.814
N=1500 (30 x 50)	0.956	1.585	1.129	1.995	0.822	0.863	1.002	1.197
N=2500 (50 x 50)	0.835	2.208	0.976	2.313	0.633	0.656	0.898	1.767
Severely nonnormal								
N=300 (30 x 10)	1.010	0.813	1.254	0.684	1.530	3.812	2.285	5.724
N=500 (10 x 50)	0.870	0.572	1.155	0.856	1.016	1.219	1.164	0.869
N=1500 (30 x 50)	0.912	1.176	1.190	2.787	0.777	0.817	0.839	0.581
N=2500 (50 x 50)	0.489	0.430	0.585	0.485	0.666	0.523	0.879	0.810

Table 10

Means and Standard Deviations of RMSD in Standard Error Estimates: Main Effect

Missingness	*HLM*				*SEM*			
Proportion	0.150		0.300		0.150		0.300	
	M	SD	M	SD	M	SD	M	SD
Normal								
N=300 (30 x 10)	0.184	0.082	0.311	0.124	0.118	0.057	0.181	0.070
N=500 (10 x 50)	0.120	0.061	0.206	0.087	0.058	0.032	0.073	0.040
N=1500 (30 x 50)	0.079	0.032	0.093	0.037	0.050	0.023	0.061	0.026
N=2500 (50 x 50)	0.063	0.024	0.072	0.029	0.044	0.017	0.050	0.022
Moderately nonnormal								
N=300 (30 x 10)	0.195	0.085	0.299	0.126	0.163	0.063	0.246	0.093
N=500 (10 x 50)	0.150	0.063	0.202	0.084	0.056	0.031	0.071	0.042
N=1500 (30 x 50)	0.080	0.034	0.098	0.046	0.028	0.016	0.032	0.017
N=2500 (50 x 50)	0.064	0.032	0.074	0.032	0.040	0.016	0.045	0.019
Severely nonnormal								
N=300 (30 x 10)	0.237	0.093	0.349	0.131	0.113	0.056	0.166	0.083
N=500 (10 x 50)	0.156	0.072	0.270	0.092	0.112	0.057	0.151	0.070
N=1500 (30 x 50)	0.106	0.040	0.114	0.049	0.031	0.016	0.037	0.018
N=2500 (50 x 50)	0.074	0.030	0.088	0.034	0.036	0.015	0.043	0.019

Table 11

Means and Standard Deviations of RMSD in Standard Error Estimates: Cross-level Interaction Effect

Missingness	*HLM*				*SEM*			
Proportion	0.150		0.300		0.150		0.300	
	M	SD	M	SD	M	SD	M	SD
Normal								
N=300 (30 x 10)	0.200	0.080	0.303	0.131	0.119	0.057	0.171	0.089
N=500 (10 x 50)	0.140	0.064	0.208	0.090	0.056	0.033	0.072	0.043
N=1500 (30 x 50)	0.081	0.035	0.099	0.042	0.050	0.023	0.061	0.026
N=2500 (50 x 50)	0.070	0.034	0.085	0.040	0.044	0.017	0.050	0.022
Moderately nonnormal								
N=300 (30 x 10)	0.204	0.091	0.361	0.111	0.165	0.064	0.271	0.111
N=500 (10 x 50)	0.188	0.078	0.236	0.096	0.056	0.031	0.073	0.042
N=1500 (30 x 50)	0.081	0.047	0.109	0.054	0.028	0.016	0.032	0.017
N=2500 (50 x 50)	0.082	0.036	0.091	0.044	0.040	0.016	0.045	0.019
Severely nonnormal								
N=300 (30 x 10)	0.148	0.119	0.430	0.219	0.113	0.056	0.175	0.092
N=500 (10 x 50)	0.223	0.113	0.282	0.117	0.115	0.057	0.150	0.070
N=1500 (30 x 50)	0.117	0.046	0.139	0.065	0.032	0.016	0.040	0.017
N=2500 (50 x 50)	0.105	0.049	0.118	0.058	0.036	0.015	0.043	0.019

Table 12

Analysis of Variance for RMSD of Main Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	5.033**	0.004	0.007
Sample (S)	3	52.397***	0.063	<.001
Analysis (A)	1	0.407	0	0.524
N x S	6	19.912***	0.048	<.001
N x A	2	1.045	0.001	0.352
S x A	3	1.068	0.001	0.361
N x S x A	6	0.274	0.001	0.949
between-subject error	2376	(1.208)		
		Within subjects		
Missingness (M)	1	14.544***	0.006	<.001
M x N	2	0.327	0	0.721
M x S	3	0.998	0.001	0.393
M x A	1	1.6	0.001	0.206
M x N x S	6	0.851	0.002	0.531
M x N x A	2	0.707	0.001	0.493
M x S x A	3	0.603	0.001	0.613
M x N x S x A	6	0.908	0.002	0.488
M within-group error	2376	(.463)		

Note. The outcome has a grand mean of .408 and standard deviation of .951. Values enclosed in parentheses represent mean square errors.

*p < .05. **p < .01. ***p < .001.

Table 13

Analysis of Variance for RMSD of Cross-level Interaction Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	1.797	0.002	0.166
Sample (S)	3	21.35***	0.026	<.001
Analysis (A)	1	0.629	0	0.428
N x S	6	0.34	0.001	0.916
N x A	2	1.592	0.001	0.204
S x A	3	5.358**	0.007	0.001
N x S x A	6	0.644	0.002	0.695
between-subject error	2376	(5.11)		
		Within subjects		
Missingness (M)	1	99.067***	0.04	<.001
M x N	2	2.854	0.002	0.058
M x S	3	5.64**	0.007	0.001
M x A	1	7.983**	0.003	0.005
M x N x S	6	0.747	0.002	0.612
M x N x A	2	1.231	0.001	0.292
M x S x A	3	11.929***	0.015	<.001
M x N x S x A	6	0.286	0.001	0.944
M within-group error	2376	(.513)		

Note. The outcome has a grand mean of 1.033 and standard deviation of .1.704. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Table 14

Analysis of Variance for RMSD of Standard Error of Main Effect

Source	df	F	partial ?^{2}	p
		Between subjects		
Normality (N)	2	46.517	0.038	<.001
Sample (S)	3	1683.439	0.68	<.001
Analysis (A)	1	1578.277	0.399	<.001
N x S	6	30.139	0.071	<.001
N x A	2	26.622	0.022	<.001
S x A	3	90.504	0.103	<.001
N x S x A	6	26.801	0.063	<.001
between-subject error	2376	(.004)		
		Within subjects		
Missingness (M)	1	584.641	0.197	<.001
M x N	2	1.754	0.001	0.173
M x S	3	134.417	0.145	<.001
M x A	1	80.829	0.033	<.001
M x N x S	6	4.51	0.011	<.001
M x N x A	2	2.468	0.002	0.085
M x S x A	3	18.175	0.022	<.001
M x N x S x A	6	1.627	0.004	0.136
M within-group error	2376	(.003)		

Note. The outcome has a grand mean of .119 and standard deviation of .100. Values enclosed in parentheses represent mean square errors. *p < .05. **p < .01. ***p < .001.

Table 15

Analysis of Variance for RMSD of Standard Error of Cross-level Interaction Effect

Source	df	F	partial ?^{2}	p
	Between subjects			
Normality (N)	2	92.833	0.072	<.001
Sample (S)	3	1343.915	0.629	<.001
Analysis (A)	1	1861.518	0.439	<.001
N x S	6	29.188	0.069	<.001
N x A	2	51.657	0.042	<.001
S x A	3	89.23	0.101	<.001
N x S x A	6	21.442	0.051	<.001
between-subject error	2376	(.005)		
	Within subjects			
Missingness (M)	1	472.472	0.166	<.001
M x N	2	3.652	0.003	0.026
M x S	3	132.085	0.143	<.001
M x A	1	65.871	0.027	<.001
M x N x S	6	4.972	0.012	<.001
M x N x A	2	1.126	0.001	0.325
M x S x A	3	14.072	0.017	<.001
M x N x S x A	6	2.913	0.007	<.001
M within-group error	2376	(.005)		

Note. The outcome has a grand mean of .130 and standard deviation of .116.

Values enclosed in parentheses represent mean square errors.

*p < .05. **p < .01. ***p < .001.

As before the LSD post-hoc multiple comparisons were employed to follow up these significant findings. The largest amount of bias on the main effect of x1 was produced by moderately nonnormal data and severely nonnormal data for the cross-level interaction even though these differences were not statistically significant. In the case of sample size, N = 500 condition (50 units with 10 subjects per unit) yielded greatest bias for the main effect of x1 while for the cross-level interaction it occurred with the smallest sample (N = 300 at 10 units with 30 subjects per unit). A higher proportion of missing data (30%) consistently generated larger bias in terms of the two parameter estimates even though these differences were not statistically significant.

RMSD of Standard Error Estimates

All the four factors had highly significant effects (p < .001) for the bias of both standard error estimates. Their effects, discovered by the post-hoc multiple comparisons, were very consistent across various designs. Severely nonnormal data produced the most bias among the three normality conditions. It was greater than that from moderately nonnormal data, which was in turn associated with more bias than normal data. It also held for sample size that the larger the sample, the smaller the estimation bias. HLM tended to generate more bias than multilevel SEM in all designs. Higher proportion of missing data had the same effect in incurring larger bias. All the above differences were statistically significant with the p values less than .001.

Modeling Performance under Different Sample Sizes

Robustness under Small Samples

Research Question #5 addressed the robustness of HLM and multilevel SEM under small samples, which are a problem commonly encountered in application of multilevel analysis in behavioral science. This issue can be evaluated from three perspectives: empirical statistical power, efficiency of parameter estimates, which was studied through the mean empirical standard errors of parameter estimates, and bias in parameter estimates and standard errors.

In term of empirical statistical power, if the normal data were taken as the baseline level, both of the nonnormality conditions overestimated the power for the main effect of x1. Additionally, the power of HLM and multilevel SEM were very similar for the main effect. The power did not vary by sample size. The only exception was the largest sample (N = 2,500 with 50 subjects within each of 50 units) under nonnormal data, in which HLM exhibited higher power for the cross-level interaction than multilevel SEM.

Similar to statistical power, not much of an effect was found for sample size with respect to efficiency of parameter estimates or estimation bias. The mean empirical standard errors in SEM were smaller for the main effect but larger for the cross-level interaction compared with HLM. Thus HLM was more efficient in estimating the cross-level effect while SEM would be better for the main effect. But this pattern held across all sample sizes. The estimation bias effects were similar.

The Effects of Sample Sizes at Different Levels

To investigate the effects of sample sizes at different levels for HLM and multilevel SEM, four sample sizes were simulated: two with a fixed number of subjects per unit and varying number of units (30 subjects per unit with 10 or 50 units) while the other two were simulated with a fixed number of units and varying number of subjects per unit (50 units with 10 or 50 subjects per unit). For empirical statistical power, generally it was found that the larger the sample the higher the power. Nevertheless, with comparatively small samples (N = 300 or 500 compared to 1,500 and 2,500) both HLM and multilevel SEM exhibited higher power with more subjects within each unit than more units with fewer subjects under all three normality conditions, even though the differences became less distinct in severely nonnormal data. As for the estimation bias (RMSD), the sample size of 300 produced less bias in the main effect of x1 than that of 500, but the larger number of units was helpful in ensuring less bias in the cross-level interaction effect.

The purpose of this research question was to find out if the sample size at one level has to be compromised, which level would lead to less loss of power. That is why the larger samples were not discussed here. Given their sizes (N = 1,500 and 2,500), which meant at least fairly adequate sample size was obtained for one level, their overall performances would be quite satisfactory if there was an effect to be detected.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The purpose of this study was to compare the performance of HLM and multilevel SEM in a two-level hierarchical model under nonnormal data with or without missing data under various sample sizes. Missing data were estimated by the EM algorithm multiple imputation and then analyzed in the same way as complete data sets. Based on the literature review, in which it was found that not much work has been done within the framework of multilevel data, it was my hope that this simulation study would provide some preliminary guidelines for empirical researchers using multilevel analyses.

This chapter summarizes the findings of the study with regard to the research questions asked in the Chapter I. Relevant implications and conclusions were drawn based on these finding in terms of their potential influence on research practice. Limitations are discussed. Finally this chapter presents some recommendations for future research related to the purpose of the study. The discussion was organized around the original research questions.

Research Question #1

HLM and multilevel SEM are two closely related statistical modeling techniques with explicitly different applications. HLM was found to have same level of empirical statistical power as multilevel SEM under both normality and nonnormality conditions studied here. For normal data and large samples higher power was consistently found with both methods. This difference became smaller under moderately nonnormal data and almost disappeared with severely nonnormal data. Since both HLM and SEM require normality to be met prior to data analysis, one possible explanation was that the statistical power in nonnormal conditions was inflated. This could due to the underestimated standard errors, which were distorted under nonnormal conditons.

A separate null case of severely nonnormal data with 2500 cases was explored in which the intra-class correlation was close to zero. The empirical type I error rate for the cross-level interaction was found to be around .10 instead of the nominal level of .05, which could be the cause of the inflated statistical power under severely nonnormal data, especially with large sample sizes. So results from multilevel analyses should be taken with great caution if the data are found to violate the normality condition.

When the number of units was fixed (at 50 in this study), power for the moderately nonnormal data was very similar to that for normal data. However, the power changes for the sample sizes with fixed number of subjects per unit were somewhat mixed. The small sample of 300 with 10 units and 30 subjects per unit showed a power increase (+15%) from normality to moderately nonnormal data conditions, while the sample with 50 units and 30 subjects per unit lost power under nonnormality and to a greater degree (-30%). This indicates that with moderately nonnormal data retaining more second-level units will help more than recruiting more subjects within-unit so to keep power at the same level.

The power for severely nonnormal data was not much affected by sample size, probably because the large deviation from normality caused serious problems with standard error estimates, so that the sample size is not a crucial issue compared to the severe nonnormality condition. The power for the cross-level interaction under severely nonnormal data was inflated overall. This might be due to application of the multiple imputation procedure to data with a serious violation of the multivariate normality assumption. By assuming the original data to be normal, the imputed data may have been greatly distorted from its original covariance matrix form.

Research Question #2

Analysis of variance with repeated measures for parameter estimates and standard errors as dependent variables indicated that both differed by all the design main effects of normality condition, sample size, analysis method, and proportion of missingness and by all their two-way interactions. This suggested that HLM and multilevel SEM, in spite of similar statistical power, can not be used interchangeably. The differences may also interact with sample size to influence parameter estimates. No consistent pattern was found for the normality conditions, which could explain the stable parameters that were produced based on multilevel analysis of nonnormal data in research practice.

Empirical researchers should pay attention to the fact that for these ANOVA analyses the “statistical differences” should be differentiated from the “actual differences”. Even though most effects were statistically significant beyond .001 level, the significance might be only due to the huge sample size used in the analysis, while the actual differences among the least square means of the estimates were barely noticeable. This also applies to all the other ANOVA analysis results.

Neither of the analysis methods seemed to be particularly sensitive to nonnormality. Compared to multilevel SEM, larger standard error estimates were found for HLM on the main effect and smaller estimates for the cross-level interaction. Thus HLM appears to be more effective in detecting the cross-level interaction effects while multilevel SEM would be more useful exploring the main effects in complex path models. The EM algorithm imputation appeared to work very well with MAR missing data with multilevel structure because the results from the imputed data were close to those from complete data.

Research Question #3

As mentioned above, the performance of the EM algorithm for multiple imputation appeared to be satisfactory with MAR data in multilevel analyses in terms of empirical statistical power and parameter estimates comparing them for complete and imputed data. More systematic investigation on this issue using RMSD (Gold, Bentler, and Kim, 2003) revealed that all three design factors produced some differences except for the analysis method. This is helpful because the degree of estimation bias resulting from data imputation did not differ by the multilevel analysis technique. HLM and multilevel SEM were not differentially sensitive to nonnormality with respect to estimation bias compared with the factors of sample sizes and proportions of missing data.

The sample size of 500 yielded the largest bias for the main effect of x1. One explanation may be that in order to get an accurate main effect estimate the unit size, which is the number of subjects per unit or cluster, cannot be too small. Otherwise, the within-unit covariance matrices at the first level are not stable and thus unable to generate proper estimates for the first level main effect. On the other hand, the data condition with 300 cases produced the greatest bias for the cross-level interaction. Similarly, this could be due to lack of second-level variance given the small number of units available in the data. These findings indicate the differentiating effects of sample sizes at different levels.

Research Question #4

Both the percentages of missing data examined in the present study are quite common in real world data. The level of 15% represents a somewhat low to moderate degree of data missingness, while 30% may be considered moderate in traditional GLM data analysis, but is regarded as high level in multilevel analyses. As expected it was consistently found that a higher proportion of missing data tended to produce larger standard errors and also cause more bias in parameter estimates and standard errors, which corresponds to findings by Gold, Bentler, and Kim (2003) and Gibson and Olejnik (2003).

Compared with results from complete data analysis, the EM algorithm multiple imputation worked well with multilevel data. Future researchers should feel confident applying this procedure with a missing data level of no more than 30%. However, since the proportion of missing data at 30% produced significantly larger standard errors and siginficantly greater estimation bias, multiple imputation should be applied cautiously to multilevel data with more than 30% missing data.

Research Question #5

Not much difference was found between HLM and multilevel SEM in terms of their statistical power under different sample sizes. Thus, neither achieved better power under small sample sizes than the other. Since power may be inflated under nonnormal data regardless of the degree of nonnormality, the significant findings found for severely nonnormal data with small samples should be interpreted with great caution. This was particularly true for the cross-level interaction effect.

On the other hand, analysis with large samples worked very well in both HLM and multilevel SEM, especially for interaction effects, under which HLM performed even better than SEM. This pattern was consistent for all the distribution conditions. Thus, even with severely nonnormal data, HLM may be effectively utilized to detect cross-level effects, although at the same time the estimates for main effects may not be very reliable. Similar situations were discovered in terms of parameter estimation efficiency and estimation bias. Based on these findings the conclusion here is that neither HLM nor multilevel SEM is robust to small sample sizes, which suggests that multilevel analyses are inappropriate for small sample data.

Research Question #6

Among the four sample designs the two comparatively small samples clearly showed the differentiating effect of sample size at different levels. The sample size of 300 was more useful to detect the main effect of x1. Possibly, this was because this sample provided more information with each unit, and thus made the first-level covariance matrices more stable and the first level distribution more variable. In considering cross-level interaction, however, it is more beneficial to include more units at second level than increasing the number of cases at the unit level. One explanation is that adequate variability of second level means was a necessary prerequisite for the cross-level interaction to be identified in spite of the number of subjects within each unit. This implies that it is a good idea to identify one's research focus based on a careful literature review because this may help to determine the sampling plan before data collection to include an appropriately large number of units or perhaps more cases per unit, depending on the need to detect a cross-level interaction or not. These findings were applicable to both HLM and multilevel SEM. On the other hand, large sample size, no matter at which level always appears to be valuable. One advantage is to offset the negative influence of nonnormal data, especially uner severely nonnormal condition.

Conclusions

Violations of normality and missing data treatment have been issues broadly explored in single-level data analysis. The present study studied these topics from the perspective of a multilevel data structure and with HLM and SEM analysis methods. The EM algorithm for multiple imputation was found to be effective dealing with MAR missing data with nesting structure. A higher proportion of missingness tended to generate imputed data more deviant from the original complete data and thus generate larger standard errors and greater estimation bias. The nonnormality conditions, severe or not, did not really affect imputation performance even though multivariate normality is assumed for the procedure. However, under severely nonnormal data condition the imputed data matrix tended to be more normal than its original form. This is probably due to the multivariate normality assumption in multiple imputation which distorts the original data to their more normal form thus exhibited abnormally high power for the main effect.

HLM and SEM had similar power to detect the main effect of first level predictor and cross-level interaction effect across various sample sizes, and higher power was found with larger samples. However, data with small sample size may not be appropriate for multilevel analyses since there was lack of sufficient information at either level for stable estimation. Under severely nonnormal data HLM had better power for the interaction effects than SEM, but neither method was very sensitive to violations of normality in terms of parameter estimates and standard errors.

Depending on the research focus, adding more units will be useful in detecting cross-level interactions while larger numbers per unit can ensure sufficient information to estimate the first level predictor main effect. However, when the sample size at either level exceeds a certain number (in the current study 30 for the unit size and 50 for the number of units) both HLM and multilevel SEM appear to have sufficient information and statistical power to detect appropriate effects presented in the sample data.

Recommendations

Since HLM has been the preferred method for multilevel analysis, this study was intended to provide some preliminary guidelines for future research to consider SEM when HLM cannot be employed, particularly for complicated path models that focus on the main effect.

Neither HLM nor multilevel SEM appeared to be particularly sensitive to nonnormality condition no matter moderately or severely nonnormal. This does not imply that the normality assumption is not important in multilevel analysis. Instead, it just suggested that nonnormality may not greatly alter the results. But when the results are quite deviant from what is expected, especially with small to moderate sample size, it would be helpful to examine whether the normality assumption is violated in the data.

The hypothesis that sample sizes at both levels influence the statistical power was supported in that the power for main effect was higher with bigger cluster size while for the cross-level interaction it was more efficient to have more clusters sampled. For either cases it was true that the more the better. The marginal increase of power stops when the sample size at the each level reaches certain number. In the two-level model of the current study it appeared that 30 was a number adequate for the first-level cluster size and 50 for the number of second-level clusters. It should be noted that due to the limited combinations of sample sizes explored in this study, these two levels did not serve as the criterion or cutoff points for sample size investigation. The actual effect size also has some role in determining the necessary sample size.

The EM algorithm for multiple imputation can be employed confidently for multilevel data with MAR missingness. Parameter estimates from imputed data proved to be quite stable despite a high proportion of missing data; since higher proportions tended to generate larger standard errors, the significance tests should be considered with caution in such cases. Analysis of complete data should always be attempted given adequate sample size. When missing data are encountered, researchers should be careful in interpreting imputed data results with a proportion of missingness higher than 30%.

Limitations

Due to practical constraints, the present study only investigated multilevel data with low to moderate intraclass correlation, which made the inference for cross-level interaction fairly difficult, because the effect was rarely detected. Future research should replicate these findings under multilevel data having stronger nesting effects. It would also be meaningful to further increase the proportion of missingness to levels that have been proven to be easily accommodated by imputation in single-level data analysis. As the present study only studied four sample size designs, more combination of sample sizes might be explored to confirm the usual power graph for multilevel research design from popular power estimation software. This would provide practical guidance for empirical researchers with a need to design the sampling procedure in hierarchical structure.

REFERENCES

Arbuckle, J. L. (1999). *Amos users' guide version 4.0*. Chicago: Small Waters Corporation.

Arminger, G., & Rothe, G. (1993). ML and pseudo-ML estimation in covariance structural models for nonnormal distribution: A Simulation Study. In Krebs, D., & Schmidt, P. (Eds.), *New directions in attitude measurement* (pp. 266-276). Berlin: De Gruyter.

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

Bentler, P. M. (1994). On the quality of test statistics in covariance structure analysis: Caveat emptor. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 237-260). New York: Plenum.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*, 563-592.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brockmeier, L., Kromrey, J. & Hogarty, K. (2003). Nonrandomly missing data in multiple regression analysis: An empirical comparison of ten missing data treatments. *Multiple Linear Regression Viewpoints*, *29* (1)**,** 8-29.

Brown, R. L. (1990). The robustness of 2SLS estimation of a non-normally distributed confirmatory factor analysis model. *Multivariate Behavioral Research, 25* (4), 455-466.

Bunting, B. P., Adamson, G., & Mulhall, P. K. (2002). A Monte Carlo examination of an MTMM model with planned incomplete data structures. *Structural Equation Modeling, 9* (3), 369-389.

Chou ,C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39,* 1-8.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of American Statistical Association, 76,* 341-353.

Enders, C. K. (2001). The impact of nonnormality on full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430-457.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified models. *Structural Equation Modeling, 5* (5), 701-735.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. (2003). *SAS for Monte Carlo studies: A guide for quantitative researchers.* Cary, NC: SAS Institute, Inc.

Finch, J. F. (1993). *The robustness of maximum likelihood parameter estimates in structural equation models with nonnormal variables.* Unpublished Doctoral Dissertation, Arizona State University, Tempe, AZ.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormallity on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4* (2), 87-105.

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

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling, 7,* 319-355.

Gold, M. S., Bentler, P. M., & Kim, K. H. (2003). A comparison of maximum-likelihood and asymptotically distribution-free methods of treating incomplete nonnormal data. *Structural Equation Modeling, 10* (1), 47-79.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical methods* (2nd ed.). London: Edward Arnold; New York: Halstead Press.

Goldstein, H., & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*, 455–467.

Graham, J.W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*, 197–218.

Hox, J. J. (2002). *Multilevel analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J. & Kreft, G. (1994). Multilevel analysis methods. *Sociological Methods & Research, 22*, 283-299

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL8: User's reference guide*. Chicago: Scientific Software International.

Julian, M. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8,* 325-352.

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

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing data treatments. *Educational and Psychological Measurement, 54* (3), 573-593.

Laird, N. M., & Ware, H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963-974.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1-41.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A.,&Rubin, D. B. (1990). The analysis of social science data with missing values. *SociologicalMethods and Research, 18,* 292–326.

Longford, N. T. and Muthen, B. O. (1992). Factor analysis for clustered populations. *Psychometrika, 57,* 581-97.

Mason, W. M., Wong, G. M., & Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72-103). San Francisco: Jossey-Bass.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Mueller, R. O. (1997). Structural equation modeling: Back to basics. *Structural Equation Modeling, 4* (4), 353-369.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations.

Psychometrika, 54, 557–585.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171-189.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus User's Guide.* (3rd Ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87–99). San Diego: Academic.

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

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

Roberts, K. (2003, February). *An introductory primer on multilevel and hierarchical linear modeling.* Article presented at the Hierarchical Linear Workshop at Texas A&M University, College of Education.

Rosenberg, B. (1973). Linear regression with randomly dispersed parameters. *Biometrika, 60,* 61-75.

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

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

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

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241–254.

SAS Institute Inc. (2002). *Statistical analysis system version 8.1.* Cary, NC: SAS Institute Inc.

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

Schafer, J.L. (1999). NORM: *Multiple imputation of incomplete multivariate data under a normal model, version 2, Software for Windows 95/98/NT.* Available from http://www.stat.psu.edu/~jls/misoftwa.html.

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

Tanguma, J. (November, 2000). *Review of literature on missing data.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling, Greem, KY.

Tabachinick, B., & Fidell, L. (1996). *Using multivariate statistics*. New York: Harper Collins College Publishers.

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling, 3* (3), 228-247.

Willson, V. L. & Zhang, D. (2003, April). *Comparing HLM, HLM-SEM, and Residual SEM Analyses for Multilevel Models.* Paper presented at the 84th Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Zhang, D. & Willson, V. L. (2004, April). *Empirical Power and Type I Error Rates for Cross-Level Interactions in Multilevel Analysis.* Paper presented at the 69th Annual Meeting of the Psychometric Society, Monterey, CA.

APPENDIX

SAS SYNTAX FOR MODERATELY NONNORMAL DATA

WITH 10 CLUSTERS AND 30 SUBJECTS PER CLUSTER

```
proc datasets kill;
options nosource nonotes;
data a (type=corr); _type_= 'corr';
 input xc1-xc3;
 cards;
 1.00   .    .
  .30  1.00  .
  .30   .30  1.00
;

%macro simu(r);
 %do k=1 %to &r;
proc factor n=3 data=a outstat=facout noprint;
run;
data pattern; set facout;
 if _type_='PATTERN';
 drop _type_ _name_;
RUN;

PROC IML;
use  pattern;
read all var _num_ into F;
F=F`;
RUN;

DATA=RANNOR(J(300,3,0));
DATA=DATA`;
Z = F*DATA;
Z = Z`;

XC1=Z[,1];
XC2=Z[,2];
XC3=Z[,3];

Z=XC1||XC2||XC3;
```

```
CREATE A FROM Z [COLNAME={XC1 XC2 XC3}];
APPEND FROM Z;

data b; set a;
XC1= -0.07311802793159 + 0.92409763318404*XC1 + 0.07311802793159*XC1**2 +
0.02298245181387*XC1**3;
XC2= -0.07311802793159 + 0.92409763318404*XC2 + 0.07311802793159*XC2**2 +
0.02298245181387*XC2**3;
XC3= -0.07311802793159 + 0.92409763318404*XC3 + 0.07311802793159*XC3**2 +
0.02298245181387*XC3**3;

proc sort data=b; by xc2;


DATA D1; SET B; id=_n_;

if _n_ < 31 then cls=1;if _n_ > 30 and _n_ < 61 then cls=2;
 if _n_ > 60 and _n_ < 91 then cls=3;if _n_ > 90 and _n_< 121 then cls=4;
 if _n_ > 120 and _n_ < 151 then cls=5;if _n_ > 150 and _n_ < 181 then cls=6;
 if _n_ > 180 and _n_ < 211 then cls=7;if _n_ > 210 and _n_ < 241 then cls=8;
 if _n_ > 240 and _n_ < 271 then cls=9;if _n_ > 270 and _n_ < 301 then cls=10;

proc sort;by cls;
proc means noprint ;by cls;var xc2 xc3; output out =d2a mean=xc2bar xc3bar;
data d2;merge d1 d2a;by cls; t=1;
data corr.mdxc1030; set d2; keep id cls xc3 xc3bar;

**HLM**;
proc mixed data=d2 ic noclprint noitprint noinfo; class cls;
model xc3=xc1 xc2bar xc1*xc2bar/solution ddfm=bw;
random intercept /sub=cls type=un;
ods output solutionf=out_hlm0 ;
data hlmout0; set out_hlm0;keep effect estimate stderr; proc sort; by effect;

**Multilevel SEM**;
data d3; set d2; proc sort; by cls;
proc means noprint; by cls; var xc1 xc3; output out=d3a mean=xc1bar xc3bar std=xc1sd
xc3sd;
data d4; merge d3 d3a; by cls; if cls ne .;
zxc1=(xc1-xc1bar)/xc1sd; zxc3=(xc3-xc3bar)/xc3sd; crossint=zxc1*xc2bar;

proc calis corr cov outest=out_sem0 noprint;
lineqs zxc3=b4 zxc1 + b6 crossint + e1;
std e1= the1; cov zxc1 crossint = 0; var zxc1 crossint zxc3;
```

```
data sema0; set out_sem0;  if _TYPE_='PARMS';
data semb0; set out_sem0;  if _TYPE_='STDERR';stb4=b4; stb6=b6;
 drop b4 b6;
data semc0; set sema0; b4_orig=b4; b6_orig = b6; keep b4_orig b6_orig;
data semd0; set semb0; stb4_orig=stb4; stb6_orig=stb6; keep stb4_orig stb6_orig;
data semout0;merge semc0 semd0;

**missing data imputation at missingness of 15***;
proc means data=d2 noprint; var xc2; output out = d2b p25 = p25_xc2;
data d5; set d2b; t=1;
data d6; merge d2 d5; by t;drop _type_ _freq_;

data d7; set d6; if xc2>p25_xc2 then z=1; else z=0;drop t ;

data d8; set d7;

proc surveyselect data=d8 sampsize=113 method=pps out=sm1; size z;
data d9; set sm1;t1=1; proc sort; by id;

data d10; set d8; proc sort; by xc1;
proc surveyselect data=d10 sampsize=113 method=pps out=sm2; size z;
data d11; set sm2; t2=1;proc sort; by id;

data d12; merge d9 d11 d8; by id; x3=xc3; x1=xc1;
data d13; set d12;
if t1=1 then x1='.';  if t2=1 then x3='.';

data d14; set d13; if x3 = '.' then r_x3=0; else r_x3=1;
if x1='.' then r_x1=0; else r_x1=1; drop t1 t2 samplingweight selectionprob;

data d15; set d14; keep id cls x1 xc2 xc2bar x3;

proc mi data=d15 seed=32173 out=out_imp  simple nimpu=5 ;
var x1 xc2bar x3;

data dimp; set out_imp; XC1=x1;XC3=x3;

**analyses of imputed datasets**;
**HLM**;
proc mixed data=dimp noclprint noitprint noinfo; class cls;
model xc3=xc1 xc2bar xc1*xc2bar/solution ddfm=bw ;by _imputation_;
random intercept /sub=cls type=un;
ods output solutionf=out_hlm1;
```

```
proc sort data=out_hlm1; by effect;
proc means noprint; by effect; var estimate StdErr; output out=est_hlm1 mean=estimate
stderr;

data hlmout1; set est_hlm1; estim_imp=estimate; stderr_imp=stderr;
keep effect estim_imp stderr_imp;
proc sort; by effect;

data d16; merge hlmout0 hlmout1; by effect;

data hlma1; set d16;if effect = 'XC1';
b1_orig=Estimate; b1_imp=estim_imp;stb1_orig=StdErr;  stb1_imp=stderr_imp;
t1_orig=Estimate/StdErr; t1_imp=estim_imp/stderr_imp;
pt1_orig= probt(t1_orig, 288); pt1_imp= probt(t1_imp, 288);
data hlmb1; set d16;if effect = 'xc2bar';
b2_orig=Estimate; b2_imp=estim_imp;stb2_orig=StdErr; stb2_imp=stderr_imp;
t2_orig=Estimate/StdErr; t2_imp=estim_imp/stderr_imp;
pt2_orig= probt(t2_orig, 8);pt2_imp= probt(t2_imp, 8);
data hlmc1; set d16;if effect = 'XC1*xc2bar';
b3_orig=Estimate; b3_imp=estim_imp; stb3_orig=StdErr; stb3_imp=stderr_imp;
t3_orig=Estimate/StdErr; t3_imp=estim_imp/stderr_imp;
pt3_orig= probt(t3_orig, 288);pt3_imp= probt(t3_imp, 288);

data hlm1; merge hlma1 hlmb1 hlmc1;
keep b1_orig b2_orig b3_orig stb1_orig stb2_orig stb3_orig t1_orig t2_orig t3_orig
pt1_orig pt2_orig pt3_orig
b1_imp b2_imp b3_imp stb1_imp stb2_imp stb3_imp t1_imp t2_imp t3_imp pt1_imp
pt2_imp pt3_imp;

data d17; set hlm1;
rmsd_b1= sqrt(abs(b1_orig-b1_imp)/abs(b1_orig));rmsd_b2= sqrt(abs(b2_orig-
b2_imp)/abs(b2_orig));
rmsd_b3= sqrt(abs(b3_orig-b3_imp)/abs(b3_orig));
rmsd_stb1= sqrt(abs(stb1_orig-stb1_imp)/abs(stb1_orig));
rmsd_stb2= sqrt(abs(stb2_orig-stb2_imp)/abs(stb2_orig));
rmsd_stb3= sqrt(abs(stb3_orig-stb3_imp)/abs(stb3_orig));

proc append base=corr.mtmp1030_hlm15 data=d17;

**SEM**;
data dimp2; set dimp; proc sort; by cls;
proc means noprint; by cls; var xc1 xc3; output out=dimp2a mean=xc1bar xc3bar
std=xc1sd xc3sd;
data dimp3; merge dimp2 dimp2a; by cls;
```

```
zxc1=(xc1-xc1bar)/xc1sd; zxc3=(xc3-xc3bar)/xc3sd; crossint=zxc1*xc2bar;

data dimp4; set dimp3; proc sort; by _imputation_;

proc calis corr cov outest=out_sem1 noprint; by _imputation_;
lineqs zxc3=b4 zxc1 + b6 crossint + e1;
std e1= the1; cov zxc1 crossint = 0; var zxc1 crossint zxc3;

data sema1; set out_sem1;  if _TYPE_='PARMS';
data semb1; set out_sem1;  if _TYPE_='STDERR';stb4=b4; stb6=b6;
 drop b4 b6;
data semc1; set sema1; b4_imp=b4; b6_imp= b6; keep b4_imp  b6_imp;
data semd1; set semb1; stb4_imp=stb4; stb6_imp=stb6; keep stb4_imp stb6_imp;
data semout1;merge semc1 semd1;
proc means noprint data=semout1; output out=est_sem1 mean=b4_imp b6_imp
stb4_imp stb6_imp;

data sem1; merge semout0 est_sem1;
t4_orig=b4_orig/stb4_orig;t6_orig=b6_orig/stb6_orig;
pt4_orig=probt(t4_orig,288);  pt6_orig=probt(t6_orig,288);
t4_imp=b4_imp/stb4_imp; t6_imp=b6_imp/stb6_imp;
pt4_imp=probt(t4_imp,288);  pt6_imp=probt(t6_imp,288);

data d18;set sem1;
rmsd_b4= sqrt(abs(b4_orig-b4_imp)/abs(b4_orig));
rmsd_b6= sqrt(abs(b6_orig-b6_imp)/abs(b6_orig));
rmsd_stb4= sqrt(abs(stb4_orig-stb4_imp)/abs(stb4_orig));
rmsd_stb6= sqrt(abs(stb6_orig-stb6_imp)/abs(stb6_orig));

proc append base=corr.mtmp1030_sem15 data=d18;

**missing data imputation at missingness of 30***;
proc surveyselect data=d8 sampsize=225 method=pps out=sm3; size z;
data d90; set sm3;t1=1; proc sort; by id;

proc surveyselect data=d10 sampsize=225 method=pps out=sm4; size z;
data d110; set sm4; t2=1;proc sort; by id;

data d120; merge d90 d110 d8; by id; x3=xc3; x1=xc1;
data d130; set d120;
if t1=1 then x1='.';  if t2=1 then x3='.';

data d140; set d130; if x3 = '.' then r_x3=0; else r_x3=1;
if x1='.' then r_x1=0; else r_x1=1; drop t1 t2 samplingweight selectionprob;
```

```
data d150; set d140; keep id cls x1 xc2 xc2bar x3;

proc mi data=d150 seed=32173 out=out_imp0  simple nimpu=5 ;
var x1 xc2bar x3;

data dimp0; set out_imp0; XC1=x1;XC3=x3;


**analyses of imputed datasets**;
**HLM**;
proc mixed data=dimp0 noclprint noitprint noinfo; class cls;
model xc3=xc1 xc2bar xc1*xc2bar/solution ddfm=bw ;by _imputation_;
random intercept /sub=cls type=un;
ods output solutionf=out_hlm10;
proc sort data=out_hlm10; by effect;
proc means noprint; by effect; var estimate StdErr; output out=est_hlm10 mean=estimate
stderr;

data hlmout10; set est_hlm10; estim_imp=estimate; stderr_imp=stderr;
keep effect estim_imp stderr_imp;
proc sort; by effect;

data d160; merge hlmout0 hlmout10; by effect;

data hlma10; set d160;if effect = 'XC1';
b1_orig=Estimate; b1_imp=estim_imp;stb1_orig=StdErr;  stb1_imp=stderr_imp;
t1_orig=Estimate/StdErr; t1_imp=estim_imp/stderr_imp;
pt1_orig= probt(t1_orig, 288); pt1_imp= probt(t1_imp, 288);

data hlmb10; set d160;if effect = 'xc2bar';
b2_orig=Estimate; b2_imp=estim_imp;stb2_orig=StdErr; stb2_imp=stderr_imp;
t2_orig=Estimate/StdErr; t2_imp=estim_imp/stderr_imp;
pt2_orig= probt(t2_orig, 8);pt2_imp= probt(t2_imp, 8);

data hlmc10; set d160;if effect = 'XC1*xc2bar';
b3_orig=Estimate; b3_imp=estim_imp; stb3_orig=StdErr; stb3_imp=stderr_imp;
t3_orig=Estimate/StdErr; t3_imp=estim_imp/stderr_imp;
pt3_orig= probt(t3_orig, 288);pt3_imp= probt(t3_imp, 288);

data hlm10; merge hlma10 hlmb10 hlmc10;
keep b1_orig b2_orig b3_orig stb1_orig stb2_orig stb3_orig t1_orig t2_orig t3_orig
pt1_orig pt2_orig pt3_orig
```

```
b1_imp b2_imp b3_imp stb1_imp stb2_imp stb3_imp t1_imp t2_imp t3_imp pt1_imp
pt2_imp pt3_imp;

data d170; set hlm10;
rmsd_b1= sqrt(abs(b1_orig-b1_imp)/abs(b1_orig));rmsd_b2= sqrt(abs(b2_orig-
b2_imp)/abs(b2_orig));
rmsd_b3= sqrt(abs(b3_orig-b3_imp)/abs(b3_orig));
rmsd_stb1= sqrt(abs(stb1_orig-stb1_imp)/abs(stb1_orig));
rmsd_stb2= sqrt(abs(stb2_orig-stb2_imp)/abs(stb2_orig));
rmsd_stb3= sqrt(abs(stb3_orig-stb3_imp)/abs(stb3_orig));

proc append base=corr.mtmp1030_hlm30 data=d170;

**SEM**;
data dimp20; set dimp0; proc sort; by cls;
proc means noprint; by cls; var xc1 xc3; output out=dimp2a0 mean=xc1bar xc3bar
std=xc1sd xc3sd;
data dimp30; merge dimp20 dimp2a0; by cls;
zxc1=(xc1-xc1bar)/xc1sd; zxc3=(xc3-xc3bar)/xc3sd; crossint=zxc1*xc2bar;

data dimp40; set dimp30; proc sort; by _imputation_;

proc calis corr cov outest=out_sem10 noprint; by _imputation_;
lineqs zxc3=b4 zxc1 + b6 crossint + e1;
std e1= the1; cov zxc1 crossint = 0; var zxc1 crossint zxc3;

data sema10; set out_sem10;  if _TYPE_='PARMS';
data semb10; set out_sem10;  if _TYPE_='STDERR';stb4=b4; stb6=b6;
 drop b4 b6;
data semc10; set sema10; b4_imp=b4; b6_imp= b6; keep b4_imp  b6_imp;
data semd10; set semb10; stb4_imp=stb4; stb6_imp=stb6; keep stb4_imp stb6_imp;
data semout10;merge semc10 semd10;
proc means noprint data=semout10; output out=est_sem10 mean=b4_imp b6_imp
stb4_imp stb6_imp;

data sem10; merge semout0 est_sem10;
t4_orig=b4_orig/stb4_orig;t6_orig=b6_orig/stb6_orig;
pt4_orig=probt(t4_orig,288);  pt6_orig=probt(t6_orig,288);
t4_imp=b4_imp/stb4_imp; t6_imp=b6_imp/stb6_imp;
pt4_imp=probt(t4_imp,288);  pt6_imp=probt(t6_imp,288);

data d180;set sem10;
rmsd_b4= sqrt(abs(b4_orig-b4_imp)/abs(b4_orig));
rmsd_b6= sqrt(abs(b6_orig-b6_imp)/abs(b6_orig));
```

```
rmsd_stb4= sqrt(abs(stb4_orig-stb4_imp)/abs(stb4_orig));
rmsd_stb6= sqrt(abs(stb6_orig-stb6_imp)/abs(stb6_orig) );

proc append base=corr.mtmp1030_sem30 data=d180;

run;
%end;
%mend simu;
%simu(100);
run;

data hlm15; set corr.mtmp1030_hlm15; m=15;
data sem15; set corr.mtmp1030_sem15; m=15;
data hlm30; set corr.mtmp1030_hlm30; m=30;
data sem30; set corr.mtmp1030_sem30; m=30;

data corr.mdfinal1030; merge hlm15 sem15 hlm30 sem30; by m; drop _type_
_freq_;run;
proc means data=corr.mdfinal1030;by m;run;
data final; set corr.mdfinal1030;
if pt1_orig >.975 or pt1_orig< .025 then cnt105a=1;else cnt105a=0;
if pt3_orig >.975 or pt3_orig< .025 then cnt305a=1;else cnt305a=0;
if pt4_orig >.975 or pt4_orig< .025 then cnt405b=1;else cnt405b=0;
if pt6_orig >.975 or pt6_orig< .025 then cnt605b=1;else cnt605b=0;
proc freq; tables cnt105a*cnt405b  cnt305a*cnt605b ;
proc means data=b n mean std skewness kurtosis; var xc1 xc2 xc3;
proc corr data=b nosimple; var xc1 xc2 xc3;
run;
```

VITA

Name: Duan Zhang

Address: Room 502 Unit 1 Building 6, Fukang Block, Wenhua East Road,
Xiaogan City, Hubei Province, P. R. China 432100

Email Address: duan.zhang@gmail.com

Education: B. S., International Economics,

University of International Business and Economics, China, 2000

M. S., Educational Psychology

Texas A&M University, 2002

Ph. D., Educational Psychology

Texas A&M University, 2005