

USING PRINCIPAL COMPONENT ANALYSIS (PCA) TO OBTAIN AUXILIARY  
VARIABLES FOR MISSING DATA IN LARGE DATA SETS

By  
Waylon J. Howard

Submitted to the graduate degree program in Psychology and the Graduate Faculty of the  
University of Kansas in partial fulfillment of the requirements for the degree of Doctor of  
Philosophy.

---

Chairperson Todd D. Little

---

Paul E. Johnson

---

Dale Walker

---

Wei Wu

---

Carol Woods

Date Defended: June 10, 2012

The Dissertation Committee for Waylon J. Howard  
certifies that this is the approved version of the following dissertation:

USING PRINCIPAL COMPONENT ANALYSIS (PCA) TO OBTAIN AUXILIARY  
VARIABLES FOR MISSING DATA IN LARGE DATA SETS

---

Chairperson Todd D. Little

Date approved: June 10, 2012

## Abstract

A series of Monte Carlo simulations were used to compare the relative performance of the inclusive strategy, where as many auxiliary variables are included as possible, with typical auxiliary variables (AUX) and a smaller set of auxiliary variables derived from principal component analysis ( $PCA_{AUX}$ ). We examined the influence of seven independent variables: magnitude of correlations, homogeneity of correlations across auxiliary variables, rate of missing, missing data mechanism, missing data patterns, number of auxiliary variables, and sample size on four dependent variables: raw parameter estimate bias, percent bias, standardized bias, and relative efficiency. Results indicated that including a single  $PCA_{AUX}$  (which explained about 40% of the total variance) is as beneficial for parameter bias as the AUX inclusive strategy. Findings also suggested the  $PCA_{AUX}$  method can capture a non-linear cause of missingness. Regarding efficiency, results indicate that the  $PCA_{AUX}$  method is at least as efficient as the inclusive strategy and potentially greater than 25% more efficient. Researchers can apply the results of this research to more adequately approximate the MAR assumption when the number of potential auxiliary variables is beyond a practical limit. The dissertation is divided into the following sections: 1) an introduction to missing data; 2) a brief review of the history of missing data; 3) a discussion of auxiliary variables; 4) an outline of principal component analysis; 5) a presentation of the  $PCA_{AUX}$  method; and finally, 7) a demonstration of the relative performance of the AUX and the  $PCA_{AUX}$  methods in the analysis of simulated and empirical data.

## **Acknowledgement**

Partial support for this project was provided by grant NSF 1053160 (Wei Wu & Todd D. Little, co-PIs), the Center for Research Methods and Data Analysis at the University of Kansas (Todd D. Little, director), grant IES R324C080011 (Charles Greenwood & Judith Carta, CoPIs), and a grant from the Society for Multivariate Experimental Psychology (SMEP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the funding agencies.

## Table of Contents

Abstract.....	iii
Acknowledgement .....	iv
Table of Contents.....	v
List of Tables .....	x
List of Figures.....	xi
Introduction.....	1
The Missing Data Problem .....	5
The Historical Context of Missing Data.....	7
The First Historical Period: Early Developments.....	9
Least squares methods for missing data.....	9
Maximum likelihood methods for missing data. ....	14
Introduction to maximum likelihood estimation with complete data. ....	16
An overview of log likelihood. ....	18
Using first derivatives to locate MLE.....	21
Introduction to maximum likelihood estimation with incomplete data. ....	23
MLE with bivariate complete data.....	23
MLE with bivariate incomplete data.....	27
Simpler systems of estimates in the presence of missing data.....	29
Pairwise deletion.....	30
Mean substitution.....	32
Missing information.....	33
Summary of early developments. ....	36

The Second Historical Period: Ignorable Missing Data and Estimation .....	37
The beginnings of a classification system. ....	37
Introduction of iterative algorithms. ....	38
Defining an iterative algorithm. ....	39
Introduction of the Expectation-Maximization (EM) algorithm. ....	40
Introduction of full-information maximum likelihood (FIML). ....	43
Summary of ignorable missing data and estimation. ....	46
The Third Historical Period – A Revolution in Missing Data .....	47
Theoretical overview of missing data mechanisms. ....	48
Defining missing data mechanisms .....	48
Missing completely at random (MCAR). ....	50
Missing at random (MAR). ....	53
Missing not at random (MNAR). ....	56
The MNAR to MAR continuum. ....	59
Missing data patterns. ....	61
Multiple imputation (MI). ....	62
The imputation step. ....	67
The analysis step. ....	70
The pooling step. ....	71
Fraction of missing information. ....	75
Relative increase in variance. ....	76

Auxiliary variables.....	78
Auxiliary variables for nonlinear causes of missingness.....	81
Auxiliary variable demonstration using simulated data. ....	82
Excluding a cause of missingness.....	84
Improving estimation with MNAR.....	85
Including non-linear MAR.....	85
Improving statistical power.....	86
Auxiliary variable simulation summary.....	87
Methods for choosing auxiliary variables.....	88
Finding a few influential auxiliary variables. ....	89
The appeal of a restrictive strategy. ....	91
Issues with the restrictive strategy. ....	93
A practical all-inclusive auxiliary variable strategy. ....	97
An Introduction to Principal Components Analysis .....	98
The Historical Context of Principle Component Analysis .....	99
Defining Principal Components Analysis.....	104
Eigenvectors.....	109
Finding eigenvectors.....	113
Illustration of principal component calculation. ....	119
Summary of principal component analysis.....	120
Method I: Simulation Studies .....	121
Design and Procedure .....	121

Convergence.....	121
Population Model.....	122
Data Generation.....	122
Analysis Models.....	123
Outcomes.....	124
Relative Performance.....	125
Population Model.....	127
Data Generation.....	128
Analysis Models.....	131
Outcomes.....	132
Parameter Bias.....	132
Relative Efficiency.....	133
Method II: Empirical Study.....	133
Procedure and variables.....	134
Empirical missing data.....	134
Outcomes.....	135
Descriptives.....	135
Fraction of missing information.....	135
Relative increase in variance.....	136
Simulation Results.....	136
Convergence.....	136
AUX strategy.....	136

PCA <sub>AUX</sub> Strategy.....	139
Linear MAR performance.....	140
Non-linear MAR performance.....	142
Relative efficiency.....	143
Empirical Example Results.....	143
Discussion.....	144
Convergence.....	146
Parameter Bias.....	148
Relative Efficiency.....	149
Empirical Example.....	150
Limitations.....	151
Conclusions.....	152
Implementing the PCA <sub>AUX</sub> Strategy.....	153
References.....	154
Appendix A: Marginal Means with Unbalanced Data.....	167
Appendix B: Yates Formula Example.....	169
Appendix C: Hartley and Hocking (1971) Illustration.....	173
Appendix D: Rubin’s (1976) Non-Ignorable Missing Data.....	175
Appendix E: Demonstration of the EM Algorithm.....	176
Appendix F: Discussion of FIML as a Direct ML Method.....	182
Appendix G: Discussion of Probability and Missing Data Mechanisms.....	184
Appendix H: Demonstration of data augmentation in MI.....	187
Appendix I: Syntax Guide to Using PCA Auxiliary Variables.....	192

## List of Tables

Table 1	<i>Data from an exemplar randomized block design taken from Allan and Wishart.</i>	197
Table 2	<i>Iterative solutions for a set of linear equations.</i>	198
Table 3	<i>Exemplar data illustrating MCAR, MAR and MNAR missing data mechanisms.</i>	199
Table 4	<i>Simulation results showing the impact of omitting a cause or correlate of missing</i>	201
Table 5	<i>Simulation results showing the influence of auxiliary variables to improve</i>	202
Table 6	<i>Simulation results showing the bias associated with ignoring a non-linear cause</i>	203
Table 7	<i>Simulated data for PCA examples.</i>	204
Table 8	<i>Monte Carlo simulation studies published in the social sciences on the topic</i>	205
Table 9	<i>Raw ECI key skills data in 3-month intervals from 9 – 36 months.</i>	206
Table 10	<i>Monte Carlo simulation results showing the impact of PCA<sub>AUX</sub> and AUX</i>	207
Table 11	<i>Monte Carlo simulation results showing the impact of PCA<sub>AUX</sub> and AUX</i>	208
Table 12	<i>Monte Carlo simulation results showing the impact of PCA<sub>AUX</sub> and AUX</i>	209
Table 13	<i>Relative Efficiency of correlation parameter estimate between X and Y</i>	210
Table 14	<i>Multiple imputation of ECI key skills data using no auxiliary variables.</i>	211
Table 15	<i>Multiple imputation of ECI key skills data using the AUX method</i>	212
Table 16	<i>Multiple Imputation of ECI key skills data using the PCA auxiliary variables.</i>	213
Table 17	<i>A Demonstration of marginal means estimation for the main effect of B.</i>	214
Table 18	<i>A Demonstration of marginal means estimation for the main effect of B.</i>	215

## List of Figures

Figure 1. The image provides a conceptual demonstration of research that leads to valid .....	216
Figure 2. A graphical representation of the relationship between a sampling density ( $n = 5$ ) ...	218
Figure 3. A plot of the likelihood values as a function of various population means .....	219
Figure 4. A plot of the log-likelihood values as a function of various population means .....	220
Figure 5. A plot of the log-likelihood values as a function of various population variance.....	221
Figure 6. A close up view of the peak of the likelihood function used to demonstrate MLE ....	222
Figure 7. A plot of the log-likelihood values as a function of various population means .....	223
Figure 8. A plot of the log-likelihood values as a function of various population means .....	224
Figure 9. A three-dimensional plot of a bivariate normal probability distribution.....	225
Figure 10. Illustration of Mahalanobis distance .....	226
Figure 11. Illustration of missing data patterns .....	227
Figure 12. A depiction of two three-dimensional plots of a bivariate normal probability .....	228
Figure 13. A graphical depiction of the log-likelihood values as a function.....	229
Figure 14. Illustration of an incomplete data matrix $Y$ where missingness is denoted .....	230
Figure 15. A modified version of Little's (2009) schematic representation of .....	231
Figure 16. A graphical representation of the MCAR missing data mechanism. ....	232
Figure 17. Univariate missing data pattern with missing on $Y_2$ but not on $Y_1$ .....	233
Figure 18. Diagram of the MAR missing data mechanism. ....	234
Figure 19. Diagram of the MNAR missing data mechanism.....	235
Figure 20. Diagram illustrating a conceptual overview of the missing data mechanisms.....	236
Figure 21. Diagram illustrating the relationship between MNAR and MAR as a continuum....	237
Figure 22. An illustration of common missing data patterns.....	238
Figure 23. A graphical demonstration of the Bayesian approach used by Rubin (1977) .....	239

Figure 24. A graphical depiction of Rubin’s (1978a) multiple imputation procedure ..... 240

Figure 25. An illustration of Rubin’s (1987) multiple imputation procedure..... 241

Figure 26. The Y and X vectors from the previous Yates method example..... 242

Figure 27. Trace plot for the simulated mean of Y..... 243

Figure 28. Illustration of a bivariate regression in the context of MI. .... 244

Figure 29. Illustration of a missing data pattern with missing on Y but not on X..... 245

Figure 30. A path diagram illustrating the range of population correlations..... 246

Figure 31. A plot of simulation results showing bias associated with the exclusion ..... 247

Figure 32. Simulation results showing the bias reduction associated with including ..... 248

Figure 33. Simulation results showing the relative power associated with ..... 249

Figure 34. Illustration of data reduction using PCA. .... 250

Figure 35. Scatterplot of the data in Table 10 (N = 25) with a correlation ellipse ..... 251

Figure 36. Illustration of the rotated scatterplot from Figure 37. .... 253

Figure 37. Illustration of data transformation process from the original data ..... 254

Figure 38. Illustration of vector  $\mathbf{u}_1$  as a trajectory in 2 dimensional space..... 255

Figure 39. Illustration of scalar multiplication of vector ..... 256

Figure 40. Geometric illustration of vector  $\mathbf{u}_1$  and vector  $\mathbf{u}_2$ ..... 257

Figure 41. Geometric illustration of the angle formed between the perpendicular vector ..... 258

Figure 42. Geometric representation of the angles of the rotated axes relative to..... 259

Figure 43. Geometric representation of the angles of the rotated axes relative to..... 260

Figure 44. Illustration of k principal components as new variables ..... 261

Figure 45. A graphical depiction of a likelihood function to depict convergence failure. .... 262

Figure 46. A graphical depiction of the data augmentation stability..... 263

Figure 47. Exemplar general missing data pattern with missing on all ..... 264

Figure 48. Illustration of the population model for the three homogeneity conditions. ....	265
Figure 49. Illustration of the missing data patterns investigated across X, Y and.....	266
Figure 50. Illustration of simulation results showing bias of the correlation .....	267
Figure 51. Illustration of simulation results showing the correlation between.....	268
Figure 52. Illustration of simulation results showing the correlation .....	269
Figure 53. The Y and X vectors for the Yates method example.....	270
Figure 54. A graphical depiction of joint events in a sample space .....	271
Figure 55. Mean plots of the ECI key skills from the empirical data example. ....	272

## Introduction

Most research in the social and behavioral sciences involves the analysis of data with missing information. Current research frequently presents numerous and often-sophisticated circumstances that preclude a researcher from obtaining complete data (Peugh & Enders, 2004; Schafer & Graham, 2002). Given that the majority of statistical methods theoretically (Little & Rubin, 2002) and computationally (Iacobucci, 1995) assume complete data, a primary objective is to handle the missing data in a way that will not hinder the researcher's ability to reach valid inferences.

Little and Rubin (1987) positioned the missing data problem within the more general context of quantitative methodology in the social and behavioral sciences, and discussed the implications of various missing data handling techniques. While there are many possible ways to treat missing data, they noted that most approaches are not recommended (e.g., long-standing traditional approaches like deletion and single imputation). Little and Rubin discussed Rubin's (1976) classification system for missing data emphasizing the importance of assumptions regarding why the data are missing as this can bias any inferences made from the data being studied. They also noted that in most incomplete datasets, observed values provide indirect information about the likely values of missing data. Numerous theoretical papers and simulation studies have demonstrated that this information, guided by statistical assumptions, can effectively recover missing data to the degree that the variables responsible for causing missing data are included in the missing data handling procedure (e.g., Allison, 2003; Baraldi & Enders, 2010; Collins, Schafer, & Kam, 2001; Enders, 2008; Enders & Bandalos, 2001; Graham, 2003).

Multiple imputation (MI), a missing data handling technique that replaces a missing value by drawing a random sample from a distribution of possible values a set number of times, and full-information maximum likelihood (FIML), a missing data handling technique that relies on a probability density function to iteratively maximize the likelihood of estimates in the

presence of missing values, are the most generally recommended techniques in the methodological literature because they enable researchers to account for missing data in a variety of conditions while maximizing statistical power and minimizing bias (e.g., Collins, Schafer, & Kam, 2001; Enders, 2010; Graham, 2009). However, the effectiveness of these statistical tools depends on the researcher's ability to meet the missing at random (MAR) assumption (Buhi, Goodson, & Neilands, 2008; Enders, 2010; Graham, 2009; Graham & Collins, 2011). That is, MI and FIML only yield unbiased parameter estimates when all variables that are causes or correlates of missingness are included in the missing data handling procedure (Enders, 2010; Little & Rubin, 2002). Methodologists have recommended auxiliary variables to address this issue. Auxiliary variables are not the focus of an analysis but are instead used to inform the missing data handling procedure (e.g., MI, FIML) by adding information about why a particular variable has missing data or describing the probability that a particular case has missing data (Collins, Schafer, & Kam, 2001; Enders, 2010). Therefore, auxiliary variables support the MAR assumption and improve estimation (Collins, et al., 2001; Enders & Peugh, 2004; Graham, 2003).

For the past two decades, missing data research has focused primarily on developing analytical and computational alternatives for traditional missing data handling techniques (e.g., listwise deletion, pairwise deletion, mean imputation, single regression imputation, etc.), implementing these procedures in easy-to-use software, and developing extensions to MI and FIML. These extensions include: multiple groups with missing data (e.g., Enders & Gottschall, 2011), categorical missing data (e.g., Allison, 2006), clustered data with missingness (e.g., Beunckens, Molenberghs, Thij, & Verbeke, 2007), incomplete non-normal data (e.g., Demirtas, Freels, & Yucel, 2008), and better approximating the MAR assumption by incorporating auxiliary variables (e.g., Collins et al., 2001; Schafer & Graham, 2002). While this list is not exhaustive, it illustrates efforts to address known limitations in MI and FIML.

This article addresses an important issue in handling missing data in large data sets. The issue arises when a large number of auxiliary variables are used to improve the quality of estimation in the presence of missing data. The methodological literature encourages an “inclusive strategy” where numerous auxiliary variables are used to reduce the chance of inadvertently omitting an important cause of missingness while reducing standard errors and gaining efficiency (Collins, Schafer, & Kam, 2001; Enders, 2010); however, the inclusive strategy can be difficult to implement in practice (Collins et al., 2001; Enders, 2010; Graham, 2009; Yoo, 2009).

More than a decade ago, Collins, Schafer, & Kam (2001) traced limitations of the inclusive strategy to statistical software and the associated documentation because “...neither of [these] encourages users to consider auxiliary variables or informs them about how to incorporate them,” (Collins et al., 2001, p. 349). Since then, much progress has been made on implementing the inclusive strategy in software packages as well as the accompanying documentation. For example, Muthén & Muthén (2011) recently integrated an auxiliary option in *Mplus* 6.0 to greatly simplify the otherwise complex implementation of auxiliary variables in full-information maximum likelihood (FIML) estimation (see Graham, 2003 for details).

However, despite these developments the number of auxiliary variables that can be feasibly included is limited (Graham, Cumsille & Shevock, 2013). The determination of this limit may relate to many dataset specific factors and is likely to vary across a specific set of variables. For instance, the amount of missing information, the number of variables in the model (i.e., complexity of the model) and the presence of high collinearity have been shown to influence the convergence of modern missing data handling techniques and may lead to estimation failure (Enders, 2010). Beyond this limit, FIML and even MI will fail to converge on an acceptable solution (Asparouhov & Muthén, 2010; Enders, 2002, 2010; Graham et al., 2013; Savalei & Bentler, 2009).

Although various guidelines to “fix” convergence failure abound (e.g., Asparouhov & Muthén, 2010; Enders, 2010; Graham et al., 2013), a practical recommendation has been to reduce the number of auxiliary variables used (Enders, 2010). As a result, researchers typically pursue a restrictive strategy, where only a few carefully selected auxiliary variables are employed (Collins, Schafer, & Kam, 2001; Enders, 2010). Large-scale projects can present a challenge because hundreds of potential auxiliary variables may provide information about why data are missing (i.e., many variables may be needed to reasonably meet the MAR assumption). The situation is complicated when the auxiliary variables themselves contain missing data (Enders, 2008) and when non-linear information (e.g., interaction and power terms) is incorporated (Collins, et al., 2001).

The methodological literature recognizes the inevitable uncertainty regarding the cause of missing data (e.g., Enders, 2010; Graham, 2012). To best approximate the MAR assumption, however, all potential causes of missingness in the data set should be incorporated (Baraldi & Enders, 2010; Enders, 2010; Schafer & Olsen, 1998). Moreover, researchers typically assume linear relationships when using MI or FIML (Collins, Schafer, & Kam, 2001; Enders, 2010); however, including product terms or powered terms in MI and FIML provides variables that approximate non-linear processes (Collins, et al., 2001; Enders & Gottschall, 2011; Graham, 2009). Typically, this non-linear information should be included in the missing data handling routine to help predict missingness.

A consequence of including all possible auxiliary variables (including non-linear information) is that some of the information may not be useful. That is, researchers can reach a point of “diminishing returns” where only a few auxiliary variables contain useful information (Graham, Cumsille, & Shevock, 2013). In this situation the addition of extra “junk” variables complicates the missing data handling model but is not harmful (see Collins, Shafer, & Kam, 2001). Because there is an upper limit on the number of auxiliary variables that can be included

in MI and FIML, the researcher is faced with a dilemma between implementing the inclusive strategy with regard to auxiliary variables and the practical limitations of current statistical software. Thus, adequately satisfying the MAR assumption, which is required for unbiased MI and FIML estimation, becomes a task for the research analyst, whom must select a subset of auxiliary variables that sufficiently capture the cause of missingness.

We present a practical solution that views auxiliary variables as a collection of useful information rather than a specific set of extra variables. The use of methods to consolidate large numbers of variables and incorporate non-linear information have come of age elsewhere in the social and behavioral sciences, and need to be incorporated more routinely into modern missing data handling procedures. Consequently, the current approach focuses on a method of using principal component analysis (PCA), a multivariate data reduction technique that involves the linear transformation of a large set of variables into a new, more parsimonious set of variables, to obtain auxiliary variables that inform missing data handling procedures.

The fundamental idea of PCA is to find (through an eigen decomposition) a set of  $k$  principal components ( $\text{component}_1, \text{component}_2, \dots, \text{component}_k$ ) that contain as much variance as possible from the original  $p$  variables ( $\text{variable}_1, \text{variable}_2, \dots, \text{variable}_p$ ), where  $k < p$ . The  $k$  principal components contain most of the important variance from the  $p$  original variables and can then replace them in further analyses (see Johnson & Wichern, 2002). Consequently, the PCA procedure reduces the original data set from  $n$  measurements on  $p$  variables to  $n$  measurements of  $k$  variables. When applied to all possible auxiliary variables (both linear and non-linear) in the original data set, a new smaller set of auxiliary variables are created (principal components) that contain all the useful variation present across all possible auxiliary variables. These new uncorrelated principal component variables are then used as auxiliary variables in an “*all-inclusive*” approach to best inform missing data handling procedure.

### **The Missing Data Problem**

Often researchers are unable to obtain all the information needed on individuals in a study. Participants may accidentally pass over questionnaire items, they may skip items that they are uncomfortable answering, they could run out of time, they might avoid confusing questions, the measurement tool may fail, or the participant may just get bored and stop answering questions. In the case of longitudinal studies, participants could move before data collection is complete; they might also get sick or be absent from a particular measurement occasion. In reality, any number of random and/or systematic circumstances could prevent a researcher from collecting data. Given that the majority of classical statistical methods assume complete data (Allison, 2001; Little & Rubin, 2002), it follows that a primary objective of the quantitative analyst is to handle the missing information in a way that will not hinder the researcher's ability to reach valid inferences.

Figure 1 provides a conceptual demonstration of a research scenario in which the researcher plans to collect data (planned data) however; the observed data contains missingness (cause of missingness; denoted by dotted lines). Missing data handling techniques are then used to obtain the complete data for subsequent statistical modeling. The straight arrow pointing to a rectangle represents a progression while the straight arrow pointing to another arrow represents an interaction. Note that the statistical inference implied from the planned data is equivalent to that implied from the complete data (denoted " $a$ ") while the regression path from the observed data (with missingness) denoted " $b$ " represents potential bias associated with deletion methods.

In order to provide a practical example, consider a simple research scenario using a simple two-group experimental design with a treatment and control group. Suppose researchers want to know how children perform on language outcomes when they are exposed to a new intervention. Therefore, researchers create two groups by randomly assigning participants to either a treatment group, that receives the new intervention, or a control group that is not exposed to the intervention. Randomization in this example should ensure that the treatment and control

groups only differ in their exposure to the intervention; thus any differences found by contrasting outcomes between the two groups can be attributed to the intervention (Hayes, 1994). In this manner, confounding factors are eliminated and the researchers can make unbiased casual statements about the effect of the intervention.

Consider that it is possible for this research scenario to generate missing values as children may not provide data on all of the variables of interest. Should this occur, missing data can be problematic for two basic reasons: (1) as Rubin (1976) noted, missing data can disrupt the randomization process such that the two groups no longer differ solely on the intervention, and (2) the loss of children from the sample can reduce statistical power to detect group differences (Buhi, Goodson, & Neilands, 2008; Schafer & Graham, 2002; Schafer & Olson, 1998). That is, missing data can introduce bias, prejudice in favor or against the treatment group compared to the control group, which may influence any assessment of the intervention's effect. The extent of this influence is related to characteristics of the missing data, such as the amount and complexity of missingness, as well as to the appropriate implementation of a missing data handling technique. The term "missing data handling" refers to a technique that corrects for missing data which encompasses either imputation-based methods, maximum likelihood-based methods, or deletion methods (following Enders, 2003). The phrase "missing data imputation" is avoided as a general term for these techniques because likelihood-based methods and deletion methods do not impute missing data.

The following discussion provides a historical account of research and theory on the problem of missing data and introduces the historical context and rationale for modern missing data handling procedures. As will be discussed, much of this work emphasized the importance of assumptions regarding why the data are missing as this can bias any inferences made from the data being studied.

### **The Historical Context of Missing Data**

Given the frequency of missing data in applied research and the current interest in appropriate missing data handling techniques, it is not surprising that many authors have discussed the history of the topic. The forthcoming discussion provides a historical perspective on the methods used to address missing data. It will become apparent that tremendous progress has been made on the technical problems associated with statistical methods for handling missing data. The statistical tools resulting from this methodological advancement carry their own assumptions and difficulty of use. We will later focus on one of these problems (convergence failure with too many auxiliary variables) and present a possible solution. The ensuing discussion contains a selective overview of foundational works and is not intended to provide a complete history of the topic.

The history of modern missing data methods in the social and behavioral sciences has generally been grouped into three (see Schafer, 1997) or four (see Little, 2010) historical periods that reflect different ways of thinking about and handling incomplete data. The following discussion will refer to three overlapping periods that differ slightly from those introduced by Little. The first and earliest period dates back to the early 20<sup>th</sup> century with the advent of many commonly used test statistics including the *t*-test (Student, 1908), analysis of variance (ANOVA; Fisher, 1921, 1924) and analysis of covariance (ANCOVA; Fisher, 1932) and can be thought to conclude sometime in the late 1970s and early 1980s. Researchers during this historical period were largely concerned with the impact of missing data on statistical methods that assume complete data and publications from this period were typically marked by complete case (deletion methods) and single imputation approaches (e.g., mean substitution, single-regression imputation). The second period seems to emerge with the advent and availability of modern computers in the late 1970s and early 1980s. This historical period can be thought to persist for less than a decade but is distinct because many publications during this time seem to turn from a more practical approach to addressing incomplete data (e.g., deletion) to more theoretical

strategies (e.g., model-based procedures; maximum likelihood, expectation-maximization algorithm). Perhaps it was during this historical period that missing data first emerged as a field of study within the social and behavioral sciences (see Molenberghs & Kenward, 2007). In the third period, which arose in the late 1980s and early 1990s, specialized computer packages for handling missing data materialized and several seminal publications popularized a “revolution in thinking about missing data,” (Graham, 2009, p, 550). These developments made “modern” missing data procedures possible for the typical researcher and discouraged the routine use of traditional methods (Little, 2010). Over the past three decades, much research in the area of missing data has focused on a continuing effort to discourage the use of over-simplified “ad hoc” missing data handling practices, has developed theoretical and computational extensions to modern methods, and has implemented these procedures in popular statistical software packages.

### **The First Historical Period: Early Developments**

**Least squares methods for missing data.** The first published account of an experiment that considered the problem of missing observations was likely that of Allan and Wishart (1930) in *A Method of Estimating the Yield of a Missing Plot in Field Experimental Work*. This work was published only five years after the first edition of Fisher’s 1925 book titled “Statistical Methods for Research Workers”, which fully explained the analysis of variance procedure (earlier editions were less complete; see Cowles, 2001) and discussed the importance of randomization (Dodge, 1985; Hunt & Triggs, 1989). Two aspects of Allan and Wishart’s paper are of historical interest. In this article, they discuss an agricultural study of hay yields that incorporated a Latin square research design that suffered from missing data (e.g., cows had broken through a fence and consumed one of the experimental plots). At the time, the computational techniques for experimental designs, including the Latin square, required complete data (i.e., balanced data; Dodge, 1985; Little & Rubin, 2002). Therefore, the typical approach was to, “...delete the whole block containing the missing plot, or the row or column of

the Latin square...,” (Allan & Wishart, p. 399-400). In this manner, standard calculations were applied to the remaining complete data (i.e., complete case analysis). Yet, this is not how Allan and Wishart proceeded. Interestingly, rather than deleting data to balance the design, they suggested, “...what is needed is a means of utilizing all the known plot values to form a best estimate of the missing yield,” (p. 400). In doing so, they introduced formulae for the replacement of a single missing observation in a Latin square design (as well as in a randomized block design). This technique introduced the concept of “filling in” missing values based on observed data, an idea that is regarded as the earliest version of missing data imputation (Dodge, 1985; Little & Rubin, 2002).

Consider the example data given in Table 1 adapted from Allan and Wishart (1930, p. 402). Note that the data in Table 1 contain missingness on in Treatment 4 of Block B (denoted “–”). To replace this missing value with an estimate, Allan and Wishart (1930) implemented the following formula:

$$k = \frac{(n + s - 1)S - s(S_t) - n(S_b)}{(n - 1)(s - 1)} = \frac{16(1417.48) - 9(1284.16) - 8(1243.11)}{(9 - 1)(8 - 1)} = 20.97 \quad (1)$$

where  $n$  is the number of blocks,  $s$  is the number of treatments,  $S$  is the total observed values excluding those in the same line and column as the missing value,  $S_t$  represents the sum of all values of the treatments excluding the treatment that contains missingness (e.g., Treatment 4), and  $S_b$  represents the sum of all values of the blocks excluding the treatment that contains missingness (i.e., Block B). As demonstrated in Equation 1, 20.97 would replace the missing value in Table 1 and the ANOVA formula can be applied to the data.

A second noteworthy feature is that Allan and Wishart discussed the use of missing data techniques to improve data quality in situations where data were actually observed. For instance, they perceived that a particular plot that was located close to a dirt road seemed to suffer from traffic that generated dust. They suggested it was reasonable to intentionally delete the data

corresponding to this plot and replace it with an estimate; what the underperforming plot would have yielded had it been located elsewhere in the field. While this point may seem subtle, Allan and Wishart made the connection that missing data are not simply accidental occurrences that indicate careless experimental design, which was a common assessment at the time (see Rubin, 1976). Rather, they suggested that missing data should be expected and knowledge of why data are missing or why data are of poor quality may allow for corrective procedures that recover the lost information. This point is notable because these ideas represent a modern way of thinking about missing data that would not fully emerge in the literature for nearly five decades.

R. A. Fisher, whom was at the forefront of experimental design at the time, appears to have had little to say about missing data. His first remarks regarding missingness may have appeared in the third edition of “Statistical Methods for Research Workers”, published in 1930. In this edition, he introduced a few paragraphs regarding “fragmentary data” which he referred to as “extremely troublesome” and best avoided (Fisher, 1930). While Fisher failed to provide a statistical solution to missing data (see Brandt, 1932), it is reasonable that he may have influenced Allan and Wishart’s work.

One approach to understand why Fisher designed the analysis of variance (ANOVA) procedure assuming that the researcher would always have complete data is to appreciate that missing data procedures were not a principal concern at the time. Rather, the concerns were with the ideas and statistical groundwork for a viable alternative to multiple regression. Cohen, Cohen, West and Aiken (2003) put their finger on these issues noting, “...multiple regression was often computationally intractable in the precomputer era: computations that would take milliseconds by computer required weeks or even months to do by hand. This led Fisher to develop the computationally simpler, equal (or proportional) sample size ANOVA/ANCOVA model, which is particularly applicable to planned experiments,” (p. 4). Fisher was devoted to keeping the hand calculations simple so the ANOVA method would be popular among applied

researchers (see Conniffée, 1991). In order to accomplish this, Fisher's ANOVA calculations were built around balanced data (i.e., data containing no missing values and an equal number of participants per condition; Dodge, 1985; Little & Rubin, 1987).

When data are unbalanced, Fisher's calculations to partition sources of variance can alter the actual hypotheses tested by generating non-orthogonal effects (i.e., marginal means that contain information from other model parameters), which alter the sum of squares estimates and bias the resulting *F*-statistic (see Dodge, 1985; Iacobucci, 1995). As Little and Rubin (2002) note, "the computational problem is that the specialized formulas and computing routines used with complete *Y* [complete data] cannot be used, since the original balance is no longer present," (p. 27). This concept can be clarified by reviewing the calculations of marginal means from an exemplar ANOVA design. As demonstrated in Appendix A, the marginal means of unbalanced data are non-orthogonal and the sums of squares computed from these means are "contaminated" with functions of other parameters (see Iacobucci, 1995; Shaw & Mitchell-Olds, 1993).

Conniffée (1991) suggests that Fisher was often criticized for presenting clever methodological ideas but failing to include sufficient details, which he often left to others to work out; a view that was reflected by those who admired and worked closely with Fisher (see reflection in Yates, 1968). As Yates and Mather (1963) note, "...it was part of Fisher's strength that he did not believe in delaying the introduction of a new method until every *i* was dotted and every *t* crossed," (p. 110). Herr (1986) suggests this was the case with "unbalanced" designs due to missing data. For instance, Herr described an account in 1931 where a graduate student named Bernice Brown used ANOVA for a rodent weight study and ended up with an unbalanced design. As there were no formulae for unbalanced designs at the time, she proceeded with the calculations laid out in Fisher's book "Statistical Methods for Research Workers" (see references in Brown, 1932) and obtained a negative sum of squares estimate for an interaction term. Rather than balancing the design by deleting cases (which would have severely limited her sample size),

Brown consulted Brandt, her academic advisor at the time, who then presented the problem to Fisher (see Brown, 1932). According to Brandt (1932), a solution was provided by Fisher that involved an adjustment so that, "...the interior 2-way means (i.e., the cell means) should differ by a constant amount," (p. 168). Fisher's adjustment was an early version of Type II sums of squares and as such was not ideal because it assumed a null interaction term between the main effect of time and rat body weight (see Brandt, 1932).

Herr (1986) submitted that Fisher realized his solution was not ideal and shortly after meeting with Brandt, "...set Yates on the scent," to work out a more practical solution to unbalanced designs (p. 266). At the time, Frank Yates was working as an assistant statistician at Rothamsted Experimental Station, an agricultural research institution, under the direction of Fisher (see Healy, 1995). Regarding his motivation for the topic, Yates (1968) would later remark, "...here again it was erroneous use of confounding in actual experiments that directed attention [from Fisher] to the need for orthogonally in the corresponding analysis of variance," (p. 464). Regardless of the initial inspiration, Yates published an influential article titled, *The Analysis of Replicated Experiments When the Field Results Are Incomplete*, two years after Fisher and Brandt met. This publication provided much needed guidance to the field regarding missingness in experimental designs and was essentially a refinement of the earlier work by Allan and Wishart (1930).

The benefit of Yates' approach over that of Allan and Wishart (1930) relates to situations with more than one missing value. Specifically, Yates (1933) incorporated an iterative (by hand calculation) least squares estimation process, which was an early precursor of the EM algorithm (Dodge, 1985; Healy, 1995). As Yates directed, researchers with multiple missing values should: (a) insert guesses for all but one missing value, (b) apply the "Yates formula" to solve for one missing value at a time, and then (c) repeat the process for all missing values where each set of updated estimates (i.e., each iteration) provide smaller sum of squared error estimates until

subsequent changes are negligible. For further insights into the parallels between Allan and Wishart's formula and that of Yates, consider the illustrative formulae and least squares estimation example, provided in Appendix B.

As demonstrated in Appendix B, least squares based imputation techniques were a feasible approach to missing data handling early in the 20<sup>th</sup> century when these procedures were carried out by hand calculations. Despite the lack of modern computers, researchers invested much effort in the refinement of these estimation routines for missing data. For instance, Yates (1936) furthered least squares estimation techniques for situations with a complete row or a complete column of data missing. Bartlett (1937) developed his own two-step least squares technique called Bartlett's ANCOVA, which was commonly used at the time (see Dodge, 1985). Also, Yates and Hale (1939) considered least squares for a situation with two or more complete rows or columns missing and Cornish (1944) introduced simultaneous linear equations to least-square estimation techniques.

While regression-based single imputation techniques remained popular among applied researchers well into the late 1980's (Rubin, 1987), many of these techniques were designed for specific situations (e.g., specific patterns of missingness, specific analytic designs, etc.) and more generalized approaches were often too complex for applied researchers to effectively implement (see Dodge, 1985 for discussion). Regardless of these factors, each regression-based technique required knowledge of background values (e.g., other treatment and block conditions) that were, in one way or another, entered into the least squares prediction equation (Cochran, 1983). In this way, imputation was carried out by borrowing information from observed data, an idea that will continue throughout the following discussion.

**Maximum likelihood methods for missing data.** During the same time that researchers advanced missing data methods using least squares for randomized experiments, incomplete data methods related to maximum likelihood were also emerging in relation to observational studies.

In 1932, Samuel S. Wilks introduced maximum likelihood for missing data and "...other less efficient, but simpler systems of estimate," (p. 164). His article titled, *Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples*, demonstrates a combination of quantitative methodology relating to maximum likelihood estimation (an idea first presented with complete data by Fisher between 1912 and 1922; see Hald, 1999), and the problem of missing data. Though he did not include data examples to demonstrate his technique, Wilks' effort provides the statistical foundation for explaining the contribution of information among a set of variables with missingness to the likelihood function. In doing so, he outlined the future potential for maximum likelihood techniques with missing data. However, realizing the impracticality of his maximum likelihood approach among applied researchers at the time (e.g., convergence was not considered in this paper because the computational demands were beyond the available technology), Wilks devoted the second half of his paper to the now infamous missing data techniques: mean substitution and pairwise deletion.

The primary purpose of the following discussion includes two points. First, it recognizes an early innovator in missing data research, demonstrates the idea behind maximum likelihood estimation (MLE), and provides an example of this procedure. The second purpose is to highlight that in the early 1930s, maximum likelihood with missing data represented an idea that was often unattainable in practical applications. More commonly, researchers had to settle for a more informal approach. Therefore, mean substitution, in which a missing value is replaced with an average (e.g., mean of the observed values for that variable), and pairwise deletion, wherein each population parameter is estimated based on some but not all cases, will be discussed.

To begin with, consider that early in the history of missing data research the vast amount of work originated from randomized experiments (especially in the area of agriculture) and Wilks (1932) represents an important exception. His approach began with missing data, "...arising from incompletely answered questionnaires," (p. 164). That is, Wilks' point of view

also marks a divergence from emphasizing the summation of experimental block rows and treatment columns in missing data handling using least squares estimation to a focus on the probability of the observed data as a random sample from an unknown population using MLE. Said differently, Wilks used MLE to address missing values by taking advantage of the concept that observed values can be used to identify the population that is most likely to have generated the sample. Thus, it is possible to summarize that population (e.g., with sufficient statistics) without actually deleting data or imputing missing values. The following is an example of the method of maximum likelihood with a single variable and one unknown population parameter. Following this example, an explanation of Wilks' theoretical application of these ideas to bivariate normal data with missingness will be presented.

***Introduction to maximum likelihood estimation with complete data.*** To illustrate how maximum likelihood estimates (MLE) are derived, let  $\mathbf{Y}$  represent a  $5 \times 1$  column vector of randomly sampled data from a population with a normal distribution:

$$\mathbf{Y} = \begin{bmatrix} -.48 \\ -.11 \\ -.82 \\ .73 \\ .94 \end{bmatrix} \quad \text{where } \mathbf{Y} \sim N(\mu_Y, \sigma_Y^2) \quad (2)$$

Suppose the variance of the population is known to be  $\sigma_Y^2 = 1.5$  but the mean is unknown. In this example (adapted from Kutner, Nachtsheim, Neter, & Li, 2005), the goal is to determine  $\mu_Y$  that is most consistent with the sample data vector  $\mathbf{Y}$ . For didactic purposes consider auditioning the following two possible values for the population mean  $\mu_Y = 0.0$  and  $\mu_Y = 1.0$ . The objective of MLE is to use the sample data to determine which guess is most likely. Figure 2 illustrates these two population means and the locations of the sample data in vector  $\mathbf{Y}$ . Panel A in Figure 2 depicts  $\mu_Y = 0$  while Panel B in Figure 2 depicts  $\mu_Y = 1.0$ . Notice that the population estimate  $\mu =$

0.0 is more realistic than population estimate  $\mu = 1.0$  given the observed random sample. That is, the height of the curve at each value of  $\mathbf{Y}$  (i.e.,  $Y_1, \dots, Y_5$ ) reflects the density of the probability distribution at that value. For example, in Panel A of Figure 2 the variable  $Y_3$  has a higher density than  $Y_3$  in Panel B of Figure 2 (i.e., the tail of the distribution is less likely than nearer to the center; Hays, 1994).

While Figure 2 provides a clear visual contrast between the two possible values for the population mean (e.g.,  $\mu_Y = 0.0$  and  $\mu_Y = 1.0$ ), a more practical assessment is provided by the density function for a normal distribution, which can be written as:

$$f(\mathbf{Y}) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-.5\frac{(Y_i - \mu_Y)^2}{\sigma_Y^2}} \quad (3)$$

where  $Y_i$  is a particular value from the column vector  $\mathbf{Y}$ ,  $\mu_Y$  is the population mean, and  $\sigma_Y^2$  is the population variance (Enders, 2010; Kutner, Nachtsheim, Neter, & Li, 2005). Applying this formula to the observed data in vector  $\mathbf{Y}$  it is possible to derive estimates of density of the probability distribution for each value as follows:

$$\begin{array}{cc} \mu_Y = 0.0 & \mu_Y = 1.0 \\ f(Y_1) = \frac{1}{\sqrt{2\pi 1.5}} e^{-.5\frac{(-.48-0.0)^2}{1.5}} & f(Y_1) = \frac{1}{\sqrt{2\pi 1.5}} e^{-.5\frac{(-.48-1.0)^2}{1.5}} \\ \vdots & \vdots \\ f(Y_5) = \frac{1}{\sqrt{2\pi 1.5}} e^{-.5\frac{(.94-0.0)^2}{1.5}} & f(Y_5) = \frac{1}{\sqrt{2\pi 1.5}} e^{-.5\frac{(.94-1.0)^2}{1.5}} \end{array} \quad (4)$$

The probability density (i.e., height of the normal curve) given  $\mu_Y = 0.0$  are as follows:  $Y_1 = .356$ ,  $Y_2 = .397$ ,  $Y_3 = .285$ ,  $Y_4 = .306$ ,  $Y_5 = .256$ . Notice that the value with the highest density value is  $Y_2$ , which was also the closest value to the center of the distribution in Panel A of Figure

2. Likewise, the probability density given  $\mu_Y = 1.0$  are as follows:  $Y_1 = .133$ ,  $Y_2 = .215$ ,  $Y_3 = .076$ ,  $Y_4 = .385$ ,  $Y_5 = .398$ . Notice that in this situation the value with the highest density value is  $Y_5$ , which was also the closest value to the center of the distribution in Panel B of Figure 2. Also, observe that  $Y_5$  in Panel B is higher than any single value in Panel A. That is,  $Y_5$  in Panel B strongly suggests that the correct population mean is  $\mu_Y = 1.0$ . However, the rest of the sample, especially  $Y_1 - Y_3$ , in Panel B does not make such a compelling case. The method of MLE searches for population parameters that are most realistic given all of the observed data so the product of the individual density estimates are used to generate an overall assessment of the greatest population density, called a likelihood value (Kutner, Nachtsheim, Neter, & Li, 2005). Mathematically, the sample likelihood value can be written as:

$$L = \prod_{i=1}^N \left\{ \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-.5\frac{(Y_i - \mu_Y)^2}{\sigma_Y^2}} \right\} \quad (5)$$

where the L represents the likelihood value, and  $\Pi$  indicates the multiplication of all density estimates from  $f(Y_1), \dots, f(Y_N)$ . The likelihood value for  $\mu_Y = 0.0$  is  $L = 0.00315$  and for  $\mu_Y = 1.0$  is  $L = 0.00034$  which indicates that the maximum likelihood (i.e., most area under the normal curve) is related to the prior parameter estimate (i.e.,  $\mu_Y = 0.0$ ). Said differently, the most likely population mean for the vector  $\mathbf{Y}$  is  $\mu_Y = 0.0$ . Figure 3 shows a more detailed plot of the likelihood value as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$ . Notice that  $L(\mu_Y = 0.0)$  is at the peak of the likelihood function.

*An overview of log likelihood.* In practice the multiplication of individual density estimates can generate extremely small numbers that are prone to rounding error (e.g., a random sample size of  $N = 30$  in the current example would generate a likelihood value of  $L = 1.44e^{-20}$  or



covariance). To illustrate let the population mean be  $\mu = 0.0$  for the data in vector  $\mathbf{Y}$ , as was suggested by the previous example, and consider substituting some guesses for the correct population variance  $\sigma_Y^2$ . Using the same MLE equations previously outlined, the log-likelihood function for the population variance can be generated (see Figure 5).

While the previous maximum likelihood estimation of  $\mu_Y$  and  $\sigma_Y^2$  were demonstrated separately, in practice both would be estimated at the same time (in addition to covariances when there are more than one variable). Consequently, a trial-and-error approach to finding maximum likelihood values for  $\mu_Y$  and  $\sigma_Y^2$  can become a tedious process (Enders, 2010). For example, the maximum likelihood variance estimate given the data was  $\sigma_Y^2 = 0.48$  given the population mean estimated in the first example (which was itself based on a population variance of  $\sigma_Y^2 = 1.50$ ). Considering Equation 6, this difference directly influences the estimate  $\mu_Y$ , which in turn influences the resulting estimates of  $\sigma_Y^2$ . Said differently, each guess made regarding a particular population parameter affects the other parameter estimates.

Further, consider the flatter shape of the likelihood function plot for the population variance  $\sigma_Y^2$  (see Figure 5) in relation to the likelihood function plot for the mean (see Figure 4). Visually, this difference illustrates that the likelihood function may not be peaked and the resulting MLE solution is not always straightforward. For instance, consider that the MLE solution for the population variance may range from  $\sigma_Y^2 = 0.36 - 0.60$  as any of these values may seem as likely as  $\sigma_Y^2 = 0.48$  (the MLE). That is when the shape of the likelihood function is fairly flat many population values are possible and the MLE may be relatively hard to locate. To further this point, consider the image in Figure 6 that zooms in on the peak of the likelihood function from Figure 5 to show the relatively flat shape of the peak. Figure 6 illustrates an additional challenge for MLE as a trial-and-error approach to finding maximum likelihood

values. Specifically, how to locate a specific MLE estimate when numerous values of the likelihood function seem plausible. As discussed next, calculus can be used to solve these problems by finding the maximum likelihood value via a closed-form mathematical function for derivatives (Enders, 2010).

*Using first derivatives to locate MLE.* Rather than using a trial-and-error approach to audition different values to locate the MLE, a much more efficient and practical approach is to use first derivatives (Enders, 2010). To introduce this concept, consider the Figure 7 that provides an illustration of log-likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$  (taken from Figure 4). This log-likelihood function plot is divided into three different sections (divided by dotted parallel lines and denoted as “A”, “B”, or “C”). As Enders (2010) noted, the first derivative is the slope of a line that is tangent to a value on the likelihood function curve. Therefore, section A of Figure 7 represents a region where all tangent lines (first derivatives) are positive. Here the slopes are increasing (e.g.,  $\mu_Y = -2.0$ ). Likewise, Section C symbolizes a region in which all tangent lines are negative. In this section the slopes are decreasing (e.g.,  $\mu_Y = 2.0$ ). Section B of Figure 7, in the middle of the likelihood function, signifies an area that contains the maximum likelihood value (associated with  $\mu_Y = 0.0$ ) where the tangent line is flat and the slope is zero. That is, the first derivative of the maximum likelihood value equals zero. Thus, as Enders noted, “...set the result of the derivative formula to zero and solve for the unknown parameter value,” (p. 63). This process includes the following sequence: (1) generate an equation for the first derivative of the log-likelihood function with respect to  $\mu_Y$  and  $\sigma_Y^2$ , (2) find the maximum of the likelihood function set both equations equal to zero (i.e., where the likelihood function slope is flat) and lastly (3) solve for  $\mu_Y$  and  $\sigma_Y^2$ , respectively, which can be written as:

$$\begin{aligned}
\frac{\partial \log L}{\partial \mu_Y} &= \frac{1}{\sigma_Y^2} \left( -N\mu_Y + \sum_{i=1}^N Y_i \right) \\
0 &= \frac{1}{\sigma_Y^2} \left( -N\mu_Y + \sum_{i=1}^N Y_i \right) \\
\hat{\mu}_Y &= \frac{1}{N} \sum_{i=1}^N Y_i
\end{aligned} \tag{7}$$

for  $\mu_Y$  where  $\partial$  denotes the derivative and as:

$$\begin{aligned}
\frac{\partial \log L}{\partial \sigma_Y^2} &= -\frac{N}{2\sigma_Y^2} + \sum_{i=1}^N \left\{ \frac{(Y_i - \mu_Y)^2}{2\sigma_Y^4} \right\} \\
0 &= -\frac{N}{2\sigma_Y^2} + \sum_{i=1}^N \left\{ \frac{(Y_i - \mu_Y)^2}{2\sigma_Y^4} \right\} \\
\hat{\sigma}_Y^2 &= \frac{1}{N} \sum_{i=1}^N (Y_i - \mu_Y)^2
\end{aligned} \tag{8}$$

for  $\sigma_Y^2$  (see Enders, 2010).

To develop the relationship between the likelihood function and the first derivatives, consider a more detailed illustration of this idea (see Figure 8). The plots in Figure 8 explicitly illustrate the first derivatives of  $\mu_Y = -2.0$ ,  $\mu_Y = 0.0$ , and  $\mu_Y = 2.0$  (see bottom panel) in relation to the likelihood function (see top panel). These Panels are aligned to intentionally show the relationship of the slope estimate (i.e., the first derivative) to the likelihood function. Note that as the regression line in the bottom panel of Figure 8 transitions from positive values to negative values it crosses the point of maximum likelihood (which lies on the slope = 0 reference line). This point corresponds with the highest point of the likelihood function in the top panel of Figure 8 (i.e., most likely value where the first derivative is as close to zero as possible).

Specifically, Section A of Figure 8 denotes an area of the likelihood function where all first derivatives are positive. Notice that the slope of a line tangent to the point on the likelihood function directly above  $\mu_Y = -2.0$  is positive (i.e.,  $\beta_{\mu(-2.0)} = 6.49$ ). Similarly, Section C of Figure 8 indicates a section in which all first derivatives are negative. Likewise, the resultant slope of a line tangent to the point corresponding to  $\mu_Y = 2.0$  is negative (i.e.,  $\beta_{\mu(2.0)} = -6.49$ ). Lastly, Section B of Figure 8 denotes an area that contains the maximum likelihood value that is correctly located by finding a first derivative as close to zero as possible.

***Introduction to maximum likelihood estimation with incomplete data.*** Given the previous discussion of maximum likelihood estimation (MLE) in relation to the  $5 \times 1$  column vector  $\mathbf{Y}$ , let the following discussion follow Wilks (1932) by moving forward with a bivariate example with complete data and then on to a bivariate example with incomplete data.

*MLE with bivariate complete data.* To illustrate MLE with complete data following the example set out by Wilks (1932), assume that the  $p \times 1$  column vectors  $\mathbf{X}$  and  $\mathbf{Y}$  are a random sample of size  $p$  from a bivariate normal distribution:

$$\mathbf{X} = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{p1} \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ \vdots \\ y_{p1} \end{bmatrix} \quad \begin{array}{l} \mathbf{X} \sim N(\mu_X, \sigma_X^2) \\ \mathbf{Y} \sim N(\mu_Y, \sigma_Y^2) \end{array}, \quad \mathbf{XY} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right] \quad (9)$$

where that  $\mu$  represents the mean,  $\sigma^2$  denotes the variance, and  $\sigma_{XY}$  is the covariance (Kutner, Nachtsheim, Neter, and Li, 2005; Raghunathan, 2004). While the current variables are denoted  $\mathbf{X}$  and  $\mathbf{Y}$  in keeping with Wilks' original example, note that both variables play a balanced role in the following discussion (i.e., the column vectors  $\mathbf{X}$  and  $\mathbf{Y}$  would perhaps be more clearly denoted as matrix  $\mathbf{Y}$  with vectors  $Y_{p1}$  and  $Y_{p2}$ ).

Recall that the goal of MLE is to identify the population that is most likely to have generated the sample data  $\mathbf{X}$  and  $\mathbf{Y}$ . This unknown population is found by means of a probability distribution, which provides a reference for the likelihood of various guesses for the correct population. Each guess to find the correct population takes the form of population parameters (i.e., values that define the shape and location of the probability distribution in multivariate space; Kutner, et al., 2005). Population parameters generate a particular probability distribution that has a given likelihood to have generated the observed sample. To demonstrate, let the random variables,  $\mathbf{X}$  and  $\mathbf{Y}$ , have a bivariate normal distribution (as suggested in the previous equation) so their joint *probability density function* (PDF; see Kutner, Nachtsheim, Neter, and Li, 2005) can be written as:

$$f(X, Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\sigma_{XY}^2}} e^{\left\{ -\frac{1}{2(1-\sigma_{XY}^2)} \left[ \left( \frac{X_i - \mu_X}{\sigma_X} \right)^2 - 2\sigma_{XY} \left( \frac{X_i - \mu_X}{\sigma_X} \right) \left( \frac{Y_i - \mu_Y}{\sigma_Y} \right) + \left( \frac{Y_i - \mu_Y}{\sigma_Y} \right)^2 \right] \right\}} \quad (10)$$

where the density function contains the five parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY}$  defined as they were in Equation 9). Let Figure 9 illustrate a bivariate density function for the previous equation. In Figure 9 notice that the probability distribution between  $\mathbf{X}$  and  $\mathbf{Y}$  can be shown as a plane surface in three-dimensional space with a height corresponding to the density of function  $f(X, Y)$  for every pair of X,Y values (Kutner, et al., 2005). The probability distribution surface is continuous and the probability corresponds to volume under the surface (i.e., height of the curve) which sums to one (Kutner, et al., 2005). Specifically, Figure 9 illustrates the point that some data values for  $\mathbf{X}$  and  $\mathbf{Y}$  are more probable than others.

As before, the likelihood function for this distribution can be expressed as  $L_i =$  likelihood function = function (data, parameters). That is,  $L_i$  indicates the likelihood of the  $i^{th}$  observed  $\mathbf{XY}$  variable combination, given a set of values for the model parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY}$ .

Regardless of the number of variables, the method of ML estimation uses the same concept: given the observed data, obtain the values of the population parameters that maximize the value of  $L_i$  (Enders, 2010; Kutner et al., 2005). This provides the ML estimates of the parameters (i.e., the parameter values that maximize the likelihood space of the observed data under the bivariate normal assumption of the population). Since  $\mathbf{X}$  and  $\mathbf{Y}$  follow a bivariate normal distribution, the likelihood (density) function can be written as:

$$L_i = \underbrace{\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}}}_{\text{Scaling Factor}} \underbrace{e^{-.5(\mathbf{XY}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{XY}_i - \boldsymbol{\mu})}}_{\text{Mahalanobis Distance}} \quad (11)$$

where  $\mathbf{XY}_i$  is a  $p \times 2$  matrix of observed values on  $\mathbf{X}$  and  $\mathbf{Y}$ ,  $\boldsymbol{\mu}$  is a  $2 \times 1$  column vector of means for  $\mathbf{X}$  and  $\mathbf{Y}$ , and  $\Sigma$  is a  $2 \times 2$  covariance matrix (replacing  $\sigma_x^2, \sigma_y^2, \sigma_{xy}$  in Equation 10).

To explain how estimates of probability density are derived for a bivariate normal distribution, consider Ender's (2010) description that breaks Equation 11 into two parts. As Enders (2010) notes, the likelihood function consists of (1) a scaling factor, which makes the area under the bivariate normal distribution curve sum to 1.0 and (2) Mahalanobis distance, a standardized measure of the distance between a particular value and the center of a multivariate distribution. Figure 10 provides an illustration of Mahalanobis distance (denoted by the correlation ellipse) in a three-dimensional scatter plot where point A and point B are the same distance from point C, the center of the multivariate distribution (multivariate centroid).

Given responses on a particular set of variables (e.g., scatter plot dots in Figure 10); the bivariate likelihood function can be explicitly illustrated as:

$$\mathbf{L} = \prod_{i=1}^N \left\{ L_i = \frac{1}{(2\pi)^{2/2} \begin{vmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_Y^2 \end{vmatrix}^{1/2}} e^{-.5 \begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}}^T \begin{vmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_Y^2 \end{vmatrix}^{-1} \begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}} \right\} \quad (12)$$

where  $L_i$  is the joint likelihood for a sample of observations (e.g.,  $X_i$  and  $Y_i$ ). Said differently,  $L$  is the relative probability of drawing values of  $X_i$  and  $Y_i$  from a bivariate normal distribution given the  $\boldsymbol{\mu}$  vector and  $\boldsymbol{\Sigma}$  covariance matrix (Enders, 2010). Let  $L$  designate the multiplication of each  $L_i$  to obtain a likelihood fit measure for the entire sample (i.e., an indicator of how well  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  “fit” the data). Computationally, the estimation procedure chooses estimates of the population parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  (which are unknown) from the observed data matrix  $\mathbf{XY}$  so that the likelihood function  $L$  is maximized (hence the term “maximum likelihood”). That is, the most likely values for the  $2 \times 1$  column vector of means ( $\boldsymbol{\mu}$ ) and the  $2 \times 2$  covariance matrix ( $\boldsymbol{\Sigma}$ ) are chosen for the observed variables  $\mathbf{X}$  and  $\mathbf{Y}$ .

Rather than maximizing the likelihood function, Wilks (1932) noted that it is more convenient to work with an alternative function that is inversely related to the likelihood function (i.e., the log-likelihood; as previously demonstrated in the univariate case). This alternative function is given by:

$$\log L = \sum_{i=1}^N \log \left\{ L_i = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-.5(\mathbf{XY}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{XY}_i - \boldsymbol{\mu})} \right\} \quad (13)$$

where “log” represents the natural logarithm. Notice that this equation replaces the previous matrices with characters representing those matrices (e.g., the covariance matrix was substituted for  $\boldsymbol{\Sigma}$ ). Specifically, the multivariate likelihood function (i.e., the bivariate likelihood function in the current example) is a direct extension of the principles demonstrated in the univariate case.

Here specific values, like a variable's mean ( $X_i$ ), are replaced with a matrix containing those values for more than one variable (e.g.,  $\mathbf{X}$ ).

As demonstrated in the univariate MLE example, the derivative formula is used to solve for the unknown parameter values. Specifically, an equation is produced for the first derivative of the log-likelihood function with respect to  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ . Then, the maximum of the likelihood function is set equal to zero. Finally, the equation is solved for to  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ .

*MLE with bivariate incomplete data.* Now, consider the estimation of MLE in a situation with missing (incomplete) data. Figure 11 represents situations where missing data may occur across the bivariate vectors  $\mathbf{X}$  and  $\mathbf{Y}$ . Following Wilks' (1932) original notation, Figure 11 demonstrates three distinct patterns of missingness where data is missing on  $\mathbf{X}$ , where data is missing on  $\mathbf{Y}$ , or where no data is missing (this example does not consider situations where data is missing on both  $\mathbf{X}$  and  $\mathbf{Y}$ ). Specifically, Figure 11 illustrates that  $n$  is the number of  $i$  cases observed on  $\mathbf{Y}$  but not on  $\mathbf{X}$ ,  $s$  represents  $i + 1$  to  $j$  cases that are observed on both  $\mathbf{X}$  and  $\mathbf{Y}$ , and  $m$  signifies  $j + 1$  to  $k$  cases observed on  $\mathbf{X}$  but not on  $\mathbf{Y}$ .

The goal of Wilks' FML technique for MLE estimation was to estimate the unknown population parameters (i.e.,  $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}$ ) despite the occurrence of missing values in the sample. Wilks accomplished this by separating the likelihood function into separate parts that correspond to  $s$ ,  $m$ , and  $n$  (the observed missing data patterns). In this way, each missing data pattern can be readily calculated and contributes information to the likelihood function. This process can be thought of as three separate FML calculations. The first FML relates to the  $n$  missing data pattern and would contribute information from a univariate normal density function including  $\mu_y$  and  $\sigma_y^2$  which can be illustrated as:

$$\text{FML}_n = \log L_y = \sum_{i=1}^N \log \left\{ L_{y_i} = \frac{1}{(2\pi)^{1/2} |\sigma_y^2|^{1/2}} e^{-.5 \frac{([Y_i] - [\mu_y])^T |\sigma_y^2|^{-1} ([Y_i] - [\mu_y])}{|\sigma_y^2|}} \right\} \quad (14)$$

Likewise, the second FML relates to the  $m$  pattern and would contribute information from a univariate normal density function including  $\mu_x$  and  $\sigma_x^2$ ,

$$\text{FML}_m = \log L_x = \sum_{i=1}^N \log \left\{ L_{xi} = \frac{1}{(2\pi)^{1/2} |\sigma_x^2|^{1/2}} e^{-.5 \left( [X_i] - [\mu_x] \right)^T |\sigma_x^2|^{-1} \left( [X_i] - [\mu_x] \right)} \right\} \quad (15)$$

Lastly, the third FML relates to the  $s$  missing data pattern and would contribute information from a bivariate normal density function including  $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}$ , which is expressed as:

$$\text{FML}_s = \log L_{xy} = \sum_{i=1}^N \log \left\{ L_{xyi} = \frac{1}{(2\pi)^{2/2} \begin{vmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{vmatrix}^{1/2}} e^{-.5 \begin{pmatrix} [X_i] - [\mu_x] \\ [Y_i] - [\mu_y] \end{pmatrix}^T \begin{vmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{vmatrix}^{-1} \begin{pmatrix} [X_i] - [\mu_x] \\ [Y_i] - [\mu_y] \end{pmatrix}} \right\} \quad (16)$$

The likelihood function for the observed data is then maximized with respect to each of the population parameters  $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$ . Specifically, Wilks (1932) used the ratios  $m/s$  and  $n/s$  to weight the first derivatives of the parameter estimates prior to setting them equal to 0 and solving by approximation. Though, this idea was ahead of computational technology at the time and would not become practical until the development of sophisticated iterative methods and modern computers. Still, by incorporating FML information from each missing data pattern, Wilks was able to demonstrate the mathematics behind a maximum likelihood missing data estimation routine that effectively weights the FML solution  $(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x^2, \hat{\sigma}_y^2, \hat{\sigma}_{xy})$  by using information from each missing data pattern (i.e.,  $\text{FML}_s, \text{FML}_m$  and  $\text{FML}_n$ ). Conceptually, the final FML solution is determined by combining all of these expressions where  $\text{FML} = \text{FML}_s + \text{FML}_m + \text{FML}_n$ . That is, each of the FML parameter estimates are located using first derivatives weighted by the observed information (see Wilks, p. 166).

Due to computational limitations, Wilks' work on maximum likelihood estimation in the presence of missing data remained largely undeveloped for nearly five decades until Anderson (1957) and a few others worked to refine the technique (see Little & Rubin, 2002 for exceptions). Anderson was among the first to apply maximum likelihood methods to an actual missing data example (e.g., a simple situation with one missing value) by a non-iterative factoring technique similar to that proposed by Wilks (see Dodge, 1985 for details). This approach allowed for an ease in computational complexity but was limited to very specific missing data patterns. A short time later, Hartley (1958) also added to work on maximum likelihood methods with an iterative technique involving contingency tables. Later, Hartley and Hocking (1971) applied an iterative maximum likelihood technique to multivariate normal data in the presence of missing data. Their basic approach was similar to the original ideas described by Wilks (1932). For instance, Hartley and Hocking separated the data into groups where each group represents a different missing data pattern. Then, they maximized the likelihood function for each group (i.e., missing data pattern) with respect to the population parameters. To illustrate how Hartley and Hocking (1971) selected information for each of the missing data patterns, consider the data example shown in Appendix C. As demonstrated in the previous discussion, Wilks (1932) can be seen as the origin for much of the work on missing data in the context of maximum likelihood estimation. Still, this was not his only contribution to the field. Mean substitution and pairwise deletion methods are discussed next.

**Simpler systems of estimates in the presence of missing data.** Wilks (1932) was well aware of the computational limitations of his time. As he noted, "in view of the difficulties connected with the foregoing maximum likelihood estimates, we shall devote the remainder of this paper to a consideration of the moments, distributions and efficiencies of simpler systems of estimates," (p. 178). The impending discussion covers Wilks' simpler systems including pairwise deletion and mean substitution. It is interesting to note that Wilks' method of pairwise deletion is

still the default missing data handling procedure in many current statistical software programs, despite its original purpose as essentially an alternative to cumbersome hand calculations.

**Pairwise deletion.** Conceptually, the maximum likelihood process described previously is similar to the *pairwise deletion* method also outlined by Wilks (1932) though these approaches are not mathematically associated (Enders, 2001). Rather than relying on a probability density function to maximize the likelihood of the parameter estimates employing all of the observed data, Wilks’ “less efficient but simpler” pairwise deletion method directly calculates each covariance estimate from the available data. To illustrate consider the sample covariance  $\hat{\sigma}_{XY}$  and correlation  $r_{XY}$  formula (Johnson & Wichern, 2002):

$$\hat{\sigma}_{XY} = \frac{\sum(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{N - 1} \quad r_{XY} = \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}} \quad (17)$$

where  $\hat{\mu}_X$  is the mean of  $X$ ,  $\hat{\mu}_Y$  is the mean of  $Y$ ,  $\hat{\sigma}_X^2$  is the variance of  $X$ ,  $\hat{\sigma}_Y^2$  is the variance of  $Y$ , and  $\hat{\sigma}_{XY}$  is the covariance between  $X$  and  $Y$ . Now consider the pairwise deletion method where  $\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\sigma}_{XY}$  are each calculated based on all available data. This estimate corresponds to the following (see Wilks, 1932):

$$\hat{\sigma}_{XY}^s = \frac{\sum(x_i - \hat{\mu}_X^s)(y_i - \hat{\mu}_Y^s)}{N^s - 1} \quad r_{XY}^s = \frac{\hat{\sigma}_{XY}^s}{\sqrt{\hat{\sigma}_X^{s2} \hat{\sigma}_Y^{s2}}} \quad (18)$$

where  $N^s$  is a set of cases with both  $x_i$  and  $y_i$  observed (e.g.,  $\mathbf{s}$  in Figure 11), and the means  $\hat{\mu}_X^s$  and  $\hat{\mu}_Y^s$  are typically calculated over that set of cases (Little and Rubin, 2002). However, it is important to note uncertainty as  $\hat{\mu}_X^s$  could be calculated over all available  $i$  cases for  $x$  (e.g.,  $\mathbf{s}$  and  $\mathbf{m}$  in Figure 10). Likewise, the mean  $\hat{\mu}_Y^s$  could also be calculated over all available  $i$  cases for  $y$  (e.g.,  $\mathbf{n}$  and  $\mathbf{s}$  in Figure 11). Therefore, pairwise deletion may use some cases in one part of

the sample covariance formula (e.g., the mean deviation) and another set of cases in another part (e.g., the sample mean).

While this approach attempts to use all the available information, the generated covariance matrix (which might then be used in a subsequent multivariate analysis) is based on different subsets of cases. Frequently this situation results in negative eigenvalues (a non-positive definite matrix; see Graham, 2009; Kline, 2011; Marsh, 1998; Wothke, 1993). As will be discussed in detail later, an eigenvalue is essentially a variance term related to an optimally weighted linear combination of the data and as such cannot be negative (see Jackson, 1991; Johnson & Wichern, 2002). Said differently, the resulting covariance matrix might be internally inconsistent because the range of possible values for a particular variance/covariance estimate are not dependent on all of the other relationships in the matrix. For example, consider the correlation between the variables  $X$  and  $Y$  ( $r_{XY}$ ) given the correlation between the variables  $X$  and  $Z$  ( $r_{XZ}$ ) and the correlation between the variables  $Y$  and  $Z$  ( $r_{YZ}$ ) which can be written as (see Kline, 2011; Kutner, Nachtsheim, Neter, and Li, 2005):

$$r_{XZ}r_{YZ} \pm \sqrt{(1-r_{XZ})(1-r_{YZ})} \quad (19)$$

Now suppose  $r_{XZ} = .20$  and  $r_{YZ} = .32$  so that  $r_{XY}$  must be greater than  $-.67$  and less than  $.80$  for the matrix to remain invertible (i.e., positive definite). Additionally, Kline (2011) notes that, "...the maximum absolute value of a covariance between any two variables is (must be) less than or equal to the square root of the product of their variances," (p. 50) which can be expressed as:

$$\max |\hat{\sigma}_{XY}| \leq \sqrt{(\hat{\sigma}_X^2)(\hat{\sigma}_Y^2)} \quad (20)$$

where  $\hat{\sigma}_{XY}$  is the covariance between  $X$  and  $Y$ ,  $\hat{\sigma}_X^2$  is the variance of  $X$ , and  $\hat{\sigma}_Y^2$  is the variance of  $Y$ . When data are missing these invalid correlation matrices are possible when pairwise deletion is used and can lead to subsequent intractable multivariate analyses since the data matrix

cannot be inverted via matrix algebra manipulations (Kline, 2011). Assuming a positive definite covariance matrix, research thus far does not clarify how standard errors should be calculated because the covariance matrix is not possible given any particular sample size (Allison, 2001; Newman, 2003). As Enders (2010) pointed out, the pairwise deletion approach may lead to bias and should be avoided.

**Mean substitution.** Another approach taken by Wilks (1932) was to replace the missing values with the means from available cases. Applying the available case idea to the mean of  $X$  ( $\hat{\mu}_X$ ) and the mean of  $Y$  ( $\hat{\mu}_Y$ ) yields:

$$\hat{\mu}_X^{\text{sm}} = \frac{1}{N^{\text{sm}}} \sum_{i=1}^{N^{\text{sm}}} x_i \quad \hat{\mu}_Y^{\text{ns}} = \frac{1}{N^{\text{ns}}} \sum_{i=1}^{N^{\text{ns}}} y_i \quad (21)$$

where  $N^{\text{sm}}$  is the set of observed  $x$  cases (i.e.,  $s + m$  from Figure 11), and  $N^{\text{ns}}$  is the set of observed  $y$  cases ( $n + s$  from Figure 11). Once calculated  $\hat{\mu}_X$  and  $\hat{\mu}_Y$  are then used to replace missing values for  $x$  and  $y$  respectively. Wilks' noted that covariances can then be constructed from the complete data such that the following formula is obtained for the covariance:

$$\hat{\sigma}_{XY}^{(sm)(ns)} = \frac{\sum (x_i - \hat{\mu}_X^{\text{sm}})(y_i - \hat{\mu}_Y^{\text{ns}})}{N - 1} \quad r_{XY} = \frac{\hat{\sigma}_{XY}^{(sm)(ns)}}{\sqrt{\hat{\sigma}_X^{2(sm)} \hat{\sigma}_Y^{2(ns)}}} \quad (22)$$

Where  $\hat{\mu}_X^{\text{sm}}$  is the mean among all  $i$  observed  $x$  cases and  $\hat{\mu}_Y^{\text{ns}}$  denotes the mean among all  $i$  observed  $y$  cases. Mean replacement may seem to make use of relevant information by replacing missing values by the “most likely” value for that variable (its mean); however, mean substitution has been shown to underestimate the variance and can attenuate relationships (see Allison, 2003; Peugh & Enders, 2004). Enders (2010) notes that mean substitution is, “...possibly the worst missing data handling method available,” (p. 43). That is, mean substitution is likely to perform worse than deletion methods, despite the reason data are missing.

**Missing information.** Of the previously mentioned innovations provided by Wilks (1932), each has found a place in current missing data handling approaches. For example, likelihood-based approaches to address missing data are now common in most current statistical software packages (e.g., Mplus, SAS, SPSS, etc.) and pairwise deletion is a popular default in many programs (e.g., SPSS). In addition to these topics, another important advancement is worth mention. Wilks (1932) introduced the idea of “estimation efficiency” in the presence of missing data and described it mathematically as, “...the ratio of the reciprocal of the determinant of its variances and covariances to that of the set of maximum likelihood estimates of the same parameters,” (p. 171). To further this idea, consider that the determinant of a symmetric matrix (such as a positive-definite covariance matrix) is equal to the product of its eigenvalues (Kolman, 1996; see subsequent PCA discussion for clarification) and reflects the amount of variance information in the data.

While the idea of estimation efficiency employed by Wilks was to demonstrate the potential effectiveness of maximum likelihood estimation over other possible methods, this paper was perhaps the first to consider the concept of missing information, which was later developed by Orchard and Woodbury (1972). Consider that Orchard and Woodbury, like Wilks, consider the amount of variance information in a data set a measure of the certainty of parameter estimates. To illustrate, consider that the information contained in a data set relates to the steepness of the likelihood function (as described by Enders, 2010). That is, the steepness of the likelihood function is related to the amount of information. Thus, the more information there is to use, the more certain the parameter estimates will be. Figure 12 provides a depiction of this concept.

Figure 12 includes two three-dimensional plots of a bivariate normal probability distribution with variables X and Y where the density function  $f(X, Y)$  represents the height of the probability surface. Panel A of Figure 12 shows a high information and small standard error

condition. Here the likelihood function is steep. This situation would result in little ambiguity regarding the most appropriate parameter estimates as the peaked likelihood surface indicates greater precision (i.e., smaller standard error). In contrast, Panel B of Figure 12 demonstrates low information and large standard errors reminiscent of a flat likelihood function. Parameter estimates associated with this likelihood function are less clear than those in Panel A due to a larger standard error.

Consider a more explicit illustration of this concept in relation to the univariate likelihood function and associated first derivatives discussed previously. Figure 13 contains a series of plots that demonstrate the relationship between the likelihood function and the first derivatives for two different random samples. Panel A of Figure 13 relates to the data previously discussed for the likelihood function of the mean from the column vector  $\mathbf{Y}$ . Notice that this plot replicates Figure 8. Following the example provided by Enders (2010, p. 65-67), Panel B of Figure 13 represents a likelihood function for the same mean but the data were generated to have a larger variance (i.e., 2.5 times as large as  $\mathbf{Y}$ ).

As Enders (2010) notes, the flatter the slope of the regression line (Figure 13 bottom of panel B), the less peaked the shape of the associated likelihood function (Figure 13 top panel of B) relative to panel A. Figure 13 demonstrates that the curvature of the log-likelihood function (i.e., the rate that the first derivative changes) directly reflects the standard error of a particular maximum likelihood estimate of the mean (i.e.,  $\hat{\mu}_Y$ ). Said differently, as the likelihood function becomes less peaked (i.e., flatter), many population values are possible which indicates more difficulty deriving a precise maximum likelihood estimate (i.e., many first derivatives are close to zero) and a larger standard error. Similarly, a likelihood function that is highly peaked offers fewer possibilities for the population estimate. Thus, the associated standard error is smaller.

That is, standard errors are due to the relationship between the estimate precision (i.e., sampling variance) and the likelihood function (see Enders, 2010).

To further this notion, note that the sampling variance of the mean can be written as:

$$\text{var}(\hat{\mu}) = -\left[\frac{\partial^2 \log L}{\partial^2 \mu}\right]^{-1} = \left[\frac{-N}{\sigma^2}\right]^{-1} = \frac{\sigma^2}{N} \quad (23)$$

where  $\text{var}(\hat{\mu})$  represents the sampling variance and  $\partial^2$  denotes the second derivative (see Enders, 2010). The second derivative estimate corresponding to Figure 13 is -3.33 and -1.33 for Panel A and Panel B, respectfully. As noted these values represent the slopes of the log-likelihood regression lines; see bottom of Panel A and B). Note that the slope plot in the bottom of Panel A in Figure 13 is steeper than the associated plot in Panel B. As Enders demonstrated, multiplying these derivative values by -1 (e.g., 3.33 and 1.33) yields an index of the information contained in each likelihood plot. Further, the inverse of the information index is the sampling variance estimates; which are .30 and .75 for Panel A and Panel B of Figure 13, respectively. Finally, taking the square root of the sampling variance generates the standard errors. Therefore, ML standard error-of-the-mean estimates are .55 for Panel A and .87 for Panel B of Figure 13.

Based on this discussion, it is appropriate to describe the steepness of the likelihood function in relation to the amount of information in the sample data (as previously implied). Orchard and Woodbury relate the concept of increased variance (i.e., larger ML standard errors) to the loss of information caused by missing data (they referred to this as “hidden variance”). As they note a small percent of missing data may introduce a large amount of hidden variance.

Missing information, Orchard and Woodbury (1972) argue, is a separate concept from percent of missing data. More specifically, missing data have an influence on the efficiency of parameter estimates to the degree that the missing information increases standard errors.

Importantly, this concept helped stimulate methodologists to develop estimation techniques (e.g.,

the EM algorithm) that include as much information as possible in the likelihood function as this would provide a more precise point estimate and a smaller standard error than is possible with less information (see Little & Rubin, 2002). Specific formulae related to formalizing the concept of missing information will be discussed in relation to multiple imputation in forthcoming discussion.

**Summary of early developments.** In the context of early developments in missing data research, it was shown that the idea of missing data as a problem (i.e., a statistical hindrance) seems to have its origins in the development and application of classical inferential statistics which were designed for complete case (i.e., balanced) data. While the earliest missing data techniques involved complete cases analysis (e.g., data deletion to obtain balance), simple imputation methods were introduced as early as 1930. Discussion acknowledged that early imputation approaches were comprehensively studied in the context of ANOVA, largely in response to requests from applied researchers. Also, a maximum likelihood-based estimation scheme appeared early on but suffered from computational limitations. Many of these early developments in maximum likelihood estimation in the presence of missing data were developed in response to computational limitations and many of the resulting solutions were situation specific. That is, they were designed for a specific and often simplified type of missing data pattern (which was not realistic in practice) or these methods were designed for a specific type of data analysis and were not generalizable to other situations (see Little and Rubin, 2002). Much of the general application of maximum likelihood estimation in the presence of missing data would require computer implementation (e.g., iterative algorithms) which will be discussed later.

Importantly, many key ideas relating to missing data theory and application appeared throughout the discussion of early methods. It seems that the idea of addressing missing data as a potentially complex problem developed as researchers began to acknowledge that results vary as a function of the information available and as a function of the analytic tools used to correct for

missing data. The next historical period will move missing data research forward by the computational advancements provided by computers and also by questioning a commonly held assumption that the process that caused missing data can be ignored.

### **The Second Historical Period: Ignorable Missing Data and Estimation**

Donald Rubin (1976) appears to be the first to question the statistical literature's adoption of missing values as ignorable information. Rubin had already published a covariance procedure to estimate missing data (see Rubin, 1972) and his 1976 article titled, "Inference and Missing Data", was a substantial achievement at the time as it effectively repositioned missing data theory. This article provided the theoretical motivation and the statistical foundation for understanding the relationship between "why the data are missing" and the ability to recover missing data (i.e., the ability to make unbiased sampling distribution inferences). Rubin emphasized that when researchers encounter missing data, incorrectly assuming that incomplete data are missing at random (i.e., ignorable) may bias any inferences made from that sample. He described that missing data cause deviations from probability sampling. Specifically, proper randomization is a fundamental aspect of classical inferential procedures (i.e., Frequentist approach to statistical inference) which can only be assumed when the missing values are themselves an ignorable (e.g., random) form of probability sampling. For clarification, see the brief example of this concept demonstrated in Appendix D.

**The beginnings of a classification system.** Rubin (1976) argues that typically when a study contains missing data, the researcher is never sure which population is represented by the observed sample. That is, the sample is typically selective due to the process that generated missingness. Therefore, the observed sampling distribution could actually be a mixture of several sampling distributions and any statistical inference based on a probability statement about a particular distribution of the population would be inaccurate. As later described by Rubin and Little (1987), "the key ingredient of the randomization approach, a known probability

distribution governing which values are unobserved, is lost when some of the data are missing,” (p. 53). Rubin’s concerns about randomization and probability sampling in the context of missing data introduced the beginnings of an essential classification system for missing data (Enders, 2010; Little & Rubin, 2002) and greatly influenced the introduction of missing data as an independent field of study within the social and behavioral sciences (see Molenberghs & Kenward, 2007). The fundamental aspects of the classification system related to missing data will be discussed and mathematically illustrated in a subsequent section based on a slightly refined version outlined several years later in an influential book by Rubin and Little (1987).

**Introduction of iterative algorithms.** In the Much of the second half of Rubin’s (1976) paper seems to formulate an argument in favor of a promising approach to maximum likelihood estimation (MLE), which was published in a more developed form the following year by Dempster, Laird, and Rubin (1977). Recall that MLE in the presence of missing data was already an established idea (e.g., Anderson, 1957; Hartley, 1958; Hartley & Hocking, 1971; Wilks, 1932); however, it was not practical without sophisticated iterative computational routines. That is, the application of MLE techniques in the presence of missing data usually require “...iterative optimization algorithms to identify the most likely set of parameter values,” (Enders, 2010, p. 65). Said differently, not all likelihood functions can provide a closed-form solution (i.e., one solvable equation) for the maximum likelihood value of a particular parameter.

Little and Rubin (2002) elaborate on the reason a closed-form MLE solution cannot be obtained with missing data. They note, “patterns of incomplete data in practice often do not have the particular forms that allow explicit ML estimates to be calculated by exploiting factorizations of the likelihood,” (p. 164). They continue this discussion noting that even if the missing data patterns were manageable (i.e., the observed data selection matrix can be specified); it is likely that each partition of the likelihood function would not be orthogonal. Therefore, maximizing the likelihood separately for each pattern (as demonstrated previously in the context of Wilks

technique for MLE estimation with missing data) would not actually maximize the overall likelihood. Thus, iterative algorithms were needed to solve otherwise inestimable or at the least extremely complex maximum likelihood calculations (Little & Rubin, 2002).

***Defining an iterative algorithm.*** To further develop this idea, consider that an iterative algorithm is a mathematical procedure that yields a sequence of improving approximate solutions for an estimation problem where a final solution is based on a pre-specified termination criterion that suggests solutions are no longer improving (see McLachlan & Krishnan, 2008). Consider the following set of linear equations that must be solved via approximation because a closed-form solution is not practical. Let these equations be:

$$\begin{aligned} 12x + 2y + 1z &= 14 \\ 6x - 12y + 8z &= 2 \\ 2x - 4y + 14z &= 8 \end{aligned} \quad (24)$$

where the values  $x$ ,  $y$ , and  $z$  are unknown. Notice that substituting a particular value in for  $x$ ,  $y$ , or  $z$  would influence the other estimates and solutions. That is, each unknown cannot be easily isolated to generate a closed-form solution (e.g.,  $x = 6 + 2$ ). Solving these linear equations by trial and error is both inefficient and challenging. Therefore, consider an iterative approach that begins with a simple rearrangement of equations such that (see Kline, 2010):

$$\begin{aligned} x &= \frac{14}{12} - \frac{2}{12}y - \frac{1}{12}z = 1.167 - 0.167y - 0.083z \\ y &= -\frac{2}{12} + \frac{6}{12}x + \frac{8}{12}z = -0.167 + 0.500x + 0.667z \\ z &= \frac{8}{14} - \frac{2}{14}x + \frac{4}{14}y = 0.571 - 0.143x + 0.286y \end{aligned} \quad (25)$$

Now, insert a start value (i.e., an initial guess) for  $x$ ,  $y$  and  $z$ . In the current example, suppose that  $x_0 = 0$ ,  $y_0 = 0$ , and  $z_0 = 0$  where the subscript “0” indicates the iteration number. Typically, zero is an undesirable starting value but is chosen here for didactic purposes. Using these starting values derives the following estimates for  $x$ ,  $y$  and  $z$ :

$$\begin{aligned}
x_0 &= 1.167 - 0.167(0) - 0.083(0) = 1.167 \\
y_0 &= -0.167 + 0.500(0) + 0.667(0) = -0.167 \\
z_0 &= 0.571 - 0.143(0) + 0.286(0) = 0.571
\end{aligned} \tag{26}$$

The next sequence to improve the solution involves inserting these estimates for the values  $x$ ,  $y$ , and  $z$  back into the equation where:

$$\begin{aligned}
x_1 &= 1.167 - 0.167(-0.167) - 0.083(0.571) = 1.147 \\
y_1 &= -0.167 + 0.500(1.167) + 0.667(0.571) = 0.798 \\
z_1 &= 0.571 - 0.143(1.167) + 0.286(-0.167) = 0.357
\end{aligned} \tag{27}$$

Notice that after the first iteration the estimates for  $x$ ,  $y$ , and  $z$  have been updated. This iterative process of estimation and updating the values of  $x$ ,  $y$ , and  $z$  is repeated until subsequent changes in the estimates fall below some pre-specified criterion. Consider the first 10 iterations in Table 2 to demonstrate this process. Here the estimated values for  $x$ ,  $y$  and  $z$  change the most in the first few iterations and do not change much after the eighth iteration. Given a termination criterion of .001 (the cut point for no significant changes in newly estimated values), the current iterative estimation routine would have converged (stopped) at the eighth iteration and exported  $x_8 = 0.986$ ,  $y_8 = 0.758$ , and  $z_8 = 0.647$  as a solution. The quality of an iterative solution depends on the complexity of the estimation routine, starting values and the specific iterative algorithm used (McLachlan & Krishnan, 2008). While the Newton-Raphson algorithm and other similar iterative techniques generally served this purpose, a new technique was needed for missing data situations that were otherwise too complicated to implement (see Watanabe & Yamaguchi, 2004 for elaboration).

*Introduction of the Expectation-Maximization (EM) algorithm.* Dempster, Laird, and Rubin's 1977 paper titled, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, provided such an iterative computation technique for MLE estimates. This paper described the expectation-maximization (EM) algorithm essentially, as we know it today, as a two-step

iterative procedure that exploits regression based imputation and maximum likelihood estimation in concert to address missing data. Dempster, et al., (1977) coined the term “EM” as an acronym for the two-step iterative nature of the procedure and popularized the technique by clarifying the relationship between iterations and the associated increase in parameter likelihood given the observed data (Little & Rubin, 2002).

Dempster, Laird, and Rubin (1977) also illustrated the iterative process for maximizing the likelihood of the covariance matrix and mean vector given the complete data and provided numerous examples. The practicality of the EM algorithm lies in its ease of implementation (Enders, 2001; 2010). As will be demonstrated, the expectation step of each iterative sequence involves taking expectations over complete-data and the following maximization step simply requires complete data maximum likelihood estimation, which is typically solved with a closed-form mathematical function for the first derivative (see Enders, 2001).

While earlier methods for addressing missing data required that missing data occur in a complete random fashion, a point criticized by Rubin (1976). The EM algorithm required a less stringent assumption. Specifically, the causes and correlates of missing data are not necessarily random but must be included in the estimation routine. This was a significant step forward in coupling missing data theory with techniques for estimation in the presence of missing data.

Briefly, Dempster, Laird, and Rubin (1977) began with an initial estimate of the means and covariance matrix, which provide sufficient statistics from the observed data. Missing values are not included at this point (i.e., listwise deletion for missing values). The Expectation step (the “E” in EM) uses these initial estimates to predict missing data in the observed data using a set of regression equations. The result is a new data set with no missing values. The Maximization step (the “M” in EM) then generates an updated set of sufficient statistics from this new data set. The EM procedure is repeated until changes in the sufficient statistics from one M-step to the next become trivial. The converged EM solution contains maximum likelihood estimates of the

sufficient statistics. Appendix E demonstrates this procedure using example data.

Importantly, rather than ignoring potential causes and correlates of missing data (variables that may explain why data are missing), the EM algorithm may include this information in the estimation of sufficient statistics. For example, if missingness on variable Y is caused by another variable X, the EM based sufficient statistics can incorporate X, or any such variable, by adding that variable to the likelihood function during the EM iterations. Said differently, the maximum likelihood estimate of the covariance matrix and mean vector generated by the EM algorithm can be informed by all available data and statistical modeling can then proceed with the specific variables of interest.

The EM algorithm is considered an “indirect” approach (Enders, 2001) because the final sufficient statistics must be used for other analyses such as factor analysis or used to impute a final complete data set (i.e., stopping after the last “E” step). As an indirect approach to handle missing data, the EM algorithm has some important limitations that did not adequately solve the problem of estimation in the presence of missing data. For instance, while the EM algorithm is fully capable of incorporating extra variables that are causes or correlates of missingness, it was not designed to generate appropriate standard errors; which require a corrective bootstrapping technique (see Enders & Peugh, 2004).

Specifically, when the EM sufficient statistics are used to impute a final complete data set (i.e., final set of linear regression equations; “E” step) the variability observed during the EM iteration process is not retained. Enders (2001; 2010) described this as filled-in values falling directly on a regression line because the final imputation is based solely on the final set of sufficient statistics (from the last “M” step).

Further, the literature is not clear about the appropriate sample size to reference with the EM algorithm. For instance, Graham and Schafer (1999) relate the EM algorithm to pairwise deletion noting, “...it is not clear what sample size should be ascribed to a covariance matrix

estimated by EM,” (p. 3). As will be subsequently demonstrated, these limitations coupled with the desire for a “direct” approach, where missing data are addressed and parameter estimates are estimated in a single step, would lead to a transition from the EM algorithm to the direct ML approach (i.e., FIML).

*Introduction of full-information maximum likelihood (FIML).* An emphasis on maximum likelihood estimation and advances in computer technology (especially in relation to matrix algebra; see Jöreskog, 1967) came together a couple years after the introduction of the EM algorithm. For instance, methodologists assimilated these ideas and considered iterative parameter estimation techniques for the factor analysis model, which resulted in Carl Finkbeiner’s, “Estimation for the Multiple Factor Model When Data Are Missing,” published from his dissertation in 1979. This paper was essentially an extension of earlier work by Wilks (1932; see acknowledgement in Finkbeiner, 1979) and introduced full-information maximum likelihood (FIML; Enders, 2001) for use in factor analysis. As Enders (2001) noted, FIML is conceptually the same approach taken by Hartley and Hocking (1971; whom extended the work of Wilks as previously noted) where the data are separated the data into groups that represent various missing data patterns. However, rather than generating a set of group-level likelihood estimates FIML was designed to use individual (i.e., case specific) likelihood functions. As Finkbeiner (1979) wrote, “the overall discrepancy function value is obtained by summing the  $n$  casewise likelihood functions...” (p. 134) which can be expressed as:

$$\log L = \sum_{i=1}^N \log \left\{ L_i = \frac{1}{(2\pi)^{p_i/2} |\Sigma_i|^{1/2}} e^{-.5(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)} \right\} \quad (28)$$

where  $p_i$  represents the number of complete data points for individual (i.e., case)  $i$ , and  $Y_i$  denotes the data for the  $i^{th}$  individual (i.e., case). In addition, the mean vector  $\mu_i$  and the covariance

matrix  $\Sigma_i$  represent parameters unique to each case. Said differently,  $\mu_i$  and  $\Sigma_i$  may contain  $j$  variables for a particular case but may contain  $k$  variables for another case where  $j \neq k$  (see Allison, 2002; Enders, 2010; Little & Rubin, 2002; Peugh & Enders, 2004). Note the similarity of this equation to the complete data MLE formula. As Enders (2010) notes, the only difference is that this formula does not contain the extra subscript  $i$  for the population parameters.

FIML is considered a model-based approach because missing data are handled within a single, iterative step where missing values are not imputed (Graham, Cumsille, & Elek-Fisk, 2002). Rather, model parameters and standard errors are directly estimated from the observed data of each individual case. In this way, the Finkbeiner paper marks a contrast from Dempster, Laird, and Rubin's (1977) EM algorithm approach to correcting for missing data. Specifically, the EM algorithm generates MLE of the population parameters ( $\mu, \Sigma$ ) which are then used as sufficient statistics for further analyses. By comparison, FIML directly estimates a statistic and its associated standard error from the data. Consequently, FIML is considered a direct approach and is often referred to in the literature as *direct*-maximum likelihood (direct ML; e.g., Enders, 2001; 2010).

To compare this approach to other techniques used previously, consider that FIML is similar to pairwise deletion in that both procedures use all the available data on a case wise basis (Enders, 2001). However, FIML does not simply perform a direct calculation of each covariance estimate from the available data. Rather, the FIML procedure relies on a probability density function to iteratively maximize the likelihood of the estimates employing, "...all of the information of the observed data, including mean and variance for the missing portions of a variable, given the observed portion(s) of other variables," (Wothke, 2000, p. 3). Consequently, FIML, unlike pairwise deletion, uses the appropriate sample size for standard error estimates (Allison, 2003; Enders, 2001) and is less likely to produce a non-positive definite matrix

(Wothke, 1993). FIML is considered “state of the art” in correcting for missing data (Enders & Peugh, 2004; Schafer & Graham, 2002) because it can adequately recover key information, and is based on a sound theory, unlike many of the more traditional methods for addressing missing data (see Baraldi & Enders, 2010; Collins, Schafer, & Kam, 2001). For these reasons, methodologists typically recommend FIML over traditional methods (see Graham, 2009 for exceptions).

Researchers have investigated many aspects of FIML estimation through simulation studies (e.g., Arbuckle, 1996; Enders & Bandalos, 2001; Peugh & Enders; Wothke, 2000) and describe the technique as unbiased and efficient (i.e., produces a small standard error relative to preceding missing data methods) given that certain assumptions are met. Specifically, FIML requires a relatively large sample size (see Schafer & Graham, 2002 for details), a multivariate normal distribution (Enders, 2010), and the cause of missingness must be included in the estimation procedure.

Of these assumptions ensuring that the causes and correlates of missing data are included in the FIML likelihood function is typically the most challenging (Enders, 2010; Enders & Bandalos, 2001; Little & Rubin, 2002; Schafer & Graham, 2002; Wothke, 2000). FIML estimation is thought to be relatively robust to violations of normality and small sample sizes. For instance, methodologists have noted that FIML is likely robust to mild and even moderate violations of multivariate normality (e.g., Enders, 2010) and severe non-normality can be addressed via one of many modified FIML estimation routines that are robust to non-normality (e.g., Muthén & Muthén, 2011).

However, as a direct approach for addressing missing data, the FIML procedure only incorporates variables that are directly modeled. For example, FIML only integrates information from variables that are referenced in the linear model investigated (e.g., multiple regression, factor analysis, etc.). To illustrate this concept, see Appendix F, which demonstrates how

Finkbeiner (1979) applied FIML to the factor analysis model.

The inability of the FIML procedure to include all possible causes or correlates of missingness was initially a significant limitation (see Enders, 2001). However, this drawback was addressed in the literature when Graham (2003) provided effective, though somewhat complex procedures to incorporate extra “auxiliary” variables into the FIML estimation routine without modifying the basic statistical model under investigation. This concept is discussed in more detail throughout the forthcoming sections.

**Summary of ignorable missing data and estimation.** As discussed, the traditional methods for handling missing data, though easy to implement by hand calculation, often place the researcher in a position to be the most wrong that he/she can possibly be because most missing data are not ignorable (Rubin, 1976). To avoid methods that attempt to edit the observed, incomplete dataset by filling in missing data or deleting data (e.g., single regression imputation, listwise deletion, pairwise deletion, and mean substitution), researchers turned toward iterative estimation techniques that are based on methodological theory rather than computational convenience.

Due to Rubin’s (1976) introduction to modern thinking about missing data and advances in computer technology, methodologists assimilated the EM algorithm and a short time later FIML. The difference in substitution procedures like single regression imputation and approaches like EM algorithm or FIML relates to the sophistication involved in generating plausible values. These methods do not assume that the data are missing completely at random and are able to incorporate the cause of missingness in the estimation procedure. That is, despite the noted limitations EM and FIML can more adequately recover key information, and are based on statistical theory, unlike many long-standing traditional approaches like deletion and single imputation (Graham, 2009; Schafer & Graham, 2002).

The following historical period will highlight further developments in missing data

estimation, including the introduction of a new technique called multiple imputation.

Additionally, the rising field of missing data research asserts influence in thinking about missing data throughout the social and behavioral sciences by producing a series of widely influential books and articles. The implementation of appropriate missing data techniques like FIML become more common with the introduction of less technical computing software programs primarily intended for applied researchers.

### **The Third Historical Period – A Revolution in Missing Data**

As Graham (2009) noted, “problems brought about by missing data began to be addressed in important ways starting in 1987, although a few highly influential articles appear before then,” (p. 550). In particular, two persuasive books appeared on the topic of missing data including Little and Rubin (1987), and Rubin (1987); additionally, two prominent papers also emerged with Allison (1987), and Tanner & Wong (1987) among others that same year. These publications provided the social and behavioral sciences with formal discussions and demonstrations that linked missing data research with work on statistical modeling and the computational advancements provided by personal computers at the time (see Enders, 2010; Graham, 2009; Little, 2011). For instance, Allison’s (1987) paper titled, “Estimation of linear models with incomplete data,” implemented maximum likelihood with missing data (as derived by Hartly & Hocking, 1971 and shown by Finkbeiner, 1979) in the LISREL software package for estimating structural equation models (SEM). In addition to a mathematical demonstration of this technique with matrix algebra, Allison also included appendices with LISREL code and example data. These appendices reflect the aspirations of the larger missing data field to bridge the gap between theoretical suggestions on handling missing data and practical applications.

The following discussion touches on some relevant aspects of these works and demonstrates a discernible turning point in the understanding and acceptance of missing data theory and the implementation of “modern” missing data handling techniques.

**Theoretical overview of missing data mechanisms.** When researchers encounter missing data, the recoverability of that information is related to the missing data mechanisms or “why the data are missing” (Enders, 2010). If the social scientist, or any other scientist, can determine why data are missing, they can recover the lost information (Little & Rubin, 2002). That is, the researcher can determine what missing scores would have been had the participant remained in the sample. However, when the cause of missingness is not explained (i.e., variables responsible for missing data are not included in the missing data handling procedure) the recoverability of lost information is severely limited, as will be explained in the forthcoming discussion. The theoretical ability of a researcher to recover lost information is important for choosing a missing data handling procedure and subsequently deriving appropriate statistical inferences (Enders, 2010).

Rubin’s (1976) earlier discussion of the recoverability of missing data in relation to the cause of missingness was formalized and termed “missing data mechanisms” in Roderick Little and Donald Rubin’s 1987 book, “Statistical Analysis with Missing Data” (the second edition was published in 2002). The missing data mechanisms outlined by Little and Rubin (1987; 2002), include concepts that are prevalent today as a theoretical foundation for understanding and appropriately handling missing data (see Enders, 2010).

***Defining missing data mechanisms.*** Little and Rubin (1987; 2002) refined Rubin’s (1976) probability theory discussion and formalized three basic mechanisms that cause missingness to arise in a research study. These missing data mechanisms include data that are *missing completely at random* (MCAR), data that are *missing at random* (MAR) and data that are *missing not at random* (MNAR). This section will introduce each of these missing data causes and demonstrate data examples in turn. As will be shown, the distinctions between these mechanisms involve the relationship between the probability of missingness and the data.

Consider that when data are missing, the original rectangular data matrix is unbalanced since the original  $i$  measurements on  $p$  variables are not completely observed. Figure 14 shows an exemplar data matrix with missing values and provides notation that will be referenced throughout the forthcoming discussion. Specifically, let  $\mathbf{Y}$  in Figure 14 illustrate an  $i \times p$  data matrix with missing values and allow matrix  $\mathbf{R}$  to be an  $i \times p$  missingness indicator matrix where observed values in matrix  $\mathbf{Y}$  are represented by 1's and missing values in matrix  $\mathbf{Y}$  are 0's (or vice versa).

The matrix  $\mathbf{R}$  indicates the probability data are missing, which is denoted by  $P(\mathbf{R})$ , because it refers to a matrix of indicator variables with a probability distribution (Little & Rubin, 1987, Rubin, 1976). To develop the concept of the probability distribution of the matrix  $\mathbf{R}$ , recall Rubin's (1976) argument that the observed probability distribution of a sample with missing data could actually be a mixture of several probability distributions. Explicitly stated, when missing data are present the researcher is unsure which population is represented by the observed sample. To address this issue, Rubin (1976) suggests defining missing data as a variable (i.e.,  $\mathbf{R}$ ) because variables have a probability distribution that can be subsequently described by an equation. That is,  $P(\mathbf{R})$  can be expressed as a function of the data through the conditional distribution of  $\mathbf{R}$  given  $\mathbf{Y}$ . As will be revealed in the following discussion, missing data mechanisms are fundamentally assumptions about the independence of these probability distributions. As Enders (2010) noted, "the formal definitions of missing data mechanisms involve different probability distributions for the missing data indicator,  $\mathbf{R}$ ," (p. 10). To solidify the idea of probability distributions in this context, consider the graphical example presented in Appendix G.

An important point made by Rubin (1976) is that variables that a researcher does not measure may significantly influence those variables that are measured. Therefore, any discussion of missing data causes must consider all possible factors that may influence the probability distribution of  $\mathbf{R}$ . Prior to expressing each missing data mechanism; consider the theoretical

overview of the missing data mechanisms outlined in Figure 15. This figure presents a framework for missing data mechanisms that highlights the relationship among  $Y_p$  and  $\mathbf{R}$  in relation to measured ( $Y_{obs}$ ) and unmeasured data ( $Y_{mis}$ ).

Specifically, Figure 15 is composed of four portions including: (1) MCAR, (2) MNAR, (3) MAR and (4) MAR / MNAR. These sections form a grid that identify the missing data mechanism associated with observed and unobserved that are either dependent or independent of the probability of missingness. For instance, the top left panel of Figure 15 corresponds to the MCAR mechanism and represents a situation with independent probability distributions (i.e., a condition of “no association”). All other panels denote some level of relationship between  $\mathbf{R}$  and the data. As will be discussed, in order for a researcher to assume that a particular population is represented by an observed sample with missing data; the data must be shown to be independent or predictive of  $\mathbf{R}$ . Otherwise, the sample would include the possibility of severe bias due to selectivity (e.g., selective attrition).

*Missing completely at random (MCAR).* Missing data are missing completely at random (MCAR) when missingness is due to a truly random process (Little & Rubin, 1987, 2002). That is, the probability that a value is missing ( $\mathbf{R}$ ) is unrelated to the observed data ( $Y_{obs}$ ) and also unrelated to unmeasured data ( $Y_{mis}$ ). Mathematically, MCAR can be expressed as:

$$P(\mathbf{R} | Y_{obs}, Y_{mis}, \phi) = P(\mathbf{R} | \phi) \quad (29)$$

where  $\mathbf{R}$  denotes the probability of missingness,  $Y_{obs}$  and  $Y_{mis}$  reference the data, and  $\phi$  is a parameter that describes the relationship between  $\mathbf{R}$  and the data (Enders, 2010; Little & Rubin, 2002). This equation indicates that the probability of missingness,  $\mathbf{R}$ , given  $Y_{obs}$  and  $Y_{mis}$  (the left side of the equal sign) is simply equal to the probability of missingness (the right side of the equal sign). Note that  $\phi$  merely describes the mathematical connection between  $\mathbf{R}$  and the data regardless of the relationships (i.e., random or predictive).

MCAR is indicative of mathematical independence between  $\mathbf{R}$  and the data where the occurrence of one event in no way affects the probability of the other event. Essentially, these independent probability distributions have nothing to do with each other and the  $\phi$  parameter cannot explain one from the other. More specifically, when  $P(\mathbf{R} | \mathbf{Y}) = P(\mathbf{R})$ , where  $\mathbf{Y}$  includes  $Y_{obs}$  and  $Y_{mis}$ , the conditional independence of  $\mathbf{R}$  and the data vector  $\mathbf{Y}$  can be expressed as:

$$P(\mathbf{R} | \mathbf{Y}) = \frac{P(\mathbf{Y} \cap \mathbf{R})}{P(\mathbf{Y})} \quad (30)$$

where the symbol  $\cap$  represents the word “and” denoting the occurrence of two events (see Hays, 1994). Note that the numerator can be expressed as:

$$P(\mathbf{Y} \cap \mathbf{R}) = P(\mathbf{Y})P(\mathbf{R}) \quad (31)$$

so that the term  $P(\mathbf{Y})$  cancels out of the equation leaving:

$$P(\mathbf{R} | \mathbf{Y}) = P(\mathbf{R}) \quad (32)$$

Practically, the idea is similar to a coin flip where knowledge that the first flip resulted in heads does not change the probability of the next flip will also be heads. For this reason, Little and Rubin (2002) describe MCAR missing data as equivalent to a simple random sample of the complete data.

To further this discussion, Figure 16 presents a couple path diagrams illustrating the MCAR mechanism. Panel A of Figure 16 provides a conceptual diagram of MCAR that shows a random process predicting the probability of missingness  $P(\mathbf{R})$ . Panel B of Figure 16 depicts  $p$  observed variables ( $Y_1 - Y_p$ ) as boxes which are related (i.e., correlated) to some degree which is denoted by solid reciprocal paths (double headed arrows) where  $r \neq 0$  (indicating the association is not zero). Additionally, notice that unmeasured variables (symbolized by a box labeled  $Y_{mis}$ ) may be related to the probability of missing data (a box labeled  $\mathbf{R}$  in Figure 16); but  $Y_{mis}$  is not

related to  $Y_I - Y_p$ . Therefore, any influence from  $Y_{mis}$  to  $\mathbf{R}$  does not impact  $Y_{obs}$  (i.e.,  $Y_I - Y_p$ ). Therefore, these components are independent (shown by dotted reciprocal paths) where  $r = 0$  (suggesting no association). Essentially,  $P(\mathbf{R})$  described by the parameter  $\phi$  is unrelated to measured or unmeasured systematic processes (i.e., missingness is completely random).

For example, suppose that data on language outcomes (i.e.,  $Y_I - Y_p$ ) among children are missing for a subset of participants. More specifically, let  $Y_2$  represent a particular variable with missing data while the other outcomes are completely observed. As previously discussed,  $\mathbf{R}$  describes the propensity for missing data. That is,  $\mathbf{R}$  represents the reason that children did not provide responses to  $Y_2$ .

Now, suppose that there are unmeasured reasons ( $Y_{mis}$ ) for this missing data. For example, assume that a few children did not understand the  $Y_2$  item, that the assessor (investigator) accidentally skipped  $Y_2$  for some participants, or that some children got bored and stopped participating in the language assessment, alternatively perhaps children were distracted by some random occurrence. These situations may cause missing data  $\mathbf{R}$  (denoted by  $r \neq 0$  in Figure 16 between  $\mathbf{R}$  and  $Y_{mis}$ ); however, notice that in Figure 16 there is no systematic relationship between  $\mathbf{R}$  and the language outcomes ( $Y_I - Y_p$ ). In words, there are reasons for missing data but missingness on a particular variable (i.e.,  $Y_2$ ) is not systematically related to measured or unmeasured variables. The cause of missingness is random.

All missing data handling procedures (e.g., likelihood-based techniques, data augmentation techniques, deletion techniques, etc.) assume at least the MCAR level of missing data and almost any missing data technique will lead to unbiased results given this type of missing data (Enders, 2010; Graham, 2009; Little & Rubin, 2002). That is, when missing data are MCAR, the researcher can completely recover the lost information from the remaining data. That is, there are no systematic causes of missingness so the probability distribution of  $\mathbf{R}$  is simply a random sample of the data. However, some missing data handling procedures are still

preferable. For instance FIML is preferred over deletion techniques due to an increase in statistical power (Enders, 2010).

To demonstrate a data example of MCAR, consider a research scenario where it is of interest to estimate the relationship between two random normal variables,  $Y_1$  and  $Y_2$ . These artificial data contain  $N = 30$  cases and are included in Table 3, where the correlation is  $r_{12} = .67$  when the data are complete. This table will be referenced throughout the discussion of all missing data mechanisms for instructive purposes. Now suppose that all values of the variable  $Y_1$  are observed and only some of the values of  $Y_2$  are observed. Figure 17 summarizes this univariate missing data pattern. Specifically, there are  $m$  cases of  $Y_2$  and  $p$  cases of  $Y_1$ .

Numerous research situations that lead to this pattern of missing data (see Little and Rubin, 1987; Enders, 2010). For instance, let  $Y_2$  represent a questionnaire item with missing data and let  $Y_1$  represent a demographic variable, like age, with no missing values. The variable  $Y_1$  could also represent a fixed variable controlled by the experimenter or any other variable with complete data. In order to illustrate the benign nature of MCAR missingness, let  $Y_2$  contain 50% MCAR missingness (i.e., missingness is generated completely at random; see Table 3). Note that in Table 3 the mean ( $M = -.02$ ) and standard deviation ( $SD = 1.0$ ) of  $Y_2$  under MCAR are close to the complete data estimates (.04 and .89, respectively) despite the sample missing half of the observations (i.e.,  $N = 15$ ). Additionally, as noted in Table 3, the correlation between  $Y_1$  and  $Y_2$  is  $r = .67$  in the complete data situation and  $r = .69$  in the 50% MCAR situation.

*Missing at random (MAR)*. Rubin (1976) noted that missing data that is a completely random process is unrealistic except under planned missing data designs (see Graham, Taylor, & Cumsille, 2001 for details). Rather missing data are typically related to selective (i.e., systematic) processes. This idea is essential for understanding the second mechanism, missing at random (MAR). MAR does not actually propose that the data are missing at random; rather, it means that the data are missing for a reason and the researcher happens to know the reason (Enders, 2010;

Little & Rubin, 2002). That is, there are variables within the observed data set that represent and are predictive of the missing data. To use the original terminology provided by Rubin (1976), MCAR data is missing at random (MAR) and observed at random (OAR) which indicates that the data were collected randomly and are not dependent on other variables. In contrast, data default to MAR when not OAR. That is, MAR data are collected randomly but dependent on other variables in the data set. In this way, MAR can be thought of as a special case of MCAR (i.e.,  $MCAR = OAR + MAR$ ; see Little & Rubin, 2002).

For example, consider the previous research scenario involving the variables,  $Y_1$  and  $Y_2$ . Suppose that participants with low SES and low motivation tend to have poorer  $Y_2$  scores and are more likely to have missing data on  $Y_2$ . These data are not OAR because SES and motivation are predictive of  $P(\mathbf{R})$ . That is, some parameter  $\phi$  (i.e., low motivation, low SES) uses  $Y_{obs}$  to explain the probability distribution of  $\mathbf{R}$  (Little & Rubin, 1987; Rubin, 1976). Stated simply, missing data are MAR (but not OAR) when missingness is due to a random process after controlling for the relation between missingness and other measured variables (Little & Rubin, 1987, 2002; Rubin, 1976). Because the missing data cause is known, it is possible to recover the missing information (Enders, 2010). This concept can be expressed mathematically as:

$$P(\mathbf{R} | Y_{obs}, Y_{mis}, \phi) = P(\mathbf{R} | Y_{obs}, \phi) \quad (33)$$

That is, the probability of missingness  $P(\mathbf{R})$  given the reference data,  $Y_{obs}$  and  $Y_{mis}$ , and a parameter  $\phi$  that describes the predictive nature of missingness via  $Y_{obs}$  is equal to the probability of missingness  $P(\mathbf{R})$  given the observed data  $Y_{obs}$  and information about how the observed data relates to missingness (denoted  $\phi$ ; see Enders, 2010; Little & Rubin, 2002). Note that this missing data mechanism is less restrictive than MCAR (i.e., more reasonable in practice) because missingness is not assumed to be a completely random sample of the observed data; rather, missingness can be random or explained by measured variables ( $Y_{obs}$ ).

Figure 18 presents a path diagram of the MAR mechanism. Panel A of Figure 18 offers a conceptual diagram MAR that shows observed variables ( $Y_{obs}$ ) predicting the missing data ( $\mathbf{R}$ ). Panel B of Figure 18 illustrates MAR in relation to  $p$  variables where  $Y_2$  contains missing data. As demonstrated,  $\mathbf{R}$  is can be associated with all variables except  $Y_2$ . Note that  $\mathbf{R}$  may influence  $Y_2$  through other correlated variables. That is, missingness on  $Y_2$  may be caused by values on any other  $p$  variable in the observed data matrix  $\mathbf{Y}$  (e.g., the lower distribution of  $Y_1$  may cause missingness on  $Y_2$ , etc.). Since any such relationship among these  $p$  variables and  $Y_2$  are observed the missing information can be recovered via  $\phi$ .

To provide a demonstration of the MAR mechanism, suppose again that data on language outcomes (i.e.,  $Y_1 - Y_p$ ) among children are missing for a subset of participants. Once more, let  $Y_2$  represent a variable with missing data while the other variables are observed. In Figure 18,  $\mathbf{R}$  represents the reason that children did not respond to  $Y_2$ . This time suppose that there is a relationship between missingness on  $Y_2$  and another measure called English Language Learner (ELL; where  $Y_p = \text{ELL}$ ), such that participants that do not speak English well are less likely to respond to  $Y_2$ . In this case, there is a systematic relationship between  $\mathbf{R}$  and the missing data. That is, this relationship can be described by the parameter  $\phi$  because ELL status is among the outcome measures. That is, the reason why data are missing is because of ELL status; and ELL status can be used to predict (i.e., explain) missing data. Modern missing data handling procedures come in handy when the data are MAR because these procedures can successfully recover missing information by utilizing the parameter  $\phi$ . In contrast, traditional missing data techniques are biased when the data are MAR because these techniques are unable to relate  $P(\mathbf{R})$  to  $Y_{obs}$  (Enders, 2010).

To demonstrate MAR with data, consider the 50% MAR data example in Table 3. Under the MAR mechanism, missingness on the variable  $Y_2$  was generated from the lower distribution of the variable  $Y_1$ . More specifically, values of  $Y_2$  were deleted if the observed values of  $Y_1$  were

less than or equal to .20 (see Table 3). Among other things, this procedure mimics a scenario in which missing values on a variable (i.e.,  $Y_2$ ) result from low scores on another variable in the data set. Because  $Y_1$  and  $Y_2$  were highly and positively correlated in the complete data example ( $r_{12} = .67$ ) deleting based on the lower distribution of  $Y_1$  effectively deletes from the lower distribution of  $Y_2$ . As demonstrated in Table 3, the result is that the mean of  $Y_2$  ( $M = .54$ ) is too large and the standard deviation ( $SD = .77$ ) is too small. Additionally, note that the correlation between  $Y_1$  and  $Y_2$  is  $r_{12} = .19$  in the 50% MAR situation (see Table 3). While these raw differences are dramatic, modern missing data handling techniques can recover the lost information (given adequate sample size). That is,  $\phi$  can explain that missingness on the variable  $Y_2$  was generated from the lower distribution of the variable  $Y_1$  (i.e., that low scores on one variable relate to low scores on the other, etc.).

*Missing not at random (MNAR).* Missing data are missing not at random (MNAR) when missingness is not completely random and not explained by measured variables (Little & Rubin, 1987, 2002). That is, there are unmeasured variables ( $Y_{mis}$ ) that explain the occurrence of missing data. Essentially, the data are missing for a reason and the research does not have a measure of that reason (i.e., no variables in the data set represent or are predictive of the missing data). According to Enders (2010) missing data default to the MNAR mechanism when the data are not randomly collected (i.e., selective attrition, etc.) and are not observed at random (i.e., no observed variables explain the probability of missingness). Stated simply, the OAR concept does not apply (see Little & Rubin, 2002). This mechanism can be written as:

$$P(\mathbf{R} | Y_{obs}, Y_{mis}, \phi) = P(\mathbf{R} | Y_{obs}, Y_{mis}, \phi) \quad (34)$$

That is, the probability of missingness  $\mathbf{R}$  is related to unmeasured variables ( $Y_{mis}$ ) and perhaps measured variables ( $Y_{obs}$ ) as well (see Enders, 2010; Little & Rubin, 2002). Importantly, the parameter  $\phi$  is unable to describe a relationship between  $\mathbf{R}$  and unmeasured variables. Said

differently, missingness on  $Y_2$  depends on  $Y_2$  itself (e.g., if participants who have lower  $Y_2$  scores are never tested because they avoid being measured). These data are not collected randomly because another process beyond  $Y_1 - Y_p$  generated missing data. If there is a reason for missing data and the researchers do not have a measure of it missingness cannot be explained and the recoverability of lost information is severely limited (Enders, 2010; Little, 2009).

Consider another example relating to the previously mentioned research scenario involving the variables,  $Y_1$  and  $Y_2$ . As before, let participants with low SES and low motivation have poorer  $Y_2$  scores, and be more likely to have missing data on  $Y_2$ . This time, however, suppose that SES and motivation are not included in the data set. In this scenario, the observed data probability distribution can be thought of as a combination of multiple probability distributions where the sample data is not representative of any particular population and therefore any statistical inferences based on a particular population are biased (Schafer, 1997). That is, because the intended sample is selective and likely unrepresentative, the MNAR mechanism is considered *non-ignorable* (Enders, 2010; Little & Rubin, 2002; Schafer, 1997).

Figure 19 illustrates a path diagram of the MNAR mechanism. Panel A of Figure 19 provides a theoretical explanation of MNAR. Here unmeasured variables (denoted  $Y_{mis}$ ) predict (i.e., cause) missingness. Also note that observed variables ( $Y_{obs}$ ) may also predict missingness but strictly speaking,  $Y_{obs}$  and  $Y_{mis}$  are not related. That is, the observed variables are unable to explain the missing data process. Panel B of Figure 19 shows the MNAR mechanism in relation to  $p$  variables where  $Y_2$  contains missing data. Notice that  $\mathbf{R}$  is associated with  $Y_2$ . This situation implies that information from  $Y_{mis}$  relates to the observed missingness because the lost information on the variable  $Y_2$  depends on the probability distribution of  $Y_2$  itself (Enders, 2010).

Continuing with the ongoing example regarding language outcomes, suppose that participants whom would have scored in the lower distribution of the variable  $Y_2$  failed to respond to that variable. This relationship is illustrated as a solid reciprocal relationship where  $r$

$\neq 0$  between  $\mathbf{R}$  and  $Y_2$  in Figure 19. In this case, missing data on  $Y_2$  is directly related to the underlying values of  $Y_2$ . In this situation, there is a systematic relationship between  $\mathbf{R}$  and the missing data but the investigators do not have a measure of it. Therefore, the parameter  $\phi$  is unable to describe the relationship that generated the missing data (the  $\phi$  parameter is uninformative). So, the sampling distribution of  $Y_2$  is biased because it does not contain the lower distribution of  $Y_2$ . Any inferences made based on  $Y_2$  are incorrect representations of that sample.

To demonstrate the MNAR mechanism with example data, consider the 50% MNAR data example in Table 3. Under the MNAR mechanism, missingness on the variable  $Y_2$  was generated from the lower distribution of the variable  $Y_2$ . More specifically, values of  $Y_2$  were deleted if the observed values of  $Y_2$  were less than or equal to .07 (see Table 3). As demonstrated in Table 3, the result is that the mean of  $Y_2$  ( $M = .71$ ) is the largest and the standard deviation ( $SD = .51$ ) is the smallest. Additionally, note that the correlation between  $Y_1$  and  $Y_2$  is  $r_{12} = .67$  in the complete data situation and  $r_{12} = .40$  in the 50% MNAR situation (see Table 3). Unfortunately, missing data handling techniques are generally unable to recover information lost through the MNAR mechanism. That is,  $\phi$  cannot explain that missingness on the variable  $Y_2$  was generated from the lower distribution of  $Y_2$  itself.

Using the previously described mechanisms, MCAR and MAR describe missing information that is recoverable which means that  $P(\mathbf{R})$  can be ignored when making statistical inferences based on the sample (Schafer, 1997). In contrast, methodologists note that the MNAR mechanism designates a severe limitation to the recoverability of lost information, which suggests that  $P(\mathbf{R})$  cannot be ignored because the resulting inferences are biased (see Enders, 2010). The fact that missing data can be grouped into ignorable and non-ignorable groups leads to the quandary of what to do when the data are non-ignorable (i.e., MNAR).

Methodologists refer to this as a widely misunderstood concept because missingness is typically a mixture of each mechanism (Enders, 2010; Graham, 2009; 2012). Pure mechanisms are unlikely in practice. The following section develops the concept of MAR and MNAR that are not distinct entities. Rather, these mechanisms are discussed in a more practical fashion in line with Graham's (2012) suggestions where MAR and MNAR form a continuum that cannot be easily teased apart.

*The MNAR to MAR continuum.* It follows from the above discussion that most, if not all, missing data situations involve a particular combination of missing data mechanisms (Graham, 2012; Yuan & Bentler, 2000). Figure 20 (adapted from Little, 2009) illustrates a theoretical overview of these missing data mechanisms, which will guide the forthcoming discussion. Note that in Figure 20, each mechanism contributes to the total percentage of missing data (illustrated by predictive arrows). Specifically, the total percent of missing data within a particular set of variables (termed, "Missing Data" in Figure 20) can be thought of as regressed on each of the three mechanisms where the weight of each regression coefficient varies depending on the particular variables selected from the larger data set. It is unclear how much of the total missing data is related to each mechanism, as this would vary in a given study.

In Figure 20, notice that MCAR and MAR are described as ignorable mechanism, while MNAR is denoted as non-ignorable; as previously discussed. Additionally, note the reciprocal relationships (i.e., double-headed arrows) in Figure 20. These paths illustrate no theoretical relationship between MCAR and the other mechanisms but a relationship of some magnitude between MAR and MNAR. This image demonstrates the point that MAR and MNAR are not distinct entities but are related.

To understand the relationship between MAR and MNAR (and lack of a relationship with MCAR), recall that each mechanism rests on an unlikely assumption. For instance, MCAR requires that missingness is *completely random*, which is only the case in planned missing data

designs that have no other missing data (Enders, 2010; Rubin, 1976). MAR requires that the researcher measure and include in the missing data estimation routine *all* reasons that cause missingness. Lastly, the MNAR situation involves unknown variables.

The only testable mechanism is MCAR (the randomness of  $\mathbf{R}$  can be assessed; see Little, 1998; Enders, 2010). Unfortunately, it is not possible to demonstrate that a particular variable (or set of variables) is the sole cause of missingness. Testing this assumption (that all variables responsible for missing data are recorded in the researcher's dataset) would require the missing values to be isolated from the influence of all other possible causes (Enders, 2010). Given that researchers are typically unaware of the precise causes of missing data, methodologists describe this as “an ultimately untestable assumption,” (see Enders, 2010).

Thinking more about the relationship between MAR and MNAR consider that they are more reasonably imagined as a continuum rather than as distinct entities because in reality missingness related to these mechanisms cannot be isolated. Figure 21 explicitly shows the relationship between MAR and MNAR to illustrate that virtually all missing data situations are partially MAR and partially MNAR (recall MCAR is ignorable and is thus not relevant for the current discussion).

Imagine that a given set of variables selected from a larger data set for analysis exists somewhere along the continuum in Figure 21. That is, the more that  $P(\mathbf{R})$  is explained by the observed data the closer the researcher moves on the continuum from MNAR toward MAR and the less likely inferences are to be biased. This situation lead Enders (2010) to note, “collecting data on the causes of missingness may effectively convert an MNAR situation to MAR,” (p. 15). Also consider that some variables in the larger data set are more informative than others and the researcher's ability to move away from biased inferences is impacted by the ability to explain the cause of missing data. As the research includes more information in the missing data handling routine he/she moves closer to unbiased inferences.

Consider that methodologists recommend modern missing data handling techniques (e.g., FIML) with the idea that the MAR assumption is satisfied. That is, researchers have in the larger data set and use in the modern missing data handling technique all relevant information such that the researcher positions him/herself completely to the MAR side of the continuum. In order to improve the likelihood of meeting this assumption (which is almost never actually met in practice), a strategy has emerged where researchers intentionally include extra variables (auxiliary variables) in hopes explaining why data are missing to generate unbiased inferences. This point was first discussed in relation to Rubin's 1976 work, again during the subsequent emergence of the EM algorithm and once more in the development of the FIML procedure. As these extra variables are included, it becomes more likely that MNAR data can be made to approximate MAR data – even when the actual cause of missingness is not observed (Enders, 2010; Little & Rubin, 2002). These correlates of missingness can be used to greatly improve estimation, an idea developed furthered in the upcoming discussion of auxiliary variables.

**Missing data patterns.** In addition to missing data mechanisms, Little and Rubin (1987) discuss missing data patterns, a topic not yet formally introduced. While missing data mechanisms describe *why* the data are missing, missing data patterns describe *where* the data are missing. As Enders (2010) noted, "...a missing data pattern simply describes the location of the "holes" in the data and does not explain why the data are missing, (p. 3). Figure 22 shows a simplified graphical depiction of four common missing data patterns including: (1) univariate, (2) multivariate, (3) monotone, and (4) general pattern (see Panels A – D, respectfully). Notice that each Panel in Figure 22 (i.e., missing data pattern) is composed of a smaller set of numbered rows where shaded regions indicate missing data. Panel A of Figure 22 represents a univariate missing data pattern. This pattern is composed of cases with complete data (pattern 1) and cases missing data on  $Y_2$  (pattern 2). Panel B simply extends this idea to more than one variable (multivariate nonresponse). Panel C represents monotone nonresponse, a situation where

participants drop out over time where  $Y_1 - Y_p$  represent repeated measures. Lastly, Panel D is the most common pattern (Enders, 2010) where missing values are scattered throughout the data set.

Methodologists are typically interested in missing data patterns in the context of how well parameters can be estimated (Little and Rubin, 2002). That is, missing data patterns provide information about variable coverage, the proportion of data available to estimate a particular relationship (Enders, 2010). For instance, notice that in Panel D of Figure 22 there is no overlap of observations between variables  $Y_2$  and  $Y_3$ . Therefore, there is no direct information available to estimate the association between these two variables. This situation represents low coverage. In contrast, the variables denoted  $Y_1$  and  $Y_p$  in Figure 22 are shown to have many overlapping observations and thus more coverage as more participants are informing the estimation of the parameter. As will be discussed in the following section, the idea of coverage presents a more technical way to think about missing data and is typically a desirable supplement to percentage missing (Enders, 2010).

**Multiple imputation (MI).** Perhaps the first discussion of the idea of multiple imputation was provided by Donald Rubin's (1977) paper titled, *Formalizing Subjective Notations About the Effect of Nonrespondents in Sample Surveys*. Rubin began with a discussion of missing data on sample surveys in which researchers must, "...make subjective judgments about the effect of nonrespondents," (p. 538). He notes that these decisions are akin to hypotheses about the likely values of a sample statistic in a situation where the data are complete. In working through this discussion, he formalized these judgments using Bayesian techniques, "...to produce a subjective probability interval for this statistic," (Rubin, p. 538). In his paper, Rubin noted that it is important to locate a "background" variable, say  $X$ , which related to (i.e., conditional on) an analysis variable, denoted  $Y$ , and which is observed for participants with and without data on  $Y$ . According to Rubin, the expected values of  $Y$  for participants with and without data can be expressed in the context of the sample mean as:

$$\begin{aligned}
E(\bar{Y}_R) &= \alpha_R + \beta_R \bar{X}_R \\
E(\bar{Y}_{NR}) &= \alpha_{NR} + \beta_{NR} \bar{X}_{NR}
\end{aligned} \tag{35}$$

where  $\bar{Y}_R$  and  $\bar{X}_R$  are the observed sample means while  $\bar{Y}_{NR}$  and  $\bar{X}_{NR}$  are the unobserved sample means of  $Y$  and  $X$ , respectively. Additionally,  $\alpha_R, \beta_R, \alpha_{NR}, \beta_{NR}$  are initially unknown parameters that relate  $X$  to  $Y$ . According to Rubin, the parameters  $\alpha_R$  and  $\beta_R$  are easily estimated from the observed sample; while the values  $\alpha_{NR}$  and  $\beta_{NR}$  require a more delicate technique to account for the uncertainty their unobserved sample means. Therefore, Rubin used the Bayesian idea of a “prior distribution” to formalize possible values for  $\alpha_{NR}$  and  $\beta_{NR}$ . As Rubin wrote the procedure required the specification of, “...subjective parameters that specify the prior distribution of the nonrespondents’ parameters [ $\alpha_{NR}, \beta_{NR}$ ] given the respondents’ parameters [ $\alpha_R, \beta_R$ ] and thus summarize subjective prior notions about the similarity of nonrespondents and respondents,” (p. 539). The subjective parameters, as Rubin called them, were used to generate a probability interval for the mean of  $Y$  using the observed data as well as other observed background variables associated with the variables of interest. He considered the use of these background variables to be important in generating a good distribution of possible values (i.e., prior distribution) which was subsequently used to estimate an expected distribution (i.e., posterior distribution) that could be summarized to obtain the unknown estimates in the presence of missing data.

Expressed mathematically the posterior distribution was derived through Bayes’ theorem, which can be demonstrated as (see Hays, 1996; Enders, 2010):

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{36}$$

where  $P(\theta|Y)$  represents the posterior distribution for a parameter given the data,  $P(Y|\theta)$  is the sampling density of the observed data (which is equivalent to the likelihood of the parameter given the data; see Enders, 2010),  $P(\theta)$  represents the prior distribution, and  $P(Y)$  is the marginal probability a scaling factor. As Enders (2010) noted this technique is essentially a likelihood function (as described previously) weighted by prior information, which can be worded as:

$$\text{Posterior Distribution} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Scaling Factor}} \quad (37)$$

Figure 23 presents a graphical depiction of the Bayesian approach used by Rubin (1977) to address missing data. Note that the multiplication of the prior distribution by the likelihood function generates the posterior distribution. In this image the relative shapes of each distribution provide information about how informative they are. That is, the height and width of each distribution reflects the probability density for that distribution. For instance, the prior distribution in this example is not very informative as it is relatively flat. An uninformative prior distribution could be represented as a uniform distribution. While Rubin’s paper does not explicitly discuss “multiple imputation”, it is an important precursor in that it marks a shift from a likelihood-based framework in thinking about missing data to a Bayesian perspective.

The first published account of multiple imputation can be traced to Rubin (1978a) with the publication of, *Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse*”. In this article Rubin discussed single imputation methods noting, “the approach that imputes one value for each missing datum is quite standard in practice, although often criticized by some mathematical statisticians who may prefer to think about estimating parameters under some model [i.e., likelihood-based methods], ” (p. 20-21). Rubin goes on to discuss two reasons why he is sympathetic to “the imputation position” over likelihood-based approach. Firstly, Rubin notes that likelihood-based techniques focus on

hypothetical models and not on the observed data which is, "...often not what the applied person wants to do since a hypothetical model is simply a structure that guides him to do sensible things with observed values," (p. 21). Secondly, he adds that in sample surveys (e.g., large public datasets) the data will be used to answer numerous questions by various researchers. Therefore, it is not practical to "remodel" the missing data process each time a new question is asked. Rather, it is more reasonable to address missing data once using all of the available data. Then, he argues, any research questions can be drawn from this complete sample.

After explaining the rationale for imputation, Rubin describes the difference in substitution procedures like single regression-based imputation and his new procedure that involves imputing a missing observation many times with each imputation differing slightly based on probability sampling. As Rubin noted, "...the missing values have a distribution given the observed values. Hence, what we really want to impute is not a single value but the predictive distribution of the missing values given the observed values," (p. 21). Rubin highlights the sophistication involved in generating plausible values from multiple "reasonable" models noting that each new dataset essentially draws a random sample from a distribution of possible values (based on a prior distribution) to replace a particular missing value. His idea was to substitute each missing value with a regression-based estimate of that value many times to capture the uncertainty surrounding any particular imputed value. Figure 24 provides a depiction of the multiple imputation process adopted from Rubin (1978a). Let rows in Figure 24 represent each of  $n$  participants and the columns are each of  $p$  variables ( $Y$ ) where the diagram represents a data set and each individual box is a specific data point. Missing data is indicated by the start point of an arrow and each arrow head points to a vector (denoted as "...") of plausible values that are imputed for a particular missing value.

There are two additional points of historical interest worth noting. First, Rubin (1978a) stated that when multiple datasets are generated; they differ only in the imputed values, as the

complete data remain unchanged across imputations. Therefore, he mentioned that it is possible to decompose the total variability in a particular data set into a between imputation (variability occurring due to repeated sampling from a probability distribution of plausible values) and a within imputation (the average variability within the imputed data sets) component. Based on these sources of variability, Rubin developed diagnostic measures for quantifying the uncertainty associated with the imputation process and thus the ability of his multiple imputation approach to recover lost information. These diagnostic tools formalized the ideas related to missing information previously mentioned with regard to the work of Orchard and Woodbury (1972), which will be later discussed in more detail.

Secondly, Rubin (1978a) made it clear that variability related to missingness is closely related to the information utilized in the imputation procedure. As he stated, "... having a large number of background variables that are recorded for all units and are highly correlated with variables that may be missing reduces the variation in the imputed values across repeated imputations," (p. 25). This idea recounts the point made previously that it is best to handle missing data once using all of the available information rather than "remodeling" for each research question. Specific research questions, he argues, cannot take full advantage of the entirety of available information. This point also supports the idea of satisfying the MAR assumption by including all relevant causes and correlates of missingness.

Rubin's (1978a) paper was motivated by sound theory and offered a novel approach for addressing missing data but does not appear to have popularized the technique. While multiple imputation (MI) can be thought of as an effective combination of the well-known likelihood-based methods and the regression imputation-based methods, the approach was initially seen as "...very exotic and computationally impractical," (Little, 2011, p. 4). Though there were a few other papers published on the idea of MI (e.g., Rubin, 1978b), it was likely the publication of Donald Rubin's (1987) book titled, "Multiple Imputation for Nonresponse in Surveys", which

largely popularized the idea of MI in the social and behavioral sciences (see Enders, 2010; Graham, 2009). The application of MI as a practical tool in statistical software can be traced back to Joe Schafer's NORM program which was developed following his 1997 book titled, *Analysis of incomplete multivariate data*. Schafer's algorithm was adapted by SAS for the Proc MI procedure; a now popular normal model based multiple imputation estimation routine (see Yuan, 2000).

Rubin (1987) described MI in relation to three sequential steps including: (1) an imputation step, (2) an analysis step, and (3) a pooling step (see Figure 25). The idea of the imputation step is to substitute each missing value with a regression-based estimate of that value. This is done more than once (usually at least 20 times; see Bodner, 2008; Graham, Olchowski, & Gilreath, 2007) because, as previously noted, a single imputation does not add enough variability to the imputed value. This process creates multiple copies of the original data set with each copy of the data set slightly varying on each of the imputed values. Then, the analysis step is used to individually analyze (i.e., run a particular statistical model of interest) each of the imputed data sets using standard techniques for complete data. Finally, the pooling step is used to combine the results from all these repeated analyses into a single result (see Rubin, 1987). In the following discussion, each of these steps is discussed in turn.

***The imputation step.*** The basic idea for the imputation phase is to inform a series of multiple regression equations by using all of the available information in the dataset (i.e., prior information) to inform the prediction of each missing value, including the observed values from cases with data on some, but not all, variables (Buhi, et al., 2008; Enders, 2006). The following discussion describes the mechanics of this procedure in more detail beginning with a discussion of the posterior (i.e., predictive) distribution.

To begin with, let  $P(Y|\theta)$  denote the sampling density of the observed data (i.e., the likelihood function) assume  $P(\theta)$  represents the prior distribution, and again let  $P(\theta|Y)$  refer to the posterior distribution. According to Rubin (1987) the  $m$  multiple imputations (e.g.,  $Y_{mis}^1, Y_{mis}^2, \dots, Y_{mis}^m$ ) are independent draws of size  $n$  from the posterior predictive distribution which can be defined as:

$$P(\theta | Y_{obs}) = \int P(\theta | Y_{obs}, Y_{mis}) P(Y_{mis} | Y_{obs}) dY_{mis} \quad (38)$$

where the posterior distribution of  $\theta$  given the observed data equals the observed data posterior distribution of  $\theta$  averaged over  $m$  imputations. That is, each imputation is an independent draw from a posterior predictive distribution of the parameter of interest given the observed data. Then, a sampling distribution is generated by averaging these parameter estimates from each of the repeated samples (Rubin, 1987). Some methodologists explain this process as a type of Monte Carlo simulation (see Tanner & Wong, 2010). The following discussion provides a brief explanation for the inclusion of Bayesian estimation techniques in the Monte Carlo process underlying MI estimation.

With simple missing data patterns, Little and Rubin (2002) state that simulating random draws from the posterior distribution can be relatively straightforward. That is, the process does not require Bayesian sampling techniques to generate a solution. However, the process grows more complex with each additional missing data pattern. Enders (2010) explains that each missing data pattern requires a different regression equation that must be solved simultaneously. Given a typical missing data pattern (e.g., a general missing data pattern) the process of sampling from the posterior distribution (in order to draw values for the imputation of missing data) generally requires an iterative algorithm (see Little & Rubin). Rubin (1987) used Bayesian inference as an iterative way to draw samples for this Monte Carlo simulation imputation

process. The general sampling approach for this purpose is the Markov chain Monte Carlo (MCMC) method (Enders, 2010).

The MCMC estimation technique allows for data augmentation, which makes sample data easier to analyze (Enders, 2010). To explain this idea, recall that the method of MLE searches for population parameters that are most likely given the observed data and that an iterative technique is usually needed to audition various population parameters. In this context, a MCMC sampling method has already been discussed in relation to the EM algorithm presented by Dempster, Laird, and Rubin (1977). That is, the EM algorithm is a MCMC method that was used to augment the MLE procedure given the lack of a closed form solution (e.g., deriving from complex missing data patterns; see Tanner & Wong, 1987). The term “data augmentation” is typically used by methodologists to refer to a variation of the EM algorithm popular in the context of Bayesian inference (Enders, 2010; Tanner & Wong, 1987; 2010). The difference in EM and data augmentation was expressed by Tanner and Wong (1987) as, “...we are interested in the entire likelihood or posterior distribution, not just the maximizer and the curvature of the maximizer [as are used with the EM algorithm],” (p. 529).

As Enders (2010) noted, data augmentation cycles between an imputation step (I-step) and a posterior step (P-step) which are similar to the EM algorithm steps discussed in the context of MLE. As stated above these methods are both MCMC sampling techniques. More precisely, the I-step corresponds to the E-step (expectation) and the P-step relates to the M-step (maximization; see Enders, 2010; Little & Rubin, 2002). The I-step generates a set of regression equations to predict (i.e., via single imputation) the incomplete data from the observed data (Enders, 2010). Afterwards, a random error term is added to each imputed value and then the following P-step uses the now complete data to estimate the population parameters, including the mean vector  $\mu$  and the variance/covariance matrix  $\Sigma$ . Next, another cycle of regression imputation is carried out which is linked to another formulation of the parameters  $\mu$  and  $\Sigma$ . In this

manner sampling from the posterior distribution is possible within the context of complex missing data patterns. Appendix H provides a demonstration of the data augmentation procedure using example data in Figure 26.

Note that a Markov Chain sequences random numbers such that newly generated values are pseudo-random because they depend on values generated just prior (Tanner & Wong, 2010). Said differently, the imputation routine implemented by data augmentation has a built-in autocorrelation associated with the iterations. Therefore, a Markov chain must be long enough (i.e., the i-step to p-step cycle) such that the parameter distribution stabilizes (Enders, 2010). Figure 27 shows a plot of the simulated parameter values from the P-step discussed previously. Notice that the Y-axis contains a range of mean estimates for  $Y$  and the X-axis displays the number of data augmentation cycles. This image demonstrates iterative variation associated with each imputation. Notice that at the 10<sup>th</sup> iteration (i.e., data augmentation cycle) the  $Y$  parameter was 19.5 and at the 30<sup>th</sup> iteration was 20.2, which are close to the estimates derived above (e.g., 19.065 and 19.413). It is worth noting that the fraction of missing information, as discussed above (and in more detail in the next section), is responsible for the sampling variability seen across iterations in Figure 27. That is, more missing information (i.e., less coverage) implies a less stable pattern (i.e., posterior distribution).

***The analysis step.*** The Given  $m$  imputed data sets, the next step is to analyze each complete dataset separately using complete case procedures (refer to Figure 25). To illustrate this step, suppose that  $m = 5$  and that it is of interest to estimate a bivariate regression between a predictor ( $X$ ) and an outcome ( $Y$ ). The bivariate linear regression equation that predicts  $Y$  from  $X$  is defined as

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + e_i \quad (39)$$

where  $\beta_1$ , is the unstandardized regression coefficient and represents the average change in  $Y$  for each unit change in  $X$ . The constant  $\beta_0$ , is the intercept (i.e., the value of  $Y$  when  $X$  is 0). The error term  $e_i$  is the error associated with the prediction of  $Y$  from  $X$  (Hays, 1994). This equation is applied to each of the  $m$  datasets. Figure 28 illustrates an exemplary bivariate regression analysis between  $X$  and  $Y$  for each of the  $m$  imputation where the one-way predictive arrow demonstrates the ability of  $X$  to predict  $Y$ .

Using the same data demonstrated previously in Appendix H, let the estimated regression coefficient  $\beta_1$  represent the parameter of interest  $\theta$  from each  $m$  imputation such that:

$$\begin{aligned}
 \beta_1^{(1)} &= 1.041 \\
 \beta_1^{(2)} &= 1.068 \\
 \beta_1^{(3)} &= 1.061 \\
 \beta_1^{(4)} &= 1.026 \\
 \beta_1^{(5)} &= 1.057
 \end{aligned}
 \tag{40}$$

where the superscript in parentheses ranges from 1 to 5 and represents the imputation number. Notice that these values vary slightly to reflect uncertainty about the estimated regression coefficient  $\beta_1$  given missing data. After the analysis step, the goal is generally to combine the  $m$  sets of results (e.g., means in the current example) into a final estimate of the parameter  $\theta$  with its associated variance.

***The pooling step.*** While advances in personal computing power in the late 1980's made MI more practical, a contributing factor for the success of Rubin's book was "Rubin's Rules", a set of formulas used to combine estimates and standard errors across  $m$  imputation results. Previously, there seems to be a noticeable lack of clear and practical guidance for combining the  $m$  imputed results. Rubin (1987) noted that this process is easy for the parameter estimate and can be expressed as:

$$E(\theta | Y_{obs}) = E\left[E(\theta | Y_{obs}, Y_{mis}) | Y_{obs}\right] \quad (41)$$

where the posterior mean of the parameter  $\theta$  equals the average complete data posterior mean of  $\theta$ . This expression can be expressed as (see Enders, 2010):

$$\hat{\theta} = \frac{1}{m} \sum_{t=1}^m \hat{\theta}_t \quad (42)$$

where  $\hat{\theta}_t$  is the point estimate (e.g.,  $\beta_1$ ) from imputation  $t$  and  $\hat{\theta}$  represents the parameter averaged over all  $m$  imputations. Continuing with the current example from Appendix H, the means are averaged such that:  $(1.041 + 1.068 + 1.061 + 1.026 + 1.057) / 5 = 1.0506$ . This value represents the pooled point estimate for  $\beta_1$  across all imputations.

Note that each of the averaged beta weights has an associated standard error that can also be pooled; however, this information cannot be simply averaged (Rubin, 1987). Said differently, the variance is composed of sampling variability as well as imputation variability. Recall that when multiple datasets are generated; they differ only in the imputed values as the complete data remain unchanged across imputations. Therefore, it is possible to decompose the total variability into a between imputation and a within imputation component (Enders, 2010; Rubin, 1987). To begin with consider the within-imputation variance as it is simply the average sampling variance (i.e., squared standard error; see Enders, 2010) across each  $m$  imputation and can be written as:

$$V_w = \frac{1}{m} \sum_{t=1}^m s_t^2 \quad (43)$$

where  $s_t^2$  is the sampling variance from imputation  $t$  and  $V_w$  is the sampling error that would have resulted given no missing data (see Enders, 2010). Using the current example, let the within-imputation variance be:

$$\begin{aligned}
s_{t=1}^2 &= 0.075 \\
s_{t=2}^2 &= 0.076 \\
s_{t=3}^2 &= 0.073 \\
s_{t=4}^2 &= 0.075 \\
s_{t=5}^2 &= 0.072
\end{aligned} \tag{44}$$

The sampling variances are averaged such that:  $V_w = (0.075 + 0.076 + 0.073 + 0.075 + 0.072) / 5 = 0.0742$ . The value 0.0742 represents the pooled point estimate for the variability that occurs within imputations. That is, this estimate does not consider the variability that occurred across each  $m$  imputation, which is considered next. The between-imputation variance is calculated by taking the sum of the squared deviations divided by the number of imputations and can be expressed as:

$$V_B = \frac{1}{m-1} \sum_{t=1}^m (\hat{\theta}_t - \bar{\theta})^2 \tag{45}$$

where  $\hat{\theta}_t$  is the parameter estimate from imputation  $t$  and  $\bar{\theta}$  is the pooled point estimate of  $\theta$  (see Enders, 2010). Given the current example data the between-imputation variance can be as:

$$\begin{aligned}
&\hat{\beta}_1^{(t)} - \bar{\beta}_1 \\
1.041 - 1.051 &= -0.010^2 = 0.00010 \\
1.068 - 1.051 &= 0.017^2 = 0.00028 \\
1.061 - 1.051 &= 0.010^2 = 0.00010 \\
1.026 - 1.051 &= -0.025^2 = 0.00063 \\
1.057 - 1.051 &= 0.006^2 = 0.00004 \\
&0.00115 / 5 = 0.00023 = V_B
\end{aligned} \tag{46}$$

The next step is to calculate the total variance. According to Rubin (1987) the final posterior variance of  $\theta$  can be found by:

$$V(\theta | Y_{obs}) = E[V(\theta | Y_{obs}, Y_{mis}) | Y_{obs}] + V[E(\theta | Y_{obs}, Y_{mis}) | Y_{obs}] \tag{47}$$

where the posterior variance of  $\theta$  equals the average repeated complete data variance of  $\theta$  plus the variance over repeated imputations. As Enders (2010) noted, this expression can be conveyed as:

$$V_T = V_W + V_B + \frac{V_B}{m} \quad (48)$$

where  $m$  denote the imputation number. Applied to the ongoing regression example yields the following estimate for the total variance:

$$V_T = .0742 + 0.00023 + \frac{0.00023}{5} = 0.074476 \quad (49)$$

Note that the overall standard error (which is used for significance tests and confidence intervals) is the square root of the total variance so that:  $\sqrt{0.074476} = 0.272903$  and confidence intervals can be constructed as usual to test the significance of a parameter (the estimated regression coefficient  $\beta_1$ ). For instance, given that  $\beta_1 = 1.051$  a 95% confidence interval can be computed as  $1.051 \pm 1.96 (0.272903) = 0.51611, 1.58589$ . Similarly, a  $t$ -value for each pooled estimate can also be calculated. This is done by dividing the mean parameter estimate (i.e.,  $\bar{\beta}_1 = 1.051$ ) by the overall standard error. The degrees of freedom for the  $t$ -test are calculated as follows (see Enders, 2010 for example):

$$df = \nu = (m - 1) \left[ 1 + \frac{V_W}{V_B + (V_B / m)} \right]^2 \quad (50)$$

The uncertainty associated with multiple imputations relates to the degree of missing information and can be expressed as the variability infused by the imputation process (Rubin, 1978a). For instance, an increase in the variability of imputed parameters (e.g.,  $V_B$ ) relates to a larger overall standard error for that estimate and thus more uncertainty surrounding any statistical inferences. The following discussion introduces two diagnostic measures including the

fraction of missing information, and the relative increase in variance in order to formalize the impact of missing data on the sampling variability (i.e., standard errors).

**Fraction of missing information.** Recall that Orchard and Woodbury (1972) discussed the idea of sampling variability due to missing data in the context of larger maximum likelihood standard errors due to inflated variance from missing data. Orchard and Woodbury understood that missing data have an influence on the efficiency of parameter estimates to the degree that the missing information increases standard errors. They discussed fraction of missing information in the context of their missing information principle.

The need to understand the impact of missing data on the parameter estimation (specifically the posterior distribution) was an important concept for both the EM Algorithm (see Dempster, Laird, & Rubin, 1977) and MI. In the context of MI (i.e., the Bayesian perspective), the fraction of missing information can be thought of as the additional variation caused by incomplete data. Rubin (1987) noted that as few as 3 to 5 imputations were adequate for MI provided that the fraction of missing information was not too high. That is, MI's efficiency for recovering missing data is related to the fraction of missing information and the number of imputations used.

Basically, datasets are of varying quality where “good quality” refers to decent coverage (as discussed previously) and informative associations among variables while “poor quality” consists of minimal coverage and a general lack of association across variables (i.e., less information to inform the missing data handling procedure). As Enders (2010) stated, the fraction of missing information (FMI) reflects the relationship between the missing data pattern and the magnitude of associations among variables where highly related variables with good coverage produce the lowest FMI. That is, the sampling variance is not greatly influenced by missing data. Conversely, weakly related variables with poor coverage generate a high FMI where the missing data may significantly inflate the sampling variance. Therefore, the data

augmentation routine does not contain enough information to appropriately handle the missing data. Conceptually, a high FMI is analogous to a relatively flat likelihood function and a low FMI with a more peaked likelihood function.

As Enders (2010) illustrated, there are two basic formulas for the fraction of missing information in the context of MI (see Savalei & Rhemtulla, 2012 for the FIML analog). Given a large number of imputations (e.g.,  $m > 100$ ), the fraction of missing information can be expressed as FMI in the equation below; or conversely as  $FMI_1$  given a small number of imputations:

$$\begin{aligned}
 FMI &= \frac{V_B + (V_B / m)}{V_T} \\
 FMI_1 &= \frac{V_B + (V_B / m) + 2(\nu + 3)}{V_T} \quad (51)
 \end{aligned}$$

where  $V_B, V_T$ , and  $m$  are defined as previously indicated while  $\nu$  represents the degrees of freedom (see Equation 63; Enders, 2010). As illustrated in Equation 64, the fraction of missing information is defined as the ratio of the additional variability caused by the missing data (i.e., between-imputation variability; which depends on the number of imputations) and the total variability. The interpretation of FMI is straightforward given the current regression example. For example,  $FMI = (.00023 + (.00023 / 5)) / .074476 = .003705$  so less than 1% (e.g., 0.4%) of the observed variance is related to missing data (while  $FMI_1$  is more appropriate for the current  $m = 5$  example, the less complex FMI formula was used for simplicity). The fraction of missing information will be considered in subsequent discussion regarding the use of auxiliary variables and data augmentation convergence.

**Relative increase in variance.** While the fraction of missing information provides the percentage of sample variance related to missing data, the relative increase in variance (RIV) yields a measure of the between-imputation variance relative to within-imputation variance such

that when  $RIV = 0$  the variance was not influenced by missing data and when  $RIV = 1$  the between-imputation variance is equivalent to the within-imputation variance. This measure provides an additional diagnostic tool for understanding of the effectiveness of recovering missing data and can be written as (see Enders, 2010):

$$RIV = \frac{V_B + (V_B / m)}{V_W} = \frac{FMI}{1 - FMI} \quad (52)$$

where  $V_B, V_W, V_T$ , and  $m$  are defined as previously noted. Here the denominator represents the variability that would have resulted had there been complete data and the numerator is the additional variability caused by the missing data. Therefore, RIV provides a measure of the proportional increase in variance due to missingness (Enders, 2010).  $RIV = (.00023 + (.00023 / 5)) / .0742 = .003719$  which suggests the sampling variance was approximately .004 times larger than it would have been had the data been complete (i.e., the sampling variance increased by .37 %).

Both FMI and RIV provide information about the influence of missing data on the variance of parameter estimates and as Tanner and Wong (1987) noted the convergence of the data augmentation algorithm. As Enders (2010) pointed out FMI is typically not the same as the percent missing because FMI is adjusted for the presence of other variables that inform what the missing values would have been had they been collected. Thus, FMI and RIV are typically preferable descriptive indices for the influence of missing data on the significance tests of parameter estimates. Because FMI depends on the specific set of variables used, including extra variables that are highly correlated with the variables containing missingness may improve estimation and result in a FMI estimate that is far lower than the observed percent missing (Enders, 2010).

The idea that observed values provide indirect information about the likely values of missing data arose early in the history of missing data research (recall Allan & Wishart, 1930)

and continues through the following discussion. The following section includes research and theory related to the use of these extra variables to improve FIML and MI estimation procedures and provides simulation examples that demonstrate these principles. As will be discussed, much of this work was developed to more reasonably approximate the MAR assumption.

**Auxiliary variables.** Little and Rubin (1987) positioned the missing data problem within the more general context of quantitative methodology in the social and behavioral sciences, and discussed the implications of various missing data handling techniques. While there are many possible ways to treat missing data, most approaches are not recommended (e.g., long-standing traditional approaches like deletion and single imputation). They emphasized the importance of assumptions regarding why the data are missing as this can bias any inferences made from the data being studied. Little and Rubin noted that in most incomplete datasets, observed values provide indirect information about the likely values of missing data. This information, guided by statistical assumptions, can effectively recover missing data to the degree that the variables responsible for causing missing data are included in the missing data handling procedure (Allison, 2003; Collins, Schafer, & Kam, 2001; Enders, 2008; 2010; Graham, Cumsille, & Shevock, in press).

As previously discussed, MI and FIML are the most generally recommended missing data handling techniques in the methodological literature (Enders, 2010; Graham, 2009; McKnight, McKnight, Sidani & Figueredo, 2007); however, the effectiveness of these statistical tools depend on the researcher's ability to meet the MAR assumption (Enders, 2010; Graham, 2009; Little & Rubin, 1987). That is, MI and FIML only yield unbiased parameter estimates when all variables that are causes or correlates of missingness are included in the missing data handling procedure (Enders, 2010; Little & Rubin, 2002). Methodologists have recommended extra variables (auxiliary variables) to address this issue. Auxiliary variables are not the focus of an analysis but are instead used to inform the missing data handling procedure (e.g., FIML, MI)

by adding information about why a particular variable has missing data or describing the probability that a particular case has missing data (Collins, Schafer, & Kam, 2001; Enders, 2010). Therefore, auxiliary variables support the MAR assumption and improve estimation (Collins, Schafer, & Kam, 2001; Enders & Peugh, 2004; Graham, 2003).

To illustrate this point, consider a simple research scenario using a two-group experimental design with a treatment and control group where researchers want to know how children perform on language outcomes when they are exposed to a new intervention. Now suppose that low socio-economic status (SES) resulted in missing data on a particular variable but the researcher did not include SES in the FIML or MI missing data handling procedure because it was not part of the analytic model of interest (e.g., no research hypotheses were interested in SES effects). The problem with omitting an influential variable (e.g., SES) from the imputation phase of MI or from the modeling phase of FIML is that this situation has been known to attenuate that variable's association with other variables in the subsequent statistical analysis (e.g., biased toward zero) leading to incorrect estimates (Enders, 2010). Said differently, in the previous example SES was not used to inform the missing data estimation process, therefore, information from SES is unavailable for explaining why the data are missing. Consequently, the estimated relationships derived from any ensuing statistical modeling are uninformed that participants with low SES look very different from participants with average or high SES. Therefore, even though SES was not a variable of interest in the analysis model, it was the cause of missingness in the researcher's dataset and may hence generate deviations from probability sampling (i.e., the sample is a mixture of several probability distributions) that lead to invalid inferences of the original intended sample (see Rubin, 1976).

Therefore, SES is an auxiliary variable and should be included in the missing data handling model but not in the analysis model so that the subsequent statistical modeling is unbiased and the actual model of interest is not altered. Current research has argued in favor of

the use of auxiliary variables to recover missing information to reduce bias (Collins, Schafer, & Kam, 2001; Enders & Peugh, 2004; Graham, 2003), decrease standard errors, and increase statistical power (Collins, Schafer, & Kam, 2001; Enders, 2010). These benefits arise even when the auxiliary variables are unrelated to the cause of missingness but correlated with the variable that has missing data (see Collins, Schafer, & Kam, 2001; Enders, 2010; Yoo, 2009).

Consider an extension of the previous SES example to advance this concept. Once more, assume that low SES was the cause of missing data on a particular variable but this time the researcher did not measure SES. However, suppose the researcher did measure a proxy of SES such as “who got free or reduced lunches” and included this variable as an auxiliary variable. The “free or reduced lunches” variable was not the cause of missingness in this example; however, it is reasonable to assume it is correlated to some degree with SES. Because the variable “free or reduced lunches” is based on salary and income information the researcher is likely to get back some of the missing information that is due to low SES, to the degree that the “free and reduced lunches” correlates with SES. Taking this idea a step further, suppose that the researcher might also have measured other variables that also correlate with SES. For instance, consider variables that might be a product of having been in a certain SES environment; like parental involvement and motivation. When used as auxiliary variables, each of these proxy variables contributes some information to the missing data estimation process that will help recover the cause of missingness even though the actual cause of missingness was not observed.

To illustrate the potential benefit of auxiliary variables mathematically, suppose the researcher in this example had measured SES and then investigated the effects of free or reduced lunch ( $W$ ), parental involvement ( $X$ ), and motivation ( $Z$ ) on SES ( $Y$ ). The regression equation for predicting the outcome variable (SES) from the three predictor variables can be written as:

$$\hat{Y} = b_o + b_1W + b_2X + b_3Z \quad (53)$$

where  $\hat{Y}$  represents the predicted value of SES ( $Y$ ), the  $b_0$  coefficient represents the intercept and the  $b_1$ ,  $b_2$ , and  $b_3$  coefficients denote the linear relationship of free or reduced lunch, parental involvement, and motivation to SES, respectively. Given this equation, assume that the proportion of variance in the dependent variable (SES) which can be explained by the independent variables (free or reduced lunch, parental involvement, and motivation) is somewhere between a reasonable range of  $r^2 = .30 - .40$ . This would indicate that 30 - 40% of the variance in these three proxy auxiliary variables will help the researcher account for the MNAR mechanism caused because SES was not actually measured. Given a large set of good auxiliary variables, the research analyst may more reasonably assume the MAR assumption because the missing data handling procedure is as informed as possible. Additionally, the resulting FMI estimates are in theory lower than the percent missing. As Enders (2009) noted, the effectiveness of auxiliary variables has been demonstrated even when the auxiliary variables, themselves, contain relatively high levels of missing data.

*Auxiliary variables for nonlinear causes of missingness.* In practice, it is often the case that complex causes of missing data are not adequately captured by additive linear regression equations (Collins, Schafer, & Kam, 2001). Assume that missing data caused by low SES (from the previous example) was actually caused by motivation as it relates to having been in a certain SES environment and additionally parental involvement alleviates the effect of motivation on the propensity of missing data. Therefore, while each individual's level of motivation is negatively related to his or her likelihood of having missing data, the strength of this relationship weakens as the level of parental involvement increases. By including an interaction term between motivation and parental involvement, the researcher can effectively remove bias from the missing data handling procedure (Collins, et al., 2001; Schafer, 1997).

Typically, researchers assume linear relationships when using MI or FIML (Little and Rubin, 2002); however, the inclusion of product terms or powered terms can add variables that

can approximate non-linear processes (Enders & Gottschall, 2011; Graham, 2009; Schafer, 1997). Therefore, missing data related to a non-linear function can be included in the missing data handling process to explain why data are missing. As Enders and Gottschall (2011) demonstrated, in some cases, knowledge of possible non-linear causes of missingness could be important to satisfy the MAR assumption and lead to unbiased inferences. It seems reasonable that a true *all*-inclusive strategy should incorporate interaction terms among the auxiliary variables as well as polynomials to capture potentially useful non-linear information.

***Auxiliary variable demonstration using simulated data.*** In order to demonstrate the bias that can result when auxiliary variables are ignored during imputation, a series of Monte Carlo simulations were conducted in line with those presented by Enders (2010). These simulations intend to imitate a research scenario where it is of interest to estimate the mean vector and covariance matrix between two variables,  $X$  and  $Y$ . These variables could be any combination of continuous or categorical scales. Of current interest is the case where  $X$  and  $Y$  are bivariate normally distributed continuous variables.

Now, suppose that  $Y$  is subject to nonresponse such that a monotone missing data pattern is obtained. Figure 29 illustrates a diagram of this monotone missing data pattern where all values of the variable  $X$  are observed and only some of the values of  $Y$  are observed. Notice that in Figure 29 there are  $m$  cases of  $Y$  and  $p$  cases of  $X$ . There are numerous research situations that lead to this pattern of missing data and it has been widely used in the literature to study missing data (Enders, 2010; Little & Rubin, 2002). For instance, let  $Y$  represent a questionnaire item with missing data and let  $X$  represent a demographic variable, like age, with no missing values. The variable  $X$  could also represent a fixed variable controlled by the experimenter or any other variable with complete data. This analysis model is certainly not comprehensive, but is sufficient to show the potential effectiveness of auxiliary variables.

For simplicity, the manipulated factors only included: (a) magnitude of correlations between the auxiliary variables and the variable with missingness, and (b) missing data mechanism. The rate of missingness was fixed at 30% and the sample size was set to  $N = 100$ . These factors were chosen due to their expected influence on the performance of missing data handling procedures that incorporate auxiliary variables. For example, the effectiveness of auxiliary variables is expected to improve as the magnitude of correlations increase given a particular rate of missing and sample size (see Collins, Schafer, & Kam, 2001; Enders, 2010). Additionally, the performance of auxiliary variables is expected to improve parameter estimation even in a MNAR condition (in the absence of a missing data cause; see Enders, 2010). The benefits of auxiliary variables are also considered in relation to non-linear MAR. Lastly, given a MCAR condition, auxiliary variables do not influence bias but can instead demonstrate relative improvements in power compared to the exclusion of auxiliary variables (Enders, 2010).

The following levels were selected: (a) magnitude of correlations ( $\rho = .10$  to  $.80$  by  $.10$ ), (b) missing data mechanism (MCAR, MAR, non-linear MAR, MNAR). For all variables the means were set at  $\mu = 0$  and the variances were set at  $\sigma^2 = 1.0$ . Results are expressed across the range of correlations for each of the missing data mechanisms. The goal is to provide a clear (though simplified) demonstration of auxiliary theory with simulated data. For each study, a bivariate normal correlation model was used to generate population data for the associated simulation studies. Figure 30 illustrates a path diagram of the population model including auxiliary variables. Notice that the population correlation between the variables  $X$  and  $Y$  were fixed at  $\rho = .30$  for simplicity as were the associations between  $X$  and the three auxiliary variables ( $Aux_1 - Aux_3$ ). In contrast, the magnitude of associations between  $Y$  (the variable with missingness) and the auxiliary variables range from  $.10$  to  $.80$  by  $.10$ .

Specifically, *Mplus* 6.0 (Muthén & Muthén, 1998–2010) was used to generate five variables:  $X$ ,  $Y$ , and three auxiliary variables ( $AUX_1 - AUX_3$ ) in 1,000 multivariate normal data

sets of size 1,000 for each set of population values noted previously. After data generation, SAS 9.3 (SAS Institute Inc., 2011) was used to impose one of the four missing data mechanisms on  $Y$ . After generating missingness, the data from all replications were analyzed with *Mplus* 6.0 where the auxiliary variables were incorporated into the FIML estimation routine with the saturated correlates model (see Graham, 2003).

*Excluding a cause of missingness.* For the first study, a MAR missing data pattern was specified where missing data were generated by deleting 30% of the data in  $Y$  based on the lower distribution of the first auxiliary variable ( $AUX_1$ ). This simulation was intended to show the impact of excluding a cause of missingness (i.e.,  $AUX_1$ ). That is, the missing data on  $Y$  are MNAR because the cause of missingness was not included in the subsequent FIML procedure. As illustrated in Table 4, the impact of omitting a cause of missingness depends on its association with the incomplete variable (i.e.,  $Y$ ). Specifically, a negative bias is present that increased as the relationship between the auxiliary variable and the variable with missingness increased.

This relationship is depicted graphically in Figure 31, where the plotted line slopes down and to the right demonstrating increasing negative bias. The direction of this bias directly relates to the portion of  $AUX_1$  that was responsible for missingness and does not suggest a typical direction for bias in relation to excluding a cause of missingness. Deleting 30% of the data in  $Y$  based on the upper distribution of the first auxiliary variable ( $AUX_1$ ) would produce positive bias. Because the missingness was entirely related to the omitted variable, including it in the estimation process would result in unbiased estimates.

These findings support and replicate those described by Enders (2010). As he found, "...omitting a correlate of missingness was not that detrimental when the correlation was weak (e.g.,  $r \leq .30$ ), but the estimates became increasingly biased as the correlation between the auxiliary variable and the incomplete analysis variable increased in magnitude," (p. 129). This

relationship explains the suggestion provided by Collins, Schafer, and Kam (2001) to locate auxiliary variables that exceed an association of about  $r = .40$  with the variables with missing data.

*Improving estimation with MNAR.* For the second study, the data were generated as previously noted except a MNAR missing data mechanism was specified where missing data were generated by deleting 30% of the data in  $Y$  based on the lower distribution of the variable  $Y$  itself. In this example, all three auxiliary variables ( $AUX_1 - AUX_3$ ) are included in the FIML estimation routine. This simulation was intended to show the ability of auxiliary variables to improve parameter estimation even in a MNAR situation. The results of this simulation are presented in Table 5 and Figure 32 and indicate auxiliary variables that are related to a variable with missingness may reduce (through not eliminate) parameter bias. As before the performance of auxiliary variables to reduce bias depended on the magnitude of the association between the incomplete variable (i.e.,  $Y$ ) and the auxiliary variables. These findings support a similar simulation by Enders (2010). He noted that under the MNAR condition, “although the average parameter estimates did get closer to the true population values as the correlation increased, the bias was still noticeable, even when the correlation between the auxiliary variable and the missing analysis variable was 0.80,” (p. 130). As most researchers are unaware of all possible causes of missingness, these findings encourage the use of auxiliary variables with strong associations despite theoretical motivation for selecting auxiliary variables.

*Including non-linear MAR.* For the third study, a non-linear MAR mechanism was generated by deleting 30% of the data in  $Y$  based on the lower distribution of an interaction term between the first and second auxiliary variable (i.e.,  $AUX_1 \times AUX_2 = AUX_{12}$ ). The primary goal of this simulation was to study parameter estimate bias when the cause of missingness was not a linear process given that FIML assumes linear relationships among the analysis and auxiliary variables.

To investigate this issue, two conditions are presented. First, the first two auxiliary variables ( $AUX_1$  and  $AUX_2$ ) were included in the FIML estimation routine. This approach did not contain the non-linear interaction term that actually caused missingness on  $Y$ ; however, it did include the linear components from which the interaction term was generated. Second, only the newly created interaction term ( $AUX_{12}$ ) was included and the linear components were excluded. This condition represents a lower bound on the effect of including a non-linear term as it used one auxiliary variable rather than two (though the included term was theoretically more meaningful).

The results of this investigation are presented in Table 6 and demonstrate the greater bias in the linear condition (first approach that included only  $AUX_1$  and  $AUX_2$ ) than in the non-linear condition (the second condition that included only  $AUX_{12}$ ). Said differently, while the average parameter estimates approached the true population values as the correlation among the auxiliary variables and  $Y$  increased across both conditions, the bias was slightly more prevalent in the linear only condition. These findings support prior work in the area of non-linear causes of missingness (see Collins, Schafer & Kam, 2001) and demonstrate the viability for non-linear auxiliary variable terms in FIML estimation. Further, these results present another problem as including non-linear terms (i.e., interactions) preclude the inclusion of the accompanying linear components because of estimation errors due to multicollinearity (see Graham, 2009 for discussion of the resulting non-positive definite covariance matrix). A solution to this problem is presented in an upcoming section where principal components are used to extract linear and non-linear variance from auxiliary variables (as will be discussed principal components are, by definition, uncorrelated).

*Improving statistical power.* For the last demonstration of the ability of auxiliary variables to improve estimation, a MCAR missing data pattern was specified where missing data were generated by deleting 30% of the data in  $Y$  based on a completely random process. Like the previous simulations, this study was modeled after work by Enders (2010) and intended to

demonstrate the impact of auxiliary variables on statistical power. While the previous examples generated biased parameter estimates by omitting the cause of missingness, this demonstration did not because missingness was related to a completely random process where any value of  $Y$  was as likely to be missing as any other value. Enders referred to this as a “benign” cause of missingness; thus an assessment of bias in this situation is uninformative. Instead, this investigation focused on estimates of statistical power relative to the association strength among the variable with missingness and all three auxiliary variables.

The results of this simulation suggest that statistical power increases as the magnitude of associations increase between the analysis variables with missingness and the auxiliary variables. These findings are shown in Figure 33 where, for example, a correlation of  $r = .30$  produces a relative power value of 1.028, which suggests that the auxiliary variable increased power by about 2.8%. Likewise, a correlation of  $r = .40$  relates to a relative power value of about 4%. As demonstrated in Figure 33 the relationship between power and association strength is non-linear where the largest gains in power occur when the correlation exceeds about  $r = .40$ . This finding further supports the suggestion of methodologists (e.g., Collins, Schafer, & Kam, 2001) to choose auxiliary variables that are highly correlated with the outcome measures.

*Auxiliary variable simulation summary.* As demonstrated, auxiliary variables used in conjunction FIML (or MI though not demonstrated here for clarity) are important tools for researchers. Methodologists have generally endorsed an “inclusive strategy” with regard to auxiliary variables, which recommends the generous use of all possible auxiliary variables from a dataset rather than a “restrictive strategy” which incorporates a selected subset of auxiliary variables (see Collins, Schafer, & Kam, 2001). As Collins, et al., note, the “inclusive strategy” is recommended to reduce the chance of inadvertently omitting an important cause of missingness while allowing for noticeable gains in terms of increased efficiency (e.g., power) and reduced bias. The simulations presented explicitly demonstrate these ideas as they are important concepts

for the following discussion. While thus far little attention was given to the process of choosing auxiliary variables aside from locating extra variables with strong associations, methodologists have given considerable attention to the topic and the issue is relevant for the ensuing discussion. Next, the topic of choosing auxiliary variables is elaborated on including rationale for not always selecting a few strongly correlated auxiliary variables.

*Methods for choosing auxiliary variables.* While the various sources of evidence presented to this point provide a clear indication that the inclusion of auxiliary variables can have an important impact on the results of any study with missing data, it is not clear what steps should be taken to ensure that *all* important information relating to missing data are included in the subsequent missing data handling procedure. Generally, it seems that the approaches taken to choose auxiliary variables can be divided into two categories. The first approach (perhaps the most popular approach; see Graham, Cumsille, & Shevock, in press) suggests that only a few auxiliary variables (if any) are necessary to adequately satisfy the MAR assumption. Collins, Shafer, and Kam (2001) refer to this approach as a “restrictive auxiliary variable strategy”. It is clear that researchers utilizing this approach must assume that all relevant information is included (Little & Rubin, 2002). Thus, from the restrictive strategy perspective, the researcher must be comfortable with the assumption that any potential cause of missing data is explained by the selected auxiliary variables.

The second approach begins with the assumption that results can vary with the auxiliary variables chosen. Therefore, instead of focusing on a few influential auxiliary variables this approach emphasizes a liberal use of auxiliary variables (see Collins, Shafer, & Kam, 2001). The emphasis of this approach is on the incorporation of all possible information; which reduces the chance of accidentally omitting an important cause of missingness. Collins, et al. refer to this approach as an “inclusive auxiliary variable strategy”. This strategy suggests a way to conduct research that is theoretically more in line with the MAR assumption because the researcher is

more likely to at least partially explain the cause of missingness through proxy variables (as discussed previously). Thus, if the data set contains useful variables which are incorporated into FIML or MI estimation process, the result should be an improvement in the quality of findings in terms of noticeable gains in efficiency and reduced bias (see Collins, et al., 2001; Enders, 2010; Enders, 2008).

Theoretically, the inclusive approach is superior to the restrictive strategy in that the methodological literature notes there are no known dubious effects of including too many auxiliary variables (e.g., Collins, Shafer, & Kam, 2001). As Collins, et al., note, “our results indicate that far from being harmful, the inclusion of these variables is at worst neutral, and at best extremely beneficial,” (p. 348). Though more work is needed to fully understand this topic (i.e., simulations are based on specific conditions that may not generalize to practical applications), the inclusive strategy seems preferable. Still, methodologists note that in practice it is the restrictive approach that is typically used (see Enders, 2010; Graham, 2009). In order to explain this discrepancy, the following section will further discuss the restrictive strategy and review recommendations that methodologists make for choosing a restrictive set of auxiliary variables.

*Finding a few influential auxiliary variables.* As Collins, Shafer, and Kam (2001) noted, the approach that focuses on finding a small set of most relevant auxiliary variables is the most common in part because this approach is favored by existing software and the associated documentation because, “...neither of which encourages users to consider auxiliary variables or informs them about how to incorporate them,” (p. 349). While progress has been made to revise maximum likelihood-based programs (see the AUXILIARY (m) option in *Mplus* 6.0; Muthén & Muthén, 1998–2010) and incorporate more informative documentation regarding the use of auxiliary variables (see SAS Institute Inc., 2011), the restrictive strategy persists in the applied literature. Therefore, it is important to consider how best to select influential auxiliary variables.

Enders (2010) acknowledged the inherent difficulty in establishing a rule of thumb for selecting a particular set of auxiliary variables but discusses a few reasonable approaches beyond simply looking for high correlations (though he does suggest this can be effective). He proposes that researchers “maximize the squared correlation between the auxiliary variables and the analysis model variables, and to do so with as few auxiliary variables as possible,” (p. 133). He also suggests a related approach using an omnibus test for the missing completely at random mechanism (Little’s MCAR test; see Little, 1998). If the data are not missing due strictly to a random process, researchers can proceed to determine which specific variables contribute to missingness (where complete and incomplete data groups are created for each variable). Then, independent *t*-tests can be used to compare mean differences among potential auxiliary variables (see Dixon, 1988 for details). Here, a significant difference suggests that cases with complete data are different than cases with incomplete data on a given variable indicating the variable should be included as an auxiliary variable. Lastly, Enders (2010), also suggests that a thorough literature review should provide the researcher with ideas for important auxiliary variables to include.

Some methodologists seem more comfortable providing rough guidelines for applied researchers. For instance, Graham (2009) noted that, “adding auxiliary variables that are correlated  $r = .50$  or better with the variables of interest will generally help the analysis to have less bias and more power,” (p. 565). This suggestion is supported by numerous simulation studies and has become a sort of standard recommendation (recall Collins, Shafer, & Kam’s similar recommendation). For instance, it is not uncommon for methodologists to define auxiliary variables as, “...variables that are highly correlated with the analysis model variables, but not necessarily part of the one’s analysis model,” (Graham, Cumsille, & Shevock, in press, p. 13). Regardless of the approach or number of auxiliary variables retained, the idea is to

acknowledge that the MAR assumption is not automatically met and a reasonable attempt should be made to approximate it.

*The appeal of a restrictive strategy.* Assuming that the cause of missing data can be captured by a few carefully selected variables, researchers can reach a point of “diminishing returns” where only a few auxiliary variables contain useful information (Graham, Cumsille, & Shevock, in press). In this situation the restrictive approach to auxiliary variables is preferable because the addition of extra “junk” variables serves only to complicate the model (see Collins, Shafer, & Kam, 2001).

Occasionally, modern missing data handling procedures fail to converge on a reliable, acceptable solution (see Enders, 2002; 2010). These models tend to experience convergence problems when there is insufficient data to estimate certain parameters (see Enders, 2010; Graham, Cumsille, & Shevock, in press), or when there is a complex missing data pattern (see Enders, 2010; Graham, 2009; Savalei & Bentler, 2009). Often these conditions arise when too many variables are included in the data augmentation algorithm of MI or the iterative optimization algorithm of FIML. For instance, Enders (2010) noted that when the number of variables is greater than the number of cases, “...the data contain linear dependencies that cause mathematical difficulties for regression-based imputation,” (p. 255). Asparouhov and Muthén (2010) provide a useful example. Assume that a dataset contains 50 variables and  $N = 1000$  participants. If the first variable of this dataset contains only 40 observations out of 1000 (i.e., 960 missing values), the imputation model is not identified because  $40 < 50$  and the distribution of parameters will fail to stabilize (i.e., not converge).

Non-convergence is not limited to situations when there are more cases than variables. Enders and Bandalos (2001) studied convergence failure in FIML across a range of samples sizes and missing data rates and found FIML to have a general (i.e., aggregated across all conditions) non-convergence rate of approximately 10% with more problems associated with higher rates of

missing data and smaller samples sizes. In effect, as researchers include variables in MI or FIML that contain high levels of missingness, the amount of missing data patterns can become cumbersome and missing information can become too large to reconstruct given the observed data (Enders, 2010; Enders & Bandalos, 2001; Savalei & Bentler, 2009). For instance, consider that regression-based imputation procedures require separate regression equations for each missing data pattern (Enders, 2010; Rubin, 1987).

Methodologists provide divergent guidelines regarding the best approach to “fix” convergence failure. For instance, Schafer (1997) and Enders (2010) suggest adding an informative ridge prior for MI. Enders described the use of a prior ridge as a technique that, “...adds a small number of imaginary data records from a hypothetical population where the variables are uncorrelated. These additional data points can stabilize estimation and eliminate convergence problems, but they do so at the cost of introducing a slight bias to the simulated parameter values (and thus the imputations),” (p. 256 – 257). Schafer provides a formula for determining the value of the ridge prior when, “...either the data are sparse (e.g.,  $n$  is substantially larger than  $p$ ), [where  $n$  is the number of cases and  $p$  is the number of variables] or because such strong relationships exist among the variables that certain linear combinations of the columns of  $Y$  (where  $Y$  is the complete data matrix comprising both observed and missing observations) exhibit little or no variability,” (p. 155 – 156).

Other recommendations in the literature that apply to both MI and FIML include increasing the number of iterations to an arbitrarily large value (e.g., Enders, 2010; Graham, Cumsille, & Shevock, in press) or decreasing the convergence criteria (e.g., Asparouhov & Muthén, 2010). Methodologists note that default convergence criteria and number of iterations differ depending on the statistical software program used (e.g., Graham, 2009) and some researchers consider any particular value to be relatively arbitrary (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

Irrespective of these recommended fixes for non-convergence, a result may not be produced or (perhaps less obvious for the researcher) an improper solution is generated. Consider that decreasing the convergence criteria, for example, may result in missing value estimates that are drawn from an instable posterior distribution (where parameter values change erratically across iterations). Consequently, the practical recommendation in these situations has always been to simply decreasing the number of auxiliary variables used (Enders, 2010). This point is furthered by considering that in practice data may be robust to minor violations of the MAR assumption (Collins, Schafer, & Kam, 2001). Also, methodologists often point out that a few carefully selected auxiliary variables are generally enough (e.g., Enders, 2010; Graham, Cumsille, & Shevock, in press). As Enders pointed out, "...my experience suggests that there is little benefit to using a large number of auxiliary variables," (p. 35). Thus the restrictive strategy can be seen as more practical and efficient than the inclusive strategy.

*Issues with the restrictive strategy.* Given that there is a reason for missing data, it is important to review the recommendations for selecting a restrictive set of auxiliary variables. To begin with, it is typically not possible to prove that a particular variable (or set of variables) is the sole cause of missingness. Consider that testing this assumption (that all variables responsible for missing data are included) would require the missing values to be isolated from the influence of all other possible causes (both observed and unobserved causes; see Enders, 2010). As researchers are typically unaware of the precise causes of missing data, methodologists describe this as "an ultimately untestable assumption," (Baraldi & Enders, 2010; Enders, 2006; Peugh & Enders, 2005; Schafer & Olsen, 1998). That is, researchers are unable to prove that any particular set of auxiliary variables adequately capture the cause of missingness. Thus, the assumption of a MAR process, which is required for unbiased FIML and MI estimation, becomes a task for the research analyst, whom must assume that all influential variables are included in the missing data handling procedure.

Recent research suggests that decisions regarding auxiliary variable selection are perhaps more complex than they might seem (e.g., Kreuter & Olson, 2011; Little & Vartivarian, 2005). The following discussion acknowledges that researchers do not always collect rich auxiliary variables; and selecting variables based only on the strength of association may overshadow important theoretical considerations relating to missingness. Further, it is reasonable to assume that researchers accepting the restrictive strategy with regard to auxiliary variables may not consider the ramifications related to rich auxiliary variables that are more likely of interest to their inclusive strategy counterparts.

According to Kreuter and Olson (2011), it seems likely that the choice of auxiliary variables for a particular study are often based on convenience (i.e., some extra correlated variables in the data set that could be used) or tradition (i.e., a set of auxiliary variables were used previously). Rarely is the choice of auxiliary variables justified by intentionally collecting variables that are theoretically important to the response variables as well as to other key variables that might also contain missingness (see Little & Vartivarian, 2005). That is, auxiliary variables that correlate highly with outcome variables (i.e., Y's) might not also relate highly to predictor variables (i.e., X's). Even in the case where there is strong theory and the researcher has a deep understanding of the processes that might cause missing data, it seems unlikely that such information would lead to the selection of a specific set of auxiliary variables that correlate highly across all key analysis variables (i.e., Y's and X's; Kreuter and Olson, 2011).

Traditionally, the methodological literature does not seem to provide clear guidance about theoretically relevant auxiliary variables or about situations where potential auxiliary variables relate highly to some but not all analysis variables. Perhaps missing data research in the social and behavioral sciences has not reached this level of sophistication. Still, if researchers acknowledge that the choice of rich auxiliary variables is important to the subsequent statistical

inferences, then choosing auxiliary variables based on convenience or tradition may not be the best approach.

A reasonable argument might even be made against many of the standard approaches for selecting auxiliary variables. For instance, consider that the multiple  $t$ -test approach for selecting auxiliary variables is based on mean differences and completely ignores covariance information (Enders, 2010) which may hinder the researcher's ability to effectively implement the restrictive auxiliary variable strategy. In addition to this and other noted statistical limitations for accurately choosing a specific set of auxiliary variables, the idea itself (that the research is able to find all possible causes of missing data) is theoretically problematic. It is clear that a restrictive strategy could lead researchers to face a dilemma regarding their results. The issue is that it is not clear if the results are due to the relationships being studied or to the fact that a certain set of auxiliary variables are chosen. Said differently, given the noted estimation improvements related to auxiliary variables, it is possible for two researchers working on the same statistical model and dataset to come to differing conclusions due to their particular use of auxiliary variables or for the same researcher to obtain differing results based on which variables were selected from a larger dataset. This situation led Enders (2010) to recommend that the researcher responsible for handling missing data be the same researcher responsible for analyzing the data.

Additionally, reducing the number of variables used in MI and FIML due to estimation failure is at odds with current recommendations in favor of an inclusive strategy with regard to auxiliary variables (i.e., the inclusion of all possible causes or correlates of missingness into the missing data estimation routine). Large data projects can present a particular challenge to software convergence because it is possible to have hundreds of potential auxiliary variables to inform the missing data estimation procedure. As Graham, et al. (in press) note, "...there is an upper limit on how many auxiliary variables can feasibly be included in the model, (p. 32). This

situation is compounded when the researcher incorporates non-linear information (e.g., powered terms and interactions), especially when the auxiliary variables also have missing data.

It is clear that auxiliary variables are important tools for theoretically approximating the MAR assumption and practically for improving parameter estimates by decreasing standard errors and increasing statistical power. Further, methodologists typically recommend an inclusive strategy concerning auxiliary variables and seem to suggest a restrictive strategy as a practical approach given current software limitations (Enders, 2010). As Collins, Schafer & Kam (2001) stated, “we wish to stress that this [the restricted use of auxiliary variables] has nothing to do with statistical theory,” (p. 349). They describe only benefits from including as many auxiliary variables as possible and note that the limitation is with software and the associated documentation.

The literature does not provide clear guidance for selecting an appropriate subset of auxiliary variables. Thus, the researcher is faced with a challenge regarding the known benefits of auxiliary variables and the practical limitations of current statistical software. Currently, methodologists recommend a solution that involves selecting a few good auxiliary variables that correlate highly with the analysis variables (e.g., Collins, Schafer, & Kam, 2001; Enders, 2010; Graham, 2009). However, simulation studies regarding auxiliary variables have typically not considered some of the complexities that may exist (e.g., non-linear information; see Enders & Gottschall, 2011). That is, selecting a few auxiliary variables that correlate most highly with the outcome measures may be effective in some situations, but is not likely the best approach in others. Therefore, it is reasonable to question the plausibility of MAR when a restrictive strategy is incorporated via MI or FIML. Specifically, the degree to which a set of auxiliary variables influence parameter estimates in relation to all possible auxiliary variables is unclear and ultimately depends on the cause of missingness, which is unknown.

*A practical all-inclusive auxiliary variable strategy.* While there will always be uncertainty regarding the cause of missing data, it is a theoretically more reasonable approximation of the MAR assumption to reveal and incorporate as many potential causes of missingness as possible (including non-linear information). Therefore, I propose a practical solution where auxiliary variables are viewed as a collection of potentially useful variance rather than as specific variables that correlate with the analysis variables.

The use of methods to consolidate large numbers of variables and incorporate non-linear information (such as interaction terms and power terms) have come of age elsewhere in the social and behavioral sciences, and need to be incorporated more routinely into modern missing data handling procedures. Consequently, the following approach focuses on a method of using principal component analysis (PCA) to obtain auxiliary variables to inform missing data handling procedures. Specifically, PCA is used to reduce the dimensionality of all possible auxiliary variables in a data set. A new smaller set of auxiliary variables are created (e.g., principal components) that contain all the useful information (both linear and non-linear) from the original dataset. These new uncorrelated principal component variables are then used as auxiliary variables in an “*all-inclusive*” approach to best inform missing data handling procedure.

The missing data literature has recognized that there will always be uncertainty regarding the cause of missing data; however, the literature’s practice of encouraging the restrictive strategy may not emphasize enough that the MAR assumption is not automatically met. In theory it is a more reasonable approximation of the MAR assumption (i.e., the research analyst must assume that all influential variables are included in the missing data handling procedure) to incorporate all potential causes of missingness by extracting variance information that would otherwise remain hidden within the data. These relationships can be important and may lead researchers to more informed inferences.

In order to present the PCA auxiliary variable approach, discussion now turns elsewhere in the social and behavioral sciences (beyond the typical missing data literature) to PCA a known method for extracting useful linear and non-linear variance information and condensing that information into a smaller subset of variables. To this end, PCA is introduced in the following section followed by an introductory to how variance can be effectively extracted from observed variables and used for other multivariate analytic techniques, like MI and FIML.

### **An Introduction to Principal Components Analysis**

In the social and behavioral sciences, researchers are frequently interested in obtaining information from participants across a large collection of variables. Fundamentally, Principal Components Analysis (PCA) is a multivariate data reduction technique that involves the linear transformation of a large set of variables into a new, more parsimonious set of variables (i.e., the principal components) that are uncorrelated and that contain most of the original information (Dunteman, 1989; Jackson, 1991; Joliffe, 2002). In this way, PCA is used to explain (through linear combinations) the variance-covariance structure of a set of variables, (Johnson & Wichern, 2002). While the covariance information informs the data reduction process (i.e., highly correlated variables can be combined because they share information), the method of PCA is based on variances (Joliffe, 2002; Preacher, MacCallum, 2003; Widaman, 2007). That is, PCA does not attempt to explain the covariance information; rather PCA depends directly on the variances. The central idea of PCA is to find (through some form of eigendecomposition) a set of  $k$  principal components (  $componet_1, componet_2, \dots, componet_k$  ) that contain as much information (i.e., variance) as possible from the original  $p$  variables (  $var_1, var_2, \dots, var_p$  ), where  $k < p$ . The  $k$  principal components contain most of the important information from the  $p$  original variables and can then replace them in further analyses (Johnson & Wichern, 2002; Joliffe, 2002).

Consequently, the original data set is reduced from  $n$  measurements on  $p$  variables to  $n$  measurements of  $k$  variables (see Figure 37).

Geometrically, PCA is carried out by projecting the original data on a smaller number of dimensions, which are specifically chosen to take advantage of the variance of the variables (Jackson, 1991; Wickens, 1995). A useful metaphor is provided by photography. Consider that a photograph is a 2-dimensional representation of a 3-dimensional object. The object in the picture, depending on the angle (i.e., rotation), is recognizable because it contains most of the original information. In other words, the third dimension is not necessary to identify the object in the picture, so it can be collapsed without losing important information. The process of projection will be illustrated mathematically in the forthcoming discussion.

An idea frequently advanced in favor of PCA is its potential to reveal variable relationships by extracting variance information that would otherwise remain hidden (see Johnson & Wichern, 2002). These relationships can be important and lead researchers to novel interpretations of the data (Johnson & Wichern, 2002). Although many applications of PCA have been developed (see Jackson, 1991; Joliffe, 2002), the common designs within the social sciences include: errantly using PCA as a variant of factor analysis (see Widaman, 2007) and using PCA as a preliminary step for other statistical techniques like multiple regression, and cluster analysis (Jackson, 1991; Johnson & Wichern, 2002; Joliffe, 2002).

### **The Historical Context of Principle Component Analysis**

In 1901, Karl Pearson introduced “the method of principal axes”, an idea that is regarded as the earliest version of principal components analysis (Cowles, 2001; Dunteman, 1989; Jackson, 1991; Joliffe, 2002). His article titled, “On Lines and Planes of Closest Fit to Systems of Points in Space,” illustrates a combination of quantitative methodology relating to statistical variance with the geometrical and graphical method of moments from mechanics (see Aldrich, 2007; Pearson, 1936; Porter, 2004). It provides a theoretical motivation and the statistical

groundwork for explaining the sample variance among a set of  $p$  variables and in doing so outlines the best fitting lines and planes (i.e., principal components) to a  $p$ -dimensional scatter plot. Pearson introduced an important statistical tool for data reduction, which was later described (and independently developed) by Hotelling (1933) as a “principal components analysis” (see Cowles, 2001).

Pearson’s work on “best fitting lines and planes” was a substantial achievement at the time, especially considering the context from which it developed in the early 20<sup>th</sup> century. The fundamental concepts leading to the development of principal components analysis can be traced as far back as the work of Leonhard Euler in the late-18<sup>th</sup> century. Euler introduced the idea of principal axis rotation and described it mathematically by the angles that the rotated axes make relative to the original coordinate position (see Anderson, 1830). Later, Joseph-Lewis Lagrange studied principal axis rotation, based on earlier work by Euler (see Anderson, 1830), and introduced orthogonal transformations (Hawkins, 1975). He also made the connection that a quadratic form can be reduced to its sum of square terms by solving the corresponding eigenvalue problem (e.g., using Lagrange multipliers). Specifically, the resulting eigenvectors are the direction cosines (i.e., angles) of the rotated axes relative to the original axis position in 3-dimensional space (Hawkins, 1974). This work was included in Lagrange’s influential 1788 publication, *Mécanique Analytique*, (analytical mechanics).

As Hawkins (1974) notes, Lagrange’s orthogonal transformation technique depends on “the reality of the eigenvalues,” (p. 564) and Lagrange only proved that eigenvalues were real in, “a case of the cubic” (i.e., 3-dimensional space). This limitation was overcome in the early 19<sup>th</sup> century when Augustin-Louis Cauchy used his theory of determinates to extend Lagrange’s principal axis theorem from three dimensions to  $p$ -dimensional space (Hawkins, 1974). That is, he proved that real eigenvalues could be derived from a symmetric matrix of any given size. Throughout the 19<sup>th</sup> century, other significant developments took place including further work on

eigenvalue decomposition (such as the development of singular value decomposition; see Jolliffe, 2002) and the publication in 1888 of Galton's, *Co-relations and Their Measurement, Chiefly from Anthropometric Data*, which introduced regression and correlation (see Cowles, 2001).

During the same time that Galton was describing correlation and regression, Karl Pearson was teaching mechanics to engineering students (among other courses) while working as a professor of applied mathematics at the University College London (see Aldrich, 2007; Porter, 2004). While referring to Pearson's use of "graphical statics" in these lectures, Aldrich describes that, "Pearson's earliest work on probability was not linked to his graphical projects; it was essentially philosophical but it was linked to mechanics because his philosophical projects came out of mechanics," (p. 5). In 1891, Pearson added courses on *the geometry of motion* and *the graphical representation of statistics* (Aldrich, 2007). Two years later, Pearson developed his "methods of moments" which borrowed heavily from mechanics, especially in relation to graphical statics (Aldrich, 2007).

By the beginning of the 20<sup>th</sup> century, Pearson was in a fitting position to develop his version of principal components analysis. In Pearson's 1901 paper, he began by describing that the best fitting regression line for a variable  $y$  given the value of another variable  $x$  is not the same as the best fitting regression line for  $x$  given a particular value of  $y$ . Essentially, a different line of best fit is obtained depending on which variable is designated as the dependent variable. Therefore, he defined a line (or plane) of best fit that minimizes the sum of squared errors from each point in a  $p$ -dimensional scatter plot to the best fitting line itself (i.e., the principal component). Intuitively, the line (or plane) of best fit for a set of  $p$  correlated variables transforms the data by rotating the original coordinate axes or vectors about their centroids (i.e., means). The axes are rotated so that they go through the clusters of points in a fashion that minimizes the sum of squared errors.

Consider a brief example to illustrate the concept. In this example, two random normal variables ( $x$  and  $y$ ) have been generated on 25 observations with a mean of zero and a correlation of .869. Geometrically, each  $p$  variable represents an axis in  $p$  dimensional space. Variables that share variance (i.e., are highly correlated) create tighter clusters of points than variables that are not related. These data are displayed in Table 7 and will be referenced throughout the discussion of principal components analysis for didactic purposes. Figure 38 illustrates a scatterplot of these data with a superimposed correlation ellipse and fitted lines.

Notice that in Figure 38 the regression of variable  $x$  on  $y$  and of variable  $y$  on  $x$  produces different regression lines. Also notice that the lines of best and worst fit for the two variables  $x$  and  $y$  are perpendicular but rotated from the original coordinate positions. As Pearson (1901) described, the best fitting line (or plane) for a series of points in  $p$ -dimensional space, "...coincides in direction with the maximum axis of the correlation ellipsoid, and the mean square residual," (see Pearson, p. 565).

Figure 39 provides a plot of the transformed data. That is, the principal axes have been rotated to minimize the sum of squared errors perpendicular to the line of best fit. The locations of the new axes define the new variables (i.e., the principal components) which are weighted according to the amount of total variance that they describe and are oriented in "...a direction of uncorrelated variation," (Pearson, 1901, p. 566).

Recall that the variables  $x$  and  $y$  are highly correlated (i.e.,  $r = .869$ ) so they share variation. The degree of overlap between these two variables directly relates to the amount of variance explained by the best and worst fitting line. Said differently, Figure 39 illustrates that the first principal component (i.e., the line of best fit) explains the greatest amount of variance while the second principal component (i.e., the line of worst fit) explains the residual variance not already explained. Consider that if the variables  $x$  and  $y$  were uncorrelated (i.e., a spherical cloud of points), the principal components would essentially explain the same amount of

variation and any transformation of the data would not result in a dimensional reduction of the data.

The next major advancement in the development of PCA came in 1933 with the publication of Harold Hotelling's paper, *Analysis of a Complex of Statistical Variables into Principal Components*, which described principal components analysis essentially as we know it today. Rather than approaching principal components from a geometric perspective, Hotelling derived principal components algebraically using Lagrange Multipliers to solve the *characteristic equation* (i.e., an eigenvalue and eigenvector problem; see Jackson, 1991). Cowles (2001) notes that there is no reason to think that Hotelling was influenced in any way by Pearson's work, which was published nearly 30 years prior.

My assumption is that Hotelling, like Pearson, was initially influenced by analytic geometry. I base this assumption on the fact that in addition to scalar algebra Hotelling also introduced a "curious" notational convention from tensor calculus (i.e., geometrical vector calculus; see Bock, 2007). As Bock notes, "surely, this is the only paper containing tensor notation in the entire psychological literature and perhaps the statistical literature," (p. 42). Hotelling also provides a brief geometric explanation of principal components (e.g., axis rotation) in his 1933 paper, which differs from Pearson's earlier description but illustrates that he was aware of a geometric interpretation of principal components. Regardless of the reason for Hotelling's unique notation or the initial inspiration for the method, he developed the idea of principal components analysis heavily from a factor analytic perspective (Bock, 2007; Cowles, 2001; Jolliffe, 2002).

Much of the confusion that currently exists in the social sciences between factor analysis and principal components analysis may derive from Hotelling's 1933 paper. In this paper Hotelling writes, "it is natural to ask whether some more fundamental set of independent variables exist, perhaps fewer in number than the  $x$ 's, which *determine* the values the  $x$ 's will

take,” (p. 417, emphasis is mine). Unfortunately, this language is misleading because principal components as Hotelling described them (and as they are currently understood) are mathematically determined by linear combinations of the  $x$ 's. That is, the  $x$ 's *cause* “the fundamental set of independent variables” rather than the other way around. Furthermore, the equations underlying these methods are quite different (see Joliffe, 2002; Widaman, 2007).

It is interesting that Hotelling does not explicitly distinguish his method from factor analysis. For instance, when naming his transformed variables Hotelling noted that the word “factor” was used in psychology to describe a similar procedure but the word “component” was most appropriate due to, “...conflicting word usage attaching to the word “factor” in mathematics” and “...in view of the prospect of application of these ideas outside of psychology,” (p. 417). In fact, as Joliffe (2002) notes, much of Hotelling's paper seems to deal with factor analysis rather than principal components analysis. Actually, Bock (2007) suggests that a misunderstanding of the fundamental differences between the method of factor analysis and that of principal components analysis (by Hotelling) created “antagonism” between Hotelling and L.L. Thurstone, a colleague of Hotelling and an essential figure in the development of factor analysis. For a more complete discussion of the relationship between factor analysis (FA) and principal component analysis, see Widaman (2007).

### **Defining Principal Components Analysis**

When a researcher collects data each variable usually contains a certain amount of variance (i.e., the distribution about the mean response). The sample covariance matrix (**S**) provides a representation of the covariation between each variable in the sample and can be represented as:

$$\mathbf{S} = \frac{1}{n-1} \sum_{p=1}^n (x_p - \bar{x})^2 = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})' (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}) = \begin{bmatrix} s_{11}^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22}^2 & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ s_{p1} & \cdots & \cdots & s_{pp}^2 \end{bmatrix} \quad (54)$$

where  $\mathbf{X}$  is a vector of order  $p$  containing scores on  $p$  variables, and  $\bar{\mathbf{x}}$  is a vector containing the means of the same  $p$  variables (Johnson & Wichern, 2002). Then, vector  $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$  represents the vector  $\mathbf{X}$  with sample means subtracted. A single number summary of this matrix can be provided by the *total sample variance* (i.e., the sum of all the  $p$  individual variances; see Johnson & Wichern, 2002) which can be written as:

$$\sum_{i=1}^p \text{var}(\mathbf{S}_p) = \text{tr}(\mathbf{S}) = s_{11} + s_{22} + \dots + s_{pp} \quad (55)$$

where  $\text{tr}(\mathbf{S})$  represents the sum of diagonal elements of  $\mathbf{S}$ . While the total sample variance describes the variability of the data without taking into account the covariance information, it does reflect the overall spread (i.e., variance) of the data (Johnson & Wichern, 2002; Smith, 2002). Principal components analysis is a multivariate technique with the fundamental goal of explaining the total sample variance of a set of variables (i.e., the variance accounted for; Jackson, 1991; Jolliffe, 2002). For  $p$  variables, a total of  $p$  principal components can be formed using PCA where the sum of the variances of  $p$  principal components is equal to the total sample variance:

$$\sum_{i=1}^p \text{var}(\mathbf{K}_p) = \text{tr}(\mathbf{S}) = s_{11} + s_{22} + \dots + s_{pp} = \sum_{i=1}^p \text{var}(\mathbf{S}_p) \quad (56)$$

where  $(\mathbf{K}_p)$  represents a principal component of order  $p$ . The goal of PCA is to use linear combinations to find a set of  $k$  principal components (where  $k < p$ ) that account for nearly all of the total sample variance (Jackson, 1991). This process of data reduction can be illustrated in Figure 40 where each element in the data matrix  $(\mathbf{X})$  is a score for an individual on a particular

variable. Note that in Figure 40  $x_{ip}$  is the score for the  $i^{\text{th}}$  individual on the  $p^{\text{th}}$  variable. Likewise, each element in the principal components matrix ( $\mathbf{K}$ ) is a score for an individual on a particular component. Therefore,  $k_{ip}$  is the score for the  $i^{\text{th}}$  individual on the  $k^{\text{th}}$  component. Another important aspect of Figure 40 is that the original data matrix is transformed into a matrix of principal components through the use of a covariance or correlation matrix. The covariance/correlation matrix is used to derive the fundamental mathematics behind principal components analysis. These fundamental aspects of PCA relate to the weighting of linear combinations of variables and will be mathematically illustrated in the following section.

In the social sciences, linear combinations are often used to form composite scores (Johnson & Wichern, 2002). For instance, a composite score might be formed as a linear combination (e.g., summation) of a set of items that measure a particular construct, which can be written as:

$$C_k = a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kp}x_p \quad (57)$$

where  $x_p$  represents each of the  $p$  items,  $a_{kp}$  represents a weight (e.g., 0, 1) associated with each of the  $k$  composite scores on each of the  $p$  items, and  $C_k$  indicates a composite score that is expected to contain all the important variance in the original  $p$  items.

Similarly, principal components are created from linear combinations of items that combine variance; however, the weights (e.g.,  $a_{pk}$ ) are derived mathematically to maximize the amount of variance that can be explained by each of the  $p$  composite scores (Dunteman, 1989). For example, letting  $\mathbf{K}$  represent a vector of the new composite variables (i.e., principal components) the linear combinations for PCA can be written as:  $\mathbf{K} = \mathbf{a}'\mathbf{X}$  or more descriptively as:

$$\begin{aligned}
K_1 &= \mathbf{a}'_1 \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\
K_2 &= \mathbf{a}'_2 \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\
&\vdots \\
K_p &= \mathbf{a}'_p \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p
\end{aligned} \tag{58}$$

where  $X_p$  represents each of the original  $p$  variables and  $a_{pp}$  represents a weight coefficient that maximizes the variance in the linear composite (see Jackson, 1991). Thus, this equation represents a set of linear equations, one for each component in  $\mathbf{K}$ . Each equation represents a composite variable in  $\mathbf{K}$  as a linear combination of the variables in  $\mathbf{X}$ . For instance, the variance of the first principal component ( $K_1$ ) is as large as possible given the restriction that the sum of the squared weight coefficients equal 1.0:

$$K_1 = a_{11}^2 + a_{12}^2 + \dots + a_{1p}^2 = 1 \tag{59}$$

This constraint (i.e., the “normalization” constraint) is imposed on the weight coefficients for a particular principal component (e.g.,  $K_1$ ) to prevent an arbitrary increase in the weights during the variance maximization process (via *Lagrange multipliers*), which would arbitrarily inflate the maximum variance of that principal component (Jackson, 1991; Jolliffe, 2002). Essentially, Lagrange multipliers locate the maximum (or minimum) value of an estimate that is subject to certain constraints (Hotelling, 1933; Jolliffe, 2002). A practical illustration for the way a Lagrange multiplier operates is provided by excise taxes. Excise taxes are special taxes on particular products and can be used to limit a particular behavior. For instance, a large excise tax on cigarettes is commonly used to limit the use of the product. Consider that this tax effectively constrains the maximum cigarettes consumed by a person. Likewise, Lagrange multipliers can be used to estimate a maximum value given a particular constraint. The constraint prevents the maximum value from unrestricted increases. Regarding PCA the constraint is that the sum of these squared weight coefficients are set to 1.0 (note that other restrictions are possible but typically not desirable; see Jackson, 1991). This allows the relative weight estimates to be

meaningful and reflect each variables contribution to the total variance of a particular principal component (e.g.,  $K_1$ ). The variance of the second principal component ( $K_2$ ) is as large as possible given the variance not already captured by  $K_1$  and is also subject to the constraint that the sum of its weights equal 1.0:

$$K_2 = a_{21}^2 + a_{22}^2 + \dots + a_{2p}^2 = 1 \quad (60)$$

Because the second principal component does not use variance that was already assumed by the first principal component, the components do not share variance, so the covariance (or correlation) among the first two principal components is zero:

$$Cov(K_1, K_2) = 0 \quad (61)$$

Likewise, the variance of the  $p^{\text{th}}$  principal component ( $K_p$ ) is as large as possible given the variance not already assumed by  $K_{p-1}$  and subject to the constraints that

$$a_{p1}^2 + a_{p2}^2 + \dots + a_{pp}^2 = 1 \quad (62)$$

Similarly, the covariance among any principal components (e.g.,  $i$  and  $j$ ) is zero:

$$Cov(K_i, K_j) = 0 \quad (63)$$

Further, the variance (i.e., var) of each successive component ( $K_p$ ) is smaller than the previous component such that:

$$\text{var}(K_1) \geq \text{var}(K_2) \geq \dots \geq \text{var}(K_p) \quad (64)$$

The key to deriving principal components algebraically is solving the *characteristic equation*, a fundamental outcome of matrix algebra where a square positive-definite matrix (e.g.,  $\mathbf{S}$ ) can be mathematically decomposed into a smaller set of foundational variance (i.e., eigenvalues and eigenvectors; Jackson, 1991; Jolliffe, 2002; Kolman, 1996). More specifically, the weight coefficients (i.e.,  $a_{pp}$ ) are derived from an *eigenvector* of matrix  $\mathbf{S}$  corresponding to the  $p^{\text{th}}$  largest *eigenvalue*. Eigenvectors are effectively the mechanics behind the estimation of

principal components (see Equation 11). The fundamental role of eigenvectors in relation to a covariance matrix is a key concept in PCA and is therefore demonstrated in the next section. The process of solving the characteristic equation to determine the weight coefficients (i.e., eigenvectors) used in PCA will be described mathematically in the following section. The following discussion presents the basic ideas behind finding principal components through the eigenvalues and eigenvectors and is intended for didactic purposes. In practice, statistical programs are used to calculate principal components via a number of possible algorithms. As Jolliffe (2002) noted, analysts interested in the last few principal components may find differences across algorithms and should consider the approach taken by the specific software program of interest. As the subsequent discussion of PCA in relation to auxiliary variables does not require the last few components, this topic is not addressed. Interested readers should see Jolliffe (2002, p. 408 – 414) for discussion.

**Eigenvectors.** A vector is a matrix with a single column (column vector) or a single row (row vector). A vector can be thought of as a line (with a direction) emanating from the origin and terminating at a point. For instance, take the column vector:

$$\mathbf{u}_1 = \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} \quad (65)$$

which is a trajectory in 2 dimensional space that represents an arrow pointing from the origin (0, 0) to a point (0.77, 0.64). The vector  $\mathbf{u}_1$  is illustrated in Figure 41.

An *eigenvector* is a specific type of vector. The word “eigen” is a German word meaning “proper” or “characteristic” (see Kolman, 1996). An eigenvector is a vector that has a “proper” (i.e., fundamental) relationship with a particular symmetric matrix which can be determined using matrix algebra (Johnson & Wichern, 2002; Kolman, 1996; Wickens, 1995). To illustrate, consider the sample covariance matrix ( $\mathbf{S}$ ) from the simulated data in Table 7:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix} = \begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} \quad (66)$$

There are  $p = 2$  variables in this covariance matrix. Matrix algebra provides a mathematical framework for manipulating matrices and vectors in a way that identifies eigenvectors (Abdi & Williams, 2010; Jackson, 1991; Smith, 2002). For instance, when  $\mathbf{S}$  is post multiplied by the vector  $\mathbf{u}_1$ , the result is another vector ( $\mathbf{u}_{1,1}$ ):

$$\begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} = \begin{bmatrix} 1.18 \\ 0.97 \end{bmatrix} \quad (67)$$

The “characteristic” of this new vector is that it is 1.53 times the original vector:

$$\begin{bmatrix} 1.18 \\ 0.97 \end{bmatrix} = [1.53] \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} \quad (68)$$

That is, the scalar multiplication of vector  $\mathbf{u}_1$  (i.e., the multiplication of a single number by a vector) results in changing the length of the vector but not the angle of the vector (see Johnson & Wichern, 2002; Kolman, 1996), which is illustrated in Figure 42 where the new vector  $\mathbf{u}_{1,1}$  is plotted as an extension (a lengthening) of the original vector  $\mathbf{u}_1$ . As noted previously, the difference in length between  $\mathbf{u}_1$  and  $\mathbf{u}_{1,1}$  is determined by a particular scalar that defines the relationship of the original vector and the covariance matrix. In other words, the covariance matrix maps  $\mathbf{u}_1$  onto a scalar (i.e., 1.53) multiple of itself, (Kolman, 1996). Therefore,  $\mathbf{u}_1$  is a “basis vector” or eigenvector of the square matrix  $\mathbf{S}$  (see Johnson & Wichern, 2002). The scalar (i.e., 1.53) is known as an eigenvalue. Regardless of the length of  $\mathbf{u}_1$  the associated eigenvalue defined by the covariance matrix is constant. That is, multiplying the vector  $\mathbf{u}_1$  by some value (e.g., 5) would produce:

$$\begin{aligned}
\begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} [5] &= \begin{bmatrix} 3.86 \\ 3.18 \end{bmatrix} \\
\begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} \begin{bmatrix} 3.86 \\ 3.18 \end{bmatrix} &= \begin{bmatrix} 5.92 \\ 4.87 \end{bmatrix} = \\
\begin{bmatrix} 5.92 \\ 4.87 \end{bmatrix} &= [1.53] \begin{bmatrix} 3.86 \\ 3.18 \end{bmatrix}
\end{aligned} \tag{69}$$

with a  $\mathbf{u}_1$  and  $\mathbf{u}_{1,1}$  vector that are 5 times longer than before but that have the same relative length of 1.53. The length of a vector does not change the basis for the relationship among the eigenvector and the covariance matrix.

A covariance matrix of order  $p$  has  $p$  eigenvectors which are all orthogonal to each other (i.e., Jolliffe, 2002; Kolman, 1996). Each eigenvector provides a new reference point (rather than the original coordinate axes) for interpreting data in  $p$  dimensional space (Johnson & Wichern, 2002). Consider the  $p = 2$  eigenvectors associated with the sample covariance matrix  $\mathbf{S}$  in the following illustration (see Figure 43) where  $\mathbf{u}_1$  represents the eigenvector for the first dimension and  $\mathbf{u}_2$  represents the eigenvector for the second dimension, which is:

$$\mathbf{u}_2 = \begin{bmatrix} -0.64 \\ 0.77 \end{bmatrix} \tag{70}$$

Notice that the vectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  appear to be orthogonal (i.e., perpendicular). The *inner product* of two vectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  is the sum of element-by-element multiplication (see Johnson & Wichern, 2002):

$$\mathbf{u}_1' \mathbf{u}_2 = u_{11}u_{21} + u_{12}u_{22} + \dots + u_{pp}u_{pp} \tag{71}$$

For instance given the vectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  the inner product can be calculated as:

$$\mathbf{u}_1 = \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix}, \mathbf{u}_2 = \begin{bmatrix} -0.64 \\ 0.77 \end{bmatrix}$$

$$\mathbf{u}_1' \mathbf{u}_2 = \begin{bmatrix} 0.77 & 0.64 \end{bmatrix} \begin{bmatrix} -0.64 \\ 0.77 \end{bmatrix} = \quad (72)$$

$$0.77 * -0.64 + 0.64 * 0.77 = 0$$

The inner product is used to compute the angle between vectors (Johnson & Wichern, 2002; Kolman, 1996). For instance, the angle between two vectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  is the  $\cos(\theta)$  where:

$$\theta = \frac{\mathbf{u}_1' \mathbf{u}_2}{\sqrt{\mathbf{u}_1' \mathbf{u}_1} * \sqrt{\mathbf{u}_2' \mathbf{u}_2}} \quad (73)$$

From the current example,

$$\cos(\theta) = \frac{\mathbf{u}_1' \mathbf{u}_2}{\sqrt{\mathbf{u}_1' \mathbf{u}_1} * \sqrt{\mathbf{u}_2' \mathbf{u}_2}} = \frac{0}{\sqrt{1} * \sqrt{1}} = 0 \quad (74)$$

$$\theta = \cos^{-1}(0) = 90^\circ$$

Thus it can be illustrated that the eigenvectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  are in fact orthogonal because their inner product is zero (i.e.,  $\mathbf{u}_1' \mathbf{u}_2 = 0$ ) which indicates the angle between these vectors is  $90^\circ$  as is illustrated in Figure 44.

While the vector  $\mathbf{u}_2$  in Figure 44 may seem to be longer than the vector  $\mathbf{u}_1$ , they are exactly the same length. The length of a particular vector emanating from the origin is given by the Pythagorean formula (see Johnson & Wichern, 2002). For instance letting  $L_u$  represent the length of a particular vector:

$$L_u = \sqrt{u_1^2 + u_2^2 + \dots + u_p^2} = \sqrt{\mathbf{u}' \mathbf{u}} \quad (75)$$

While the length of a particular eigenvector is not particularly relevant (as illustrated in Equation 22), the relative length of an eigenvector in relation to the length of another eigenvector is meaningful when the eigenvectors are used as reference points for interpreting data (Jackson,

1991). To standardize these reference points, the length of each  $p$  eigenvectors are set equal to one. That is, if  $\mathbf{u}'\mathbf{u} = 1$  (i.e., the sum of squares of elements of  $\mathbf{u}$  is equal to 1) then  $\mathbf{u}$  is said to be a standardized eigenvector of  $\mathbf{S}$  (see Jackson, 1991). Standardized eigenvectors may be obtained by a simple transformation of the eigenvector:

$$\frac{1}{\sqrt{\mathbf{u}'\mathbf{u}}}\mathbf{u} \quad (76)$$

For instance, the standardized length of the vector  $\mathbf{u}_1$  is:

$$\frac{1}{\sqrt{0.77^2 + 0.64^2}} \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} = 1.0 \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} \quad (77)$$

which, in the current example, is the same length as the original vector. Typically, this is not the case; however, the current data (see Table 7) were simulated to have a mean of 0 and a variance of 1.0; therefore, the eigenvectors  $\mathbf{u}_1$  and  $\mathbf{u}_2$  are effectively already standardized.

As Joliffe (2002) notes, occasionally researchers refer to the standardized eigenvectors as “principal components” even though, this practice is not technically accurate and leads to confusion. Rather, “it is preferable to reserve the term “principal components” for the derived variables  $\mathbf{a}_p'\mathbf{X}$ , and refer to  $\mathbf{a}_p$  as the vector of coefficients or loadings for the  $p$ th PC,” (p. 6).

This point is further developed in the following section.

***Finding eigenvectors.*** Assuming In the previous section, the eigenvectors of the example covariance matrix  $\mathbf{S}$  were provided as:

$$\mathbf{u} = [\mathbf{u}_1 \quad \mathbf{u}_2] = \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix} \quad (78)$$

In practice, matrix algebra is needed to uncover the eigenvectors associated with  $\mathbf{S}$ . The following section introduces eigendecomposition (reduction of the covariance matrix to its characteristic roots and vectors) by Lagrange multipliers (see Jackson, 1991) using the data in Table 7. Essentially, Lagrange multipliers locate the maximum (or minimum) of a linear function

that is subject to certain constraints (Hotelling, 1933; Joliffe, 2002). As will soon be illustrated, Lagrange multipliers are used to maximize  $\mathbf{a}'\mathbf{X}$  subject to the constraint that

$a_{p1}^2 + a_{p2}^2 + \dots + a_{pp}^2 = 1$ . While these illustrations involve only two variables these concepts generalize to  $p$  variables.

To begin, consider a sample covariance matrix ( $\mathbf{S}$ ) of order  $p$ . If  $\mathbf{u}$  is a column vector of order  $p$ , and  $\ell$  is a scalar such that  $\mathbf{S}\mathbf{u} = \ell\mathbf{u}$ , then  $\ell$  ( $\ell_1, \ell_2, \dots, \ell_p$ ) is said to be an eigenvalue (or characteristic root) of  $\mathbf{S}$  and  $\mathbf{u}$  ( $u_1, u_2, \dots, u_p$ ) is said to be the corresponding eigenvector (or characteristic vector) of  $\mathbf{S}$  (Johnson & Wichern, 2002; Jackson, 1991). This expression can be illustrated in expanded form as:

$$\begin{aligned} s_{11}u_1 + s_{12}u_2 + \dots + s_{1p}u_p &= \ell u_1 \\ s_{21}u_1 + s_{22}u_2 + \dots + s_{2p}u_p &= \ell u_2 \\ &\vdots \\ s_{p1}u_1 + s_{p2}u_2 + \dots + s_{pp}u_p &= \ell u_p \end{aligned} \quad (79)$$

Determining the eigenvalues and eigenvectors in this formula is known as solving a classical eigenproblem (see Joliffe, 2002; Voelz, 2006). The equation  $\mathbf{S}\mathbf{u} = \ell\mathbf{u}$  can be rearranged into a homogeneous form (i.e., setting the right side of the equation equal to zero) known as the characteristic equation, which can be written as (see Jackson, 1991):

$$|\mathbf{S} - \ell\mathbf{I}| = 0 \quad (80)$$

where  $\mathbf{I}$  is an identity matrix of order  $p$ . This equation derives  $p$  eigenvalues that can be ordered in sequence from largest to smallest (i.e.,  $\ell_1 \geq \ell_2 \geq \dots \geq \ell_p$ ). Consider the following illustration of the characteristic equation to solve for the eigenvalues of the example covariance matrix  $\mathbf{S}$  (see Jackson, 1991):

$$\begin{aligned}
\mathbf{S} &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\
\det(\mathbf{S} - \ell \mathbf{I}) &= \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} \ell & 0 \\ 0 & \ell \end{bmatrix} = \det \begin{bmatrix} a - \ell & b \\ c & d - \ell \end{bmatrix} \quad (81) \\
&= (a - \ell)(d - \ell) - bc \\
&= \ell^2 - (a + d)\ell + (ad - bc) = \frac{a + d}{2} \pm \frac{\sqrt{4bc + (a - d)^2}}{2}
\end{aligned}$$

Notice that the matrix algebra produces a quadratic equation that is solved by factoring. To further illustrate the derivation of eigenvalues, let the covariance matrix  $\mathbf{S}$  be defined by the simulated data in Table 7, the sample covariance matrix is:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix} = \begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} \quad (82)$$

There are  $p = 2$  variables in this covariance matrix. Applying  $\mathbf{S}$  to the characteristic equation produces:

$$\begin{aligned}
(\mathbf{S} - \ell \mathbf{I}) &= \left\{ \begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} - \begin{bmatrix} \ell & 0 \\ 0 & \ell \end{bmatrix} \right\} = \begin{bmatrix} 0.95 - \ell & 0.71 \\ 0.71 & 0.67 - \ell \end{bmatrix} \\
\det \begin{bmatrix} 0.95 - \ell & 0.71 \\ 0.71 & 0.67 - \ell \end{bmatrix} &= (0.95 - \ell)(0.67 - \ell) - 0.71 * 0.71 \\
&= \ell^2 - (0.95 + 0.67)\ell + (0.95 * 0.67 - 0.71 * 0.71) \quad (83) \\
&= \frac{0.95 + 0.67}{2} \pm \frac{\sqrt{4 * 0.71 * 0.71 + (0.95 - 0.67)^2}}{2} \\
&= 1.533, 0.086
\end{aligned}$$

where the eigenvalues are  $\ell_1 = 1.533$  and  $\ell_2 = 0.086$ . These eigenvalues represent the variance of the  $p = 2$  principal components with the first component ( $\ell_1$ ) explaining most of the variance.

Now that the eigenvalues are determined, the associated eigenvectors can be calculated by applying the sample covariance matrix  $\mathbf{S}$  to the following equations (see Jackson, 1991):

$$[\mathbf{S} - \ell \mathbf{I}] \mathbf{u} = 0 \quad (84)$$

and the normalizing equation:

$$\frac{1}{\sqrt{\mathbf{u}'\mathbf{u}}}\mathbf{u} \quad (85)$$

For the first eigenvalue ( $\ell_1 = 1.533$ ) the corresponding eigenvectors are:

$$\begin{aligned} [\mathbf{S} - \ell_1 \mathbf{I}]\mathbf{u} &= \begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} - \begin{bmatrix} 1.533 & 0 \\ 0 & 1.533 \end{bmatrix} \\ &= \begin{bmatrix} 0.95 - 1.533 & 0.71 \\ 0.71 & 0.67 - 1.533 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \quad (86)$$

To solve this equation Lagrange multipliers are used (Jackson, 1991; Joliffe, 2002). Specifically,

set  $u_{11} = 1$  and use the first equation:  $(0.95 - 1.533)(1) + (0.71)(u_{21}) = 0$  which is

$-0.58 / -0.71 = 0.82$ . Then the normalizing equation is used such that:

$$\frac{1}{\sqrt{\mathbf{u}'\mathbf{u}}}\mathbf{u} = \frac{1}{\sqrt{\begin{bmatrix} 1.0 & 0.82 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.82 \end{bmatrix}}} \begin{bmatrix} 1.0 \\ 0.82 \end{bmatrix} = \begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} \quad (87)$$

Notice, that the outcome vector in Equation 40 is the same eigenvector from previous examples.

In a similar manner, the second eigenvalue ( $\ell_1 = 0.086$ ) can be used to estimate its corresponding eigenvectors as:

$$\begin{aligned} [\mathbf{S} - \ell_1 \mathbf{I}]\mathbf{u} &= \begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} - \begin{bmatrix} 0.086 & 0 \\ 0 & 0.086 \end{bmatrix} \\ &= \begin{bmatrix} 0.95 - 0.086 & 0.71 \\ 0.71 & 0.67 - 0.086 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \quad (88)$$

To solve this equation set  $u_{21} = 1$  and use the first equation:  $(0.95 - 0.086)(u_{11}) + (0.71)(1.0) = 0$

which is  $0.71 / -0.86 = -0.82$ . Again the normalizing equation is used to determine the set of

eigenvalues:

$$\frac{1}{\sqrt{\mathbf{u}'\mathbf{u}}}\mathbf{u} = \frac{1}{\sqrt{\begin{bmatrix} 1.0 & -0.82 \end{bmatrix} \begin{bmatrix} 1.0 \\ -0.82 \end{bmatrix}}} \begin{bmatrix} 1.0 \\ -0.82 \end{bmatrix} = \begin{bmatrix} 0.77 \\ -0.64 \end{bmatrix} \quad (89)$$

The eigenvalues obtained in Equation 36 could be arranged in descending order as diagonal entries in a diagonal matrix ( $\mathbf{D}_\ell$ ) and the corresponding eigenvectors arranged as columns in  $\mathbf{U}$  such that:

$$\mathbf{D}_\ell = \begin{bmatrix} 1.533 & 0 \\ 0 & 0.086 \end{bmatrix}, \mathbf{U} = \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix} \quad (90)$$

The *eigenstructure* of  $\mathbf{S}$  can then be illustrated in matrix form as  $\mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{D}_\ell$  (see Jackson, 1991) such that:

$$\begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix} = \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix} \begin{bmatrix} 1.533 & 0 \\ 0 & 0.086 \end{bmatrix} \quad (91)$$

By post multiplying this matrix equation by  $\mathbf{U}'$  the matrix  $\mathbf{S}$  is isolated and it can be illustrated that the eigendecomposition of  $\mathbf{S}$  is:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\ell\mathbf{U}' \quad (92)$$

Therefore,

$$\begin{bmatrix} 0.95 & 0.71 \\ 0.71 & 0.67 \end{bmatrix} = \begin{bmatrix} 0.77 & -0.64 \\ 0.64 & 0.77 \end{bmatrix} \begin{bmatrix} 1.533 & 0 \\ 0 & 0.086 \end{bmatrix} \begin{bmatrix} 0.77 & 0.64 \\ -0.64 & 0.77 \end{bmatrix} \quad (93)$$

Note that if both sides of the equation were post multiplied by  $\mathbf{U}$ , we would have the eigenstructure again.

If all  $p$  eigenvalues of a symmetric matrix  $\mathbf{S}$  are positive, then  $\mathbf{S}$  is said to be positive definite (Johnson & Wichern, 2002). That is, the covariance matrix does not contain linear dependencies. The determinant,  $|\mathbf{S}|$ , of a symmetric matrix  $\mathbf{S}$  is equal to the product of its eigenvalues (Kolman, 1996). If any of the  $p$  eigenvalues of  $\mathbf{S}$  are zero, then  $\mathbf{S}$  is a singular matrix (Kolman, 1996). As previously noted, the trace (sum of diagonal elements) of a symmetric matrix  $\mathbf{S}$  is equal to the sum of its eigenvalues. Thus,  $\mathbf{S} = \mathbf{U}\mathbf{D}_\ell\mathbf{U}'$  implies that  $tr(\mathbf{S}) = tr(\mathbf{D}_\ell)$ . The sum of the eigenvalues is equal to the total sample variance:

$$\ell_1 + \ell_2 + \dots + \ell_p = \sum_{i=1}^p \ell_i = \text{tr}(\mathbf{S}) \quad (94)$$

The corresponding percentage of total sample variance accounted for by each eigenvalue,  $p$ , can be obtained from (see Johnson & Wichern, 2002):

$$\frac{\ell_p}{\ell_1 + \ell_2 + \dots + \ell_p} \quad (95)$$

Recall that the eigenvalue associated with the largest percentage of the total sample variance directly links to a set of eigenvectors. Geometrically, these vectors are the angles of the rotated axes relative to the original  $p$  coordinate positions of  $\mathbf{S}$  (Jackson, 1991; Jolliffe, 2002). Figure 45 illustrates this relationship using the data in Table 7. Given that the largest eigenvalue is 1.533 and is associated with the eigenvector:

$$\begin{bmatrix} 0.77 \\ 0.64 \end{bmatrix} \quad (96)$$

where the angles  $\theta_{11}$  and  $\theta_{21}$  are estimated as (i.e., Jackson, 1991; Johnson & Wichern, 2002; Kolman, 1996):

$$\theta_{11} = \cos(0.77) = 39.65^\circ, \quad \theta_{21} = \cos(0.64) = 50.21^\circ \quad (97)$$

and it can be shown that these angle add up to  $90^\circ$  (within rounding error). Likewise, the eigenvalue associated with the *next* largest percentage of the total sample variance is 0.086 and is associated with the eigenvector:

$$\begin{bmatrix} -0.64 \\ 0.77 \end{bmatrix} \quad (98)$$

where the angles  $\theta_{12}$  and  $\theta_{22}$  are estimated as:

$$\theta_{12} = \cos(0.77) = 39.65^\circ, \quad \theta_{22} = \cos(-0.64) = 129.80^\circ \quad (99)$$

and it can be shown that  $\theta_{22}$  minus  $\theta_{12}$  equals  $90^\circ$  (within rounding error). These angles are illustrated in Figure 46. Figures 45 and 46 are illustrated with a “line of best fit” and a “line of worst fit”, respectively. These lines represent the first (see Figure 45) and second (see Figure 46) principal component of the  $2 \times 2$  covariance matrix  $\mathbf{S}$ . The following section illustrates the simple mathematical computation of these new uncorrelated variables from the associated eigenvectors.

**Illustration of principal component calculation.** A For  $p$  variables, the  $p \times p$  covariance/correlation matrix has a set of  $p$  eigenvalues  $(\ell_1, \ell_2, \dots, \ell_p)$  and  $p$  eigenvectors  $(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p)$ . Each principal component is formed by taking the eigenvalues as the weights of the linear combination. These weights are defined by eigenvalues are multiplied by the original data matrix to construct principal component scores. The following matrix equation then represents the variables in  $\mathbf{k}$  as a linear function of the variables in  $\mathbf{x}$ , weighted by the corresponding eigenvectors  $\mathbf{u}'$  (Dunteman, 1989; Jackson, 1991; Jolliffe, 2002):

$$\mathbf{k} = \mathbf{u}'\mathbf{x} \quad (100)$$

which can be expanded to:

$$\begin{bmatrix} k_{11} & \cdots & k_{1k} \\ k_{21} & \cdots & k_{2k} \\ k_{31} & \cdots & k_{3k} \\ \cdots & \cdots & \cdots \\ k_{i1} & \cdots & k_{ik} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} & u_{31} & \cdots & u_{i1} \\ u_{12} & u_{22} & u_{32} & \cdots & u_{i2} \\ u_{13} & u_{23} & u_{33} & \cdots & u_{i3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_{1p} & u_{2p} & u_{3p} & \cdots & u_{ip} \end{bmatrix} * \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & x_{i3} & \cdots & x_{ip} \end{bmatrix} \quad (101)$$

Following the rules for multiplication, we can write this equation for any variable in  $\mathbf{k}$  as:

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{u}'_1 \mathbf{X}_1 = u_{11}x_{11} + u_{21}x_{21} + \dots + u_{p1}x_{p1} \\ \mathbf{k}_2 &= \mathbf{u}'_2 \mathbf{X}_2 = u_{12}x_{12} + u_{22}x_{22} + \dots + u_{p2}x_{p2} \\ &\vdots \\ \mathbf{k}_p &= \mathbf{u}'_p \mathbf{X}_p = u_{1p}x_{1p} + u_{2p}x_{2p} + \dots + u_{pp}x_{pp} \end{aligned} \quad (102)$$

Thus, the matrix equation above actually represents a set of linear equations, one for each variable in  $\mathbf{k}$ . Each such equation represents a variable in  $\mathbf{k}$  as a linear combination of the

variables in  $\mathbf{x}$ , multiplied by a weighting coefficient in  $\mathbf{u}$ . These equations define the principal component scores, which reflect the variation that each original variable  $x_{pp}$  contributes to a particular principal component  $k_p$ . The sum of all of the contributions (i.e., weights) of a particular variable (e.g.,  $x_1$ ) for  $p$  principal components is equal to 1.0 indicating that *all* of the variance in that variable is accounted for by the set of principal components (e.g., Dunteman, 1989). Said differently, each variable contributes a certain amount of variance to the definition of each principal component. The total variance of a particular principal component is equal to the eigenvalue associated with the eigenvectors that define that component. So, the new transformed variables are referred to as principal components while the values that do the transforming are referred to as the principal component scores (see Figure 47). For a complete guide to working with principal components in the social and behavioral sciences, see Raykov and Marcoulides (2008).

**Summary of principal component analysis.** A Principal Component Analysis (PCA) is a multivariate data reduction technique and is typically recommended as a precursor to other multivariate analytic methods. As previously discussed, PCA has a long history with a complementing geometric and algebraic interpretation. The ability to uncover variance that is otherwise hidden within a large number of variables is one of the strengths associated with PCA. Another asset of PCA is the capacity to retain variation from a larger set of correlated variables in a few new variables that are uncorrelated with each other. These new variables provide a more parsimonious description of the original data without losing useful variance (Jolliffe, 2002). The calculations behind PCA are closed form linear transformations that are directly related to eigenvectors (Jackson, 1991; Jolliffe, 2002) where the variance of each successive component is maximized. Consequently, PCA is *not* a form of factor analysis. While it is possible to interpret principal components, this practice is typically irrelevant in most applications of PCA (see

Widaman, 2007). Although the illustrations of PCA computations based on Table 7 were relatively straightforward (i.e.,  $p = 2$  dimensions), more realistic numbers of variables (e.g.,  $p = 40, 60$ , etc.) are difficult to transform by hand calculations and impossible to plot in  $p$  dimensional space. Still, the basic ideas presented generalize to any number of variables (see Jackson, 1991).

### **Method I: Simulation Studies**

After discussing the theoretical rationale for PCA as a method to extract useful variance from a large set of variables, and the practical estimation limitations of the inclusive auxiliary variable strategy, the following discussion more specifically address a method of using principal component analysis (PCA) to obtain auxiliary variables to better approximate an inclusive strategy in MI and FIML. The PCA auxiliary variable method ( $PCA_{AUX}$ ) is assessed using a series of Monte Carlo simulation studies as well as an in an empirical data example. The purpose of these simulation studies was to address the following questions: (1) Does the PCA auxiliary variable ( $PCA_{AUX}$ ) method converge in a situation where a typical inclusive strategy fails? (2) Is the inclusion of a smaller set of PCA auxiliary variables ( $PCA_{AUX}$ ) as beneficial as including all possible auxiliary variables (AUX; an inclusive strategy)? (3) does this relationship change as a function of sample size, missing data rate, or the magnitude of variable correlations? (4) does a non-linear MAR mechanism influence the performance of  $PCA_{AUX}$  relative to AUX? and (5) what is the relative performance of  $PCA_{AUX}$  and AUX in an empirical data example?

#### **Design and Procedure**

**Convergence.** As previously noted, research suggests that MI and FIML tend to experience convergence problems in situations with a complex missing data pattern and insufficient coverage for a particular parameter estimate (Enders, 2010; Graham, Cumsille, & Shevock, in press; Graham, 2009; Savalei & Bentler, 2009). Although these discussions are useful to highlight a known limitation in the field, they may not offer compelling evidence

regarding the practical performance of MI and FIML estimation routines across various software packages. In the following section, a small simulation study was used to demonstrate convergence issues in both MI and FIML even when reasonable resolutions are pursued across selected statistical software packages. In order to demonstrate the potential for PCA auxiliary variables ( $PCA_{AUX}$ ) to convergence in a situation where a large number of typical auxiliary variables ( $AUX$ ) fail, these simulations were repeated using the  $PCA_{AUX}$  variables.

**Population Model.** To demonstrate a situation where modern methods fail to converge, consider a research scenario where it is of interest to estimate the mean vector and covariance matrix between two variables,  $X$  and  $Y$ . Additionally, suppose an inclusive strategy where 48 additional auxiliary variables ( $AUX_1 - AUX_{48}$ ) are included. Let the associations across all variables be set to  $\rho = .30$  and the means and variances set at  $\mu = 0.0$  and  $\sigma^2 = 1.0$ .

**Data Generation.** *Mplus* 6.0 (Muthén & Muthén, 1998–2010) was used to generate the population model outlined previously using one multivariate normal data set of size  $N = 1,000$ . After data generation, each dataset was saved to a file and then imported into SAS 9.3 (SAS Institute Inc., 2011) for sample size selection and missing data generation. In line with Asparouhov and Muthén (2010), three different sample sizes including  $N = 50$ ,  $N = 100$ , and  $N = 500$  were considered. These values were intended to represent a practical set of possible sample sizes. In order to obtain a particular sample size, the SURVEYSELECT procedure in SAS was used to draw a random sample from the largest sample size investigated (i.e.,  $N = 1,000$ ). This method was used to minimize the influence of sampling variability in the simulated results. After sample size selection, SAS was used to impose MAR missing data rate of 30% on all variables except  $X$ . The MAR mechanism was generated by deleting values based on the lower distribution of  $X$ , which resulted in a general missing data pattern. Figure 47 presents a graphical depiction of the missing data pattern. While these simulation conditions were not intended to be comprehensive, it was of interest to provide a practical demonstration of convergence failure

given a representative missing data rate (e.g., 30%) and pattern (e.g., general pattern) for a moderate to large number of variables (e.g.,  $p = 50$ ).

After generating the data and implementing the specified missing data rate, the data were read into a selected set of popular software programs including: (a) SAS 9.3, (b) *Mplus* 6.12, and (c) SPSS 20. For this demonstration three sequential imputation methods were used including (1) the first method was based on default settings in each program, (2) the next method increased the iterations to 10,000, (3) the following methods either added a ridge prior equal to 10 (in SAS) or increased the MCMC iterations (*Mplus*). Similarly, three sequential likelihood-based methods were used including: (1) default settings, (2) increasing the iterations to 10,000, and (3) lowering the convergence criteria to .01. These conditions were based on recommendations in the literature (see Enders, 2010; Graham, 2009; Savalei & Bentler, 2009; Schafer, 1997). For example, Schafer (1997) suggested that a ridge prior be used when, "...either the data are sparse, or because such strong relationships exist among the variables that certain linear combinations of the columns of  $Y$  exhibit little or no variability," (p. 155-156). The three sequential imputation approaches and the likelihood-based approaches were intended to show that convergence failure may occur even when reasonable steps are taken to generate a solution.

***Analysis Models.*** While MI was employed in all three programs, FIML was only available in *Mplus*. Therefore, MI represents the only method that was assessed across each of the selected software packages. Importantly, this data example was not intended to compare the various software packages as they utilize differing algorithms for MI. Rather, the goal was to generate data that produced estimation problems to highlight the idea that MI and FIML may fail to converge on an acceptable solution. For each condition, the missing data model incorporated  $X$ ,  $Y$  and all auxiliary variables ( $p = 48$ ). In order to complicate the model, two additional interaction terms were generated representing variables with non-linear variation that add complexity by infusing high collinearity into the estimation process. The first interaction term

was created by multiplying the first and the second auxiliary variables together (i.e.,  $AUX_{12}$ ) while the second interaction term was generated by multiplying the third and the fourth auxiliary variables together (i.e.,  $AUX_{34}$ ). These new variables were included among the original 48 auxiliary variables in the missing data model (i.e., 50 total auxiliary variables) and along with the two analysis variables (i.e.,  $X$  and  $Y$ ).

As previously mentioned, statistical theory and simulation based research supports an “inclusive strategy” where as many auxiliary variables are included as possible (e.g., Collins, Schafer, & Kam, 2001; Enders, 2010). However, the complexity of the missing data handling procedure typically increases with the addition of each additional auxiliary variable, especially when the auxiliary variables have missing data. Practically, there is a limit to the number of auxiliary variables that can be incorporated before the missing data handling routine (e.g., FIML, MI) will fail to converge on a solution (e.g., Enders, 2010; Graham, Cumsille, & Shevock, in press). The determination of this limit may relate to many dataset specific factors and is likely to vary across a specific set of variables. For instance, the amount of missing information, the number of variables in the model (i.e., complexity of the model) and the presence of high collinearity have been shown to influence the convergence of FIML and MI and may lead to estimation failure (e.g., Enders, 2010). Large data sets can present an obvious challenge, as it is possible to have hundreds of potential auxiliary variables with various rates of missing data and a greater potential for multicollinearity problems. Therefore, the goal of this small simulation was simply to demonstrate that the  $PCA_{AUX}$  method converges in situations where a typical inclusive strategy fails to converge. An investigation into the quality of a  $PCA_{AUX}$  relative to a typical inclusive strategy (AUX) is presented next.

**Outcomes.** In addition to assessing convergence based on software error messages and graphical displays (discussed previously), the mean squared error, a measure of the quality of missing data recovery, was calculated based on the formula (see Asparouhov & Muthén, 2010):

$$MSE = \sqrt{\frac{1}{p} \sum_{j=1}^p (\bar{\mu}_j - \mu_j)^2} \quad (103)$$

where  $p$  is the number of variables,  $\mu_j$  represents the true population value and  $\bar{\mu}_j$  is the average mean over  $Y_j$  imputed data sets. In the case of FIML,  $\mu_j$  represents the true population value as before and  $\bar{\mu}_j$  is the estimated variable mean for the  $j^{th}$  variable.

**Relative Performance.** The relative performance of AUX and PCA<sub>AUX</sub> was examined across seven independent variables (magnitude of correlations, homogeneity of correlations across auxiliary variables, rate of missing, missing data mechanism, missing data patterns, number of auxiliary variables, and sample size) on four dependent variables (raw parameter estimate bias, percent bias, standardizes bias, & relative efficiency). The manipulated independent variables represent factors known to influence the performance of auxiliary variables.

For example, the effectiveness of auxiliary variables is expected to improve as the magnitude of correlations increase given a particular rate of missing and sample size (Collins, Schafer, & Kam, 2001). As Enders (2010) noted, "...omitting a correlate of missingness was not that detrimental when the correlation was weak (e.g.,  $r \leq .30$ ), but the estimates became increasingly biased as the correlation between the auxiliary variable and the incomplete analysis variable increased in magnitude," (p. 129). This relationship explains the suggestion provided by many methodologists (e.g., Collins, et al., 2001) to locate auxiliary variables that exceed an association of about  $r = .40$  with the variables with missing data. This idea is so prevalent in the literature that it is not uncommon for methodologists to define auxiliary variables as, "...variables that are highly correlated with the analysis model variables, but not necessarily part of the one's analysis model," (Graham, Cumsille, & Shevock, 2013, p. 13).

Additionally, numerous Monte Carlo simulation studies on the topic of missing data assess the influence of sample size and rate of missing data on parameter bias, relative

efficiency, and convergence failure (e.g., Bodner, 2008; Enders & Bandalos, 2001; Enders & Peugh, 2004; Graham, Olchowski, & Gilreath, 2007; Wothke & Arbuckle, 1996; Yuan, 2009). For instance, Collins, Schafer and Kam (2001) noted that when the MAR assumption is violated the resulting estimation may not be substantially biased unless the rate of missing data is high. Bias has been shown to increase with the missing data rate (Enders & Bandalos, 2001).

Some research suggests that sample size may have little effect on relative efficiency and bias (Enders & Bandalos, 2001); though bias expressed relative to the standard error may be more sensitive (Enders, 2003). Further, small sample sizes are also associated with lower statistical power which may make bias difficult to detect even when the rate of missing is relatively high (Yuan, 2009). Research also notes that small sample sizes may increase estimation failure (Enders & Bandalos, 2001; Yoo, 2009) especially with high missing data rates (Newman, 2003).

The choice to examine two missing data mechanisms was influenced by Collins, Schafer, and Kam (2001), who noted that non-linear causes of missingness may generate bias even with a missing data rate less than 25%. They also discussed the potential to minimize this bias by including the non-linear information in the missing data handling procedure. Auxiliary variables that include variance from a non-linear MAR process are expected to improve estimation across various sample size and missing data rates.

For instance, the performance of both  $PCA_{AUX}$  and  $AUX$  are expected to be similar as  $PCA_{AUX}$  should contain the important information from  $AUX$ . That is, the bias associated with  $AUX$  should be similar to the bias observed with  $PCA_{AUX}$  even though the latter contains far fewer variables (e.g., 1  $PCA_{AUX}$  vs. 8  $AUX$ ). This relative relationship should improve in parallel as the magnitude of correlations (among the auxiliary variables and the variable with missing data) increase, and with sample size increases given a particular rate of missing and number of auxiliary variables. Both techniques are expected to outperform the absence of any auxiliary

variables. Additionally,  $PCA_{AUX}$  is expected to improve parameter estimation beyond the improvement of AUX given a non-linear MAR type of missingness. This situation mimics a scenario where the cause of missingness is an unknown non-linear process that is only captured via the *all-inclusive*  $PCA_{AUX}$  approach.

In order to specify the reasonable values for each condition, Monte Carlo simulation studies published in the social sciences on the topic of missing data handling between 1996 and 2011 were reviewed (see Table 8 for a summary). Based on this literature review, the following levels were selected: (a) magnitude of correlations ( $\rho = 0$  to  $.80$ ), (b) homogeneity of correlations across auxiliary variables (uniform, moderate homogeneity, low homogeneity), (c) rate of missing (10% to 80%), (d) missing data mechanism (MAR, non-linear MAR), (e) missing data patterns (monotone, and general), (f) number of auxiliary variables (0 to 8), and (g) sample size ( $N = 50$  to 1000). These levels represent a wide range of conditions found in the reviewed literature and were selected to provide a useful context in which to assess the relative performance of  $PCA_{AUX}$  and AUX.

***Population Model.*** For the first study, a bivariate normal correlation model between the variables X and Y were used to generate population data for the associated simulation studies. Figure 48 illustrates this population model including the 8 additional auxiliary variables that were also incorporated. This data generation model was an expanded version of the model used by Collins, Schafer, and Kam (2001) and Enders (2010). As Demonstrated in Figure 48, within a given cell design, the population correlation between X and Y was fixed at  $\rho = .30$  as was the relationship between the auxiliary variables and X. To manipulate the homogeneity of associations among the auxiliary variables, three correlation patterns were introduced. The population model for the uniform condition had identical correlations of  $\rho = .30$  for all eight auxiliary variables. In the moderate condition, one fourth of the correlations were set at  $\rho = .20$ ; another fourth was set at  $\rho = .25$ ; a third fourth was set at  $\rho = .35$ ; and the last fourth was set to  $\rho$

= .40. In the low homogeneity condition one fourth of the correlations were set at  $\rho = .10$ ; another fourth was set at  $\rho = .20$ ; a third fourth was set at  $\rho = .40$ ; and the last fourth was set to  $\rho = .50$ . The moderate and low homogeneity conditions generated model-implied correlations that varied in magnitude among each fourth of the specified correlations. Regardless of the auxiliary variable magnitude, the average association between the auxiliary variables and  $Y$  was about  $\rho = .30$  which is considered a conservative lower bound for exerting influence on the imputation model (see Collins, Schafer, & Kam, 2001; Graham, 2009). As illustrated in Figure 48, the magnitude of correlations ranged from  $\rho = 0$  to  $.80$  rather than from  $\rho = 0$  to  $1.0$  based on the limitations of Equation 20. That is, because some associations were set to  $\rho = .30$  the range of other relationships were constrained so that the population variance and covariance matrix was positive definite.

**Data Generation.** Data generation involved two separate sets of simulation studies including simulations using FIML and simulations using MI. In the following discussion, each approach will be discussed in turn. For the first study, *Mplus* 6.0 (Muthén & Muthén, 1998–2010) was used to generate ten variables:  $X$ ,  $Y$ , and eight auxiliary variables in 1,000 multivariate normal data sets of size 1,000 for each set of population values using the MONTECARLO command. These simulation studies were *external* Monte Carlo simulations (i.e., the data are *not* generated and analyzed in one step within *Mplus* (see Muthén & Muthén, 2010). Rather the simulated data from all replications were saved to a file. This was done primarily for two reasons: (1) The *Mplus* statistical program has a built in tool that generates missing values from a logistic regression model which methodologists note is difficult to accurately control (see Enders, 2009; Muthén & Muthén, 2010); therefore, the SAS 9.3 statistical program was chosen to implement missing data and (2) *Mplus* does not currently support principal component analysis, which was relevant to the current study.

To obtain a particular sample size, the SURVEYSELECT procedure in SAS 9.3 (SAS Institute Inc., 2011) was used to draw a random sample from the largest sample size investigated (i.e.,  $N = 1,000$ ). This method was used to minimize the influence of sampling variability in the simulated results. After data generation and sample size selection, SAS 9.3 was used to generate principal component scores using the PROC PRINCOMP procedure. Note that the variance used to generate the principal components did not include the variables  $X$  and  $Y$  as this would not provide a low bound for the effectiveness of the imputation model. That is, it could be argued that the relative performance of  $PCA_{AUX}$  is driven by the inclusion of variance from  $X$  and  $Y$ .

Then, SAS was used to impose one of the proposed missing data mechanisms (MCAR, MAR, non-linear-MAR; Little & Rubin, 2002) on  $Y$  with one of two missing data patterns. The first pattern was a monotone missing data pattern (i.e., where missingness occurs for some of the values of  $Y$ ; Enders, 2010; Little & Rubin, 2002) and the second pattern was a general missing data pattern (i.e., with missingness scattered throughout the data in a seemingly random fashion; see Figure 49). There are numerous research situations that lead to the monotone pattern of missing data and it has been widely used in the literature to study missing data (e.g., Little and Rubin, 1987; Enders, 2010). For example, let  $Y$  represent a questionnaire item with missing data and let  $X$  represent a demographic variable, like age, with no missing values. The variable  $X$  could also represent a fixed variable controlled by the experimenter or any other variable with complete data. While a monotone pattern is of primary interest for the majority of the current simulation studies, the general missing data pattern will provide useful information regarding convergence rates for various observed data coverage rates and presents a more appropriate missing data situation (Enders, 2010).

For each missing data mechanism, missing data were imposed by generating a random variable and then distributing it through FORMAT statement using SAS 9.3. This procedure allows for the specification of a particular missing data pattern and the random audition of values

of selected variables (e.g.,  $Y$ ) for deletion in order to obtain a particular missing data percentage. To select a variable (or set of variables) for deletion, another DATA statement was used to generate an array that referenced the variable of interest. Then, a uniform random number vector was used to pair the probability of missing data with the specified missing data pattern.

Under the MCAR mechanism, a data point was subsequently deleted if the corresponding uniform random number auditioned a value from the “misspattern” FORMAT statement that related to the specified missing data rate. In this way the probability that  $Y$ , for instance, contains missingness can be written as:  $p(Y_{mis} | \text{data}) = p(Y_{mis})$  where the propensity for missingness is unrelated to any measured or unmeasured variables (a random process). A key point of this study was to assess the power of  $\text{PCA}_{\text{AUX}}$  relative to all possible typical auxiliary variables (AUX;  $\text{Aux}_1 - \text{Aux}_8$ ). MCAR missingness is unbiased and is useful to assess statistical power associated with the inclusion of an auxiliary variable (or set of auxiliary variables) relative to a model where auxiliary variables are omitted. That is, including auxiliary variables ( $\text{PCA}_{\text{AUX}}$  or AUX) into a MCAR situation will not impact parameter estimate bias but should influence statistical power.

Another main point of this study was to assess the relative performance of  $\text{PCA}_{\text{AUX}}$  and AUX across a set of conditions. Therefore, another set of simulations required a MAR missing data mechanism. The MAR mechanism required a slightly different step than the MCAR mechanism previously described. Specifically, the deciles of a selected variable (e.g., the first auxiliary variable;  $\text{Aux}_1$ ) were found using the SAS UNIVARIATE procedure. These deciles were then be used to divide the distribution of  $\text{Aux}_1$  into 10 equal portions. This was done so that the  $k^{\text{th}}$  decile for  $\text{Aux}_1$  is a value (e.g.,  $p$ ) such that the probability that  $\text{Aux}_1$  will be less than  $p$  is at most  $k / p$  and the probability that  $\text{Aux}_1$  will be more than  $p$  is at most  $(p - k) / p$ . Thus, a data point was deleted if  $\text{Aux}_1$  was less than or equal to a particular decile. In this manner the probability that  $Y$  contains missingness can be written as  $P(Y_{mis} | \text{Data}) = P(Y_{mis} | q_i, \phi_{qu_i})$ . That is, the probability of missingness ( $Y_{mis}$ ) is related to the  $i^{\text{th}}$  decile of  $\text{Aux}_1$ . Among other things, this

procedure mimics a scenario in which missing values on a variable (i.e.,  $Y$ ) result from low scores on another variable in the data set. This MAR process is similar to that modeled by Muthén et al. (1987) and Enders and Bandalos (2001).

Also of interest in the current study was the influence of a non-linear MAR mechanism on the performance of  $PCA_{AUX}$  relative to  $AUX$ . This scenario required further simulations where missingness was produced in a similar manner to that described previously for the MAR condition. However, this time rather than utilizing the decile of a particular auxiliary variable (e.g.,  $Aux_1$ ), the decile of an interaction term was used (e.g.,  $Aux_1 * Aux_2$ ). A data point was deleted if  $Aux_1 * Aux_2$  was greater than or equal to a particular decile.

For the MI approach, the *simsem* package in the R program (Pornprasertmanit, Miller, & Schoemann, 2012) was used to generate ten variables:  $X$ ,  $Y$ , and eight auxiliary variables in 1,000 multivariate normal data sets of size 1,000 for each set of population values, as was specified for the FIML condition. The MI method utilized the same simulation designs (i.e., missing data mechanism, principal components analysis, etc.) were carried out within the R program.

***Analysis Models.*** The analytic model for these simulations is the association between  $Y$  and  $X$ . The variables  $Aux_1 - Aux_8$  were used as auxiliary variables. For the FIML approach, after generating the principal component scores, the data from all replications were saved to a file. Then, *Mplus* 6.0 was used to analyze the data for each sample replicate. Auxiliary variables were incorporated into the FIML estimation routine via the *AUXILIARY* (m) option (via the saturate correlates model; see Graham, 2003; Asparouhov & Muthén, 2008). Then, the results of each simulation (i.e., parameter estimates and standard errors) were saved to a data file and collected using a series of SAS macros. For the MI approach, the *simsem* package in R was used to analyze the data for each sample replicate. Auxiliary variables were incorporated into the MI estimation routine using *AMELIA* with 20 imputations (Honaker, King, & Blackwell, 2011).

Then, the results of each simulation were pooled using Rubin's Rules (Enders, 2010; Rubin, 1987).

***Outcomes.***

*Parameter Bias.* The main outcome of interest in the current study was parameter bias. Therefore, the average parameter estimate from each simulation condition was compared to its associated population parameter (the true value in the population) resulting in an estimate of raw bias. More specifically, four dependent variables related to bias were examined including (1) raw parameter estimate bias, (2) percent bias, (3) standardize bias, and (4) relative efficiency. Each of these variables has been previously assessed in Monte Carlo simulation work on missing data (see Table 8). Parameter estimate bias can be examined by comparing the average estimates from each cell design to the corresponding population parameters. More specifically, raw bias can be calculated as the average estimate minus the parameter population value. In order to relate these findings to prior research, raw bias is expressed as a percentage relative to the true parameter value (see Enders & Bandalos, 2001) which can be written as:

$$\% BIAS = \left[ \frac{\hat{\theta}_{ij} - \theta_i}{\theta_i} \right] * 100 \quad (104)$$

where  $\theta_i$  is the true population parameter for the correlation between  $X$  and  $Y$  (i.e.,  $\rho = .30$ ) and  $\hat{\theta}_{ij}$  is the corresponding parameter estimate from the  $j$ th iteration. In the current study the average percent bias across all 1000 replications was taken within a particular cell design.

Additionally, to aid in determining practical significance, standardized bias can be determined by dividing the raw bias estimate by the standard deviation of each estimate within a particular set of conditions (see Collins, Schafer, & Kam, 2001; Enders & Gottschall, 2011) which can be expressed as:

$$\text{Standardized } BIAS = 100 * \left[ \frac{\hat{\theta}_{ij} - \theta_i}{SE_{ij}} \right] \quad (105)$$

where  $SE_{ij}$  is the standard error from the  $j^{th}$  iteration. Collins, Schafer and Kam (2001) report that standardized bias estimates that exceeds 40% - 50%, “begin to have noticeable adverse impacts on efficiency, convergence, and error rates,” (p. 340). They suggest a standardized bias absolute value greater than 40% as practically significant.

*Relative Efficiency.* Another prevalent standard metric reported below is the Root-mean-squared error (RMSE), the average squared difference between the estimate  $\hat{\theta}_{ij}$  and the population value  $\theta_i$  (Collins, Schafer, & Kam, 2001; Enders, 2003; MacDonald, Thurston & Nelson, 2000; Roth, Switzer & Switzer, 1999; Zhang & Walker, 2008). This metric combines efficiency and bias information and is typically reported in relation to accuracy (Collins, Schafer, & Kam, 2001). A purer measure of efficiency is given by the ratio of sampling variances for the parameter of interest (i.e.,  $r_{xy}$ ) called relative efficiency (RE; Arbuckle, 1996; Enders & Bandalos, 2001; Wothke, 2000) which can be illustrated by:

$$RE = \frac{\sigma^2 q_i}{\sigma^2 AUX_{PCA}} \quad (106)$$

where  $\sigma^2 q_i$  is the sampling variance of the parameter of interest ( $r_{xy}$ ) from the  $i$ th auxiliary variable and  $\sigma^2 AUX_{PCA}$  is the corresponding sampling variance of the principal component auxiliary variable method. The average standard error estimates under each condition can be compared to the true standard error (i.e., based on the Monte Carlo simulation standard deviation across replications) in a similar fashion as the parameter bias to determine the degree of over or under estimation (see Newman, 2003).

## **Method II: Empirical Study**

In addition to the simulation work, we investigated the relative performance of typical auxiliary variables (AUX) and a smaller set of PCA auxiliary variables (PCA<sub>AUX</sub>) in an empirical data example. Using actual data we investigated three conditions: (a) no auxiliary variables were employed, (b) all possible AUX were used, and (c) the PCA<sub>AUX</sub> approach was used. This example uses data from a previously reported sample assessing the Early Communication Indicator (ECI), a measure for intervention decision-making in infants and toddlers (see Greenwood & Walker, 2010). Participants were infants and toddlers recruited from Early Head Start (EHS) programs and included children between the ages of 8 and 37 months and the percent missing was 57%.

### **Procedure and variables**

Researchers collected key communicative behaviors including: gestures, vocalizations, single words, and multiple words in 1-month intervals representing 33 separate occasions between the ages of 5 to 37 months (see Greenwood & Walker, 2010 for details). The data were then transformed to reflect mean estimates separated by three months which generated 11 measurement occasions between ages of 5 to 37 months. For the current example, the first measurement occasion (including months 5, 6 and 7) for each key skill variable was not used because single and multiple word usage was not observed during this developmental period resulting in 40 key skill variables (10 measurement occasions x 4 variables; see Table 9). Raw data plots are shown in Figure 55.

**Empirical missing data.** All 40 variables had at least one missing value on a case, 569 of the 570 cases had at least one missing value on a variable, and that 13,003 of the 22,800 values (cases × variables) are missing; therefore, 57% of the data were missing. There were 308 missing data patterns dispersed throughout the data in a seemingly random fashion (see Enders, 2010); however, as demonstrated in Table 9, much of the missing data occurred at the 10<sup>th</sup> measurement occasion (i.e., 35 – 37 months).

*Mplus* 6.0 was used to calculate means and variances for each of the 40 variables via FIML using the AUXILIARY (m) option. Additionally, the SAS 9.2 MI procedure (PROC MI) with the Markov Chain Monte Carlo (MCMC) estimation routine was used to create 100 imputations. Regarding the relative performance of the PCA<sub>AUX</sub> and AUX method, the results from the FIML and MI approaches were essentially the same; therefore, only the MI results are presented to take advantage of the fraction of missing information and relative increase in variance diagnostic measures that are specific to MI. The multiple imputation procedure incorporated: an inclusive auxiliary variable strategy with 40 analysis variables (ECI key skill elements) and 706 auxiliary variables, and an all-inclusive auxiliary variable strategy utilizing a set of PCA auxiliary.

## Outcomes

**Descriptives.** Bias cannot be assessed in an empirical example because we do not know the true population values. Thus, we focus our investigation on efficiency: more efficient estimates have smaller standard errors and confidence intervals, increasing power to test hypotheses. Therefore, it was of interest to review the standard errors of the ECI measures (i.e., gestures, vocalizations, single words, and multiple words) across AUX and PCA<sub>AUX</sub> conditions.

**Fraction of missing information.** An outcome of interest in the empirical study was the fraction of missing information (FMI). Following Enders (2010), when the number of imputations (denoted  $m$ ) is large (e.g.,  $\geq 100$ ), FMI can be expressed as:

$$FMI = \frac{V_B + (V_B / m)}{V_T} \quad (107)$$

where  $V_B$  denotes the between-imputation variance, which is calculated by taking the sum of the squared deviations divided by the number of imputations, and  $V_T$  represents the total variance, which is the sum of  $V_B$  and the average estimated sampling variance (squared standard error;

Enders, 2010) across  $m$  imputations. FMI reflects the relationship between the missing data pattern and the magnitude of associations among variables where highly related variables with good coverage produce the lowest FMI value. That is, missing data does not greatly influence the sampling variance. Conversely, weakly related variables with poor coverage generate a high FMI. In this situation the missing data may significantly inflate the sampling variance. The interpretation of FMI is straightforward. For instance, when  $FMI = .004$  less than 1% (e.g., 0.4%) of the sample variance is related to missing data.

**Relative increase in variance.** Another important outcome was the relative increase in variance (RIV). While the FMI provides the percentage of sample variance related to missing data, the RIV yields a measure of the relative increase in variance (between-imputation variance relative to within-imputation variance) such that  $RIV = 0$  indicates that the variance was not influenced by missing data. Conversely,  $RIV = 1$  suggests that the between-imputation variance is equivalent to the within-imputation variance (see Enders, 2010). As illustrated by Enders, RIV can be written as:

$$RIV = \frac{V_B + (V_B / m)}{V_W} = \frac{FMI}{1 - FMI} \quad (108)$$

where  $V_B$ ,  $V_T$ , and  $m$  are defined as previously indicated while  $V_W$  is the sampling error that would have resulted given no missing data.

### Simulation Results

For ease of presentation, a select set of conditions are presented to highlight the relative performance of AUX and  $PCA_{AUX}$  in relation to parameter estimate bias.

#### Convergence

**AUX strategy.** FIML in *Mplus* was assessed for each sample size by sequentially implementing methods 1 – 3 (as previously described) until convergence was reached. Given the smallest sample size  $N = 50$  none of the likelihood-based models (i.e., 1 – 3 defined above)

converged. Note that this model was not identified, as there were more variables than cases. Further, this model had less information to inform the likelihood-based estimation than models with more cases. That is, 30% missing data from  $N = 50$  provides less coverage than 30% missing from  $N = 100$  or  $N = 500$ . These points are reflected in the error code generated by *Mplus* that noted,

*“THE MISSING DATA EM ALGORITHM FOR THE HI MODEL HAS NOT CONVERGED WITH RESPECT TO THE PARAMETER ESTIMATES. THIS MAY BE DUE TO SPARSE DATA LEADING TO A SINGULAR COVARIANCE MATRIX ESTIMATE. INCREASE THE NUMBER OF HI ITERATIONS. NOTE THAT THE NUMBER OF HI PARAMETERS (MEANS, VARIANCES, AND COVARIANCES) IS GREATER THAN THE NUMBER OF OBSERVATIONS. NO CONVERGENCE. SERIOUS PROBLEMS IN ITERATIONS.”*

With a sample size of  $N = 100$ , FIML failed to converge using the default settings and an error message noted no convergence and the presence of a singular covariance matrix. However, FIML was able to converge when the number of iterations were increased resulting in  $MSE = 0.4221$ . Further, given the largest sample size  $N = 500$  the default settings converged with  $MSE = 0.3981$ . While this demonstration of convergence failure is not intended to be inclusive, it does indicate that failure may occur with only 30% missing data given a large enough set of auxiliary variables. As shown above a larger sample size was associated with a lower number of iterations needed to reach convergence. Additionally, the model conditions including  $N = 500$  and the default estimation model provided a lower MSE estimate indicating a more accurate imputation procedure than the conditions including  $N = 100$  and the estimation model with increased iterations.

MI within the SAS statistical program was assessed for each sample size by sequentially implementing methods 1 – 3 (as described above) until convergence was reached. As with the FIML method, the smallest sample size  $N = 50$  did not converge across any of the imputation

models. The error code generated by SAS included, “WARNING: The EM algorithm (MLE) fails to converge after 200 iterations. You can increase the number of iterations (MAXITER= option) or increase the value of the convergence criterion (CONVERGE= option). WARNING: The EM algorithm (posterior mode) fails to converge after 200 iterations. You can increase the number of iterations (MAXITER= option) or increase the value of the convergence criterion (CONVERGE= option). WARNING: The initial covariance matrix for MCMC is singular. You can use a PRIOR = option to stabilize the inference. WARNING: the posterior covariance matrix is singular. Imputed values for some variables may be fixed.” While the sample size of  $N = 100$  demonstrated convergence failure under the default settings, and marked instability under increased iterations, convergence was reached via a prior ridge which resulted in  $MSE = 0.4326$ . The largest sample size  $N = 500$  also revealed marked instability until a prior ridge was added resulting in  $MSE = 0.4009$ . Given that the  $N = 100$  and the  $N = 500$  sample sizes both needed a prior ridge to successfully converge, the larger sample size provided a lower MSE estimate indicating a more accurate imputation procedure (see Asparouhov & Muthén, 2010). Additionally, the convergence speed of the imputation model based on  $N = 500$  was much faster than the imputation model based on  $N = 100$ . Specifically, the sample size of  $N = 500$  converged within 116 MCMC iterations while the sample size of  $N = 100$  resulted in convergence within 342 MCMC iterations.

MI within the *Mplus* statistical program was assessed for each sample size by sequentially implementing methods 1 – 3 (as described above) until convergence was reached. As with the prior models, none of the imputation models properly converged with the smallest sample size of  $N = 50$ . *Mplus* generated the following error code,

*“PROBLEM OCCURRED DURING THE DATA IMPUTATION. YOU MAY BE ABLE TO RESOLVE THIS PROBLEM BY SPECIFYING THE USEVARIABLES OPTION TO REDUCE*

*THE NUMBER OF VARIABLES USED IN THE IMPUTATION MODEL. SPECIFYING A DIFFERENT IMPUTATION MODEL MAY ALSO RESOLVE THE PROBLEM.”*

Given the sample size of  $N = 100$  MI in *Mplus* converged after increasing the MCMC iterations to 10,000 resulting in a  $MSE = 0.4228$ . Moreover, the sample size of  $N = 500$  converged by increasing the iterations where the  $MSE = 0.3983$ .

MI within the SPSS statistical program was assessed sequentially as previously discussed in relation to the SAS program. However, none of the SPSS imputation models (e.g., Imputation Model 1 – 3) converged for any of the specified sample sizes. The error message generated by SPSS for each condition simply stated, “Not Imputed (Too Many Missing Values)”.

**PCA<sub>AUX</sub> Strategy.** Regarding the PCA<sub>AUX</sub> method, principal component scores were generated from the  $p = 50$  auxiliary variables, which resulted in 10 PCA variables that accounted for approximately 80% of the variance in the auxiliary variables. Each of the FIML and MI models were run where 10 PCA<sub>AUX</sub> variables were used in place of the 50 AUX variables. The purpose of this approach was to determine whether the PCA<sub>AUX</sub> method converged in conditions where a typical inclusive auxiliary variable strategy failed.

FIML in *Mplus* was assessed as before for each sample sizes by sequentially implementing methods 1 – 3 (i.e., default settings, increased iterations, decreased convergence criteria) until convergence was reached. Given the smallest sample size of  $N = 50$  the default model converged under the PCA<sub>AUX</sub> approach resulting in  $MSE = 0.0240$ . Note that this model was identified due to the use of PCA<sub>AUX</sub> variables, as there were more cases (i.e., 50) than variables (i.e., 12). With a sample size of  $N = 100$ , FIML was able to converge under the default settings resulting in  $MSE = .0225$ . Further, given the largest sample size  $N = 500$  the default settings converged with  $MSE = .0125$ . While this demonstration is not comprehensive, it does suggest that to the use of PCA<sub>AUX</sub> variables (a smaller uncorrelated subset of the original

variables) converged faster and related to a much smaller MSE estimate than incorporating all possible auxiliary variables.

MI within the SAS statistical program was assessed for each sample size by sequentially implementing methods 1 – 3 (i.e., default settings, increase iterations, utilize a prior ridge) until convergence was reached. As with the FIML method, the smallest sample size  $N = 50$  converged within 31 MCMC iterations using the default settings which resulted in  $MSE = .0387$ . Further, the sample size  $N = 100$  converged within 27 MCMC iterations using the default settings which resulted in  $MSE = .0350$ . The sample size  $N = 500$  converged within 25 MCMC iterations using the default settings which resulted in  $MSE = .0182$ . As demonstrated, the convergence speed of the  $PCA_{AUX}$  models was fast and increased with sample size. Additionally, larger sample sizes provided a lower MSE estimate signifying more accuracy in the imputation process (Asparouhov & Muthén, 2010).

MI within *Mplus* was assessed for each sample size by sequentially implementing methods 1 – 3 (as described previously) until convergence was reached. As with the prior models, smallest sample size of  $N = 50$  converged with  $MSE = .0243$ . Given the sample size of  $N = 100$  MI in *Mplus* converged with  $MSE = .0159$ . Moreover, the sample size of  $N = 500$  converged where the  $MSE = .0112$ .

MI within the SPSS statistical program was assessed size by sequentially implementing methods 1 – 3 (as described previously) until convergence was reached. The  $N = 50$  sample size converged under the default settings with  $MSE = .0394$ . Similarly, the  $N = 100$  sample size converged under the default settings with  $MSE = .0239$  and the largest sample size  $N = 500$  converged using the default settings resulting in  $MSE = .0133$ . The quality of the information retained in the  $PCA_{AUX}$  approach is addressed next.

### **Linear MAR performance**

A key point of this study was to determine if the inclusion of a PCA auxiliary variable

(PCA<sub>AUX</sub>) is as beneficial as an inclusive strategy using typical auxiliary variables (AUX), given that the PCA<sub>AUX</sub> approach represents a reduced set of variables that are less likely to cause convergence failure. The following discussion highlights results from the condition with a 60% MAR missing data rate and a monotone missing data pattern where the population correlation between the auxiliary variables (Aux<sub>1</sub> – Aux<sub>8</sub>) and  $Y$  was fixed at  $\rho = .60$  in the population model. The presented relationships hold across sample size and missing data rates though large sample sizes decrease the observed bias (illustrated in Table 10) and larger missing percentages result in more bias.

Further, regarding the relative performance of the PCA<sub>AUX</sub> and AUX method, the results from the FIML and MI studies were essentially the same; therefore, only the FIML results are presented. Also, while the correlation magnitude between the auxiliary variables and the variable with missingness influenced parameter estimates, this study found no differences in the homogeneity of auxiliary variable correlations. Consequently, the subsequent presentation is limited to the uniform model.

Table 10 contains results showing the impact of PCA<sub>AUX</sub> and AUX on parameter estimation raw bias, standardized bias, and percent bias across various sample sizes. The AUX and the PCA<sub>AUX</sub> approach demonstrate essentially no bias across various sample sizes, though the PCA<sub>AUX</sub> method performs slightly better with smaller sample size relative to the AUX method. For instance, the AUX method that included all possible auxiliary variables (Aux<sub>1</sub> – Aux<sub>8</sub>) produced a slight bias of 1% or less with more bias associated with lower sample sizes. In comparison, the PCA<sub>AUX</sub> approach that utilized a single auxiliary variable (which accounted for approximately 80% of variance among the 8 simulated auxiliary variables) produced a bias of 0.16% or less with more bias associated with lower sample sizes.

Also of interest is the “no auxiliary variable” condition which is used as a point of comparison between PCA<sub>AUX</sub> and AUX beyond relative contrasts. The most notable bias is

found when no auxiliary variables are used and the sample size was small (see Table 10). This result is not surprising because excluding correlates of missingness conclusively shows parameter estimate bias and is not in line with current missing data theory regarding the use of auxiliary variables (see Collins et al., 2001; Enders, 2010). Said differently, when the auxiliary variables are not included the MAR mechanism effectively becomes a MNAR mechanism and the parameter bias is maximized for each condition.

### **Non-linear MAR performance**

Another goal of this study was to assess the relative performance of  $PCA_{AUX}$  and AUX when the cause of missingness is nonlinear. As illustrated in Table 11, the AUX method that did not include the  $AUX_1 * AUX_2$  interaction term (i.e., the cause of missingness) produced a standardized bias of up to .10 (i.e., 1/10 of a standard error) with an associated 22% bias. In comparison, the  $PCA_{AUX}$  approach utilized a single principal component auxiliary variable that included the nonlinear information from the  $AUX_1 * AUX_2$  interaction term produced a standardized bias of no more than .005 with the largest percent bias reaching 1.7%. As seen in Table 11, the magnitude of bias across both methods decreased with sample size.

Table 12 presents the relative performance of  $PCA_{AUX}$  and AUX where the AUX variable list incorporated the additional non-linear cause of missingness (i.e.,  $AUX_1 * AUX_2$ ). This condition is perhaps a more reasonable comparison among AUX and  $PCA_{AUX}$  because both procedures reflect the MAR mechanism. The addition of the interaction term decreased bias. For example, the sample size of  $N = 200$  in Table 11 included an 73.5% bias when no auxiliary variables were included, an 8% bias when all possible auxiliary variables were included excluding the interaction term, and less than 1% bias when all possible auxiliary variables including the interaction term was included or when the one PCA auxiliary variable was included. This finding was consistent across sample sizes except under small sample size conditions where the  $PCA_{AUX}$  approach shows less bias than the AUX approach with the

interaction term.

**Relative efficiency.** Following Enders and Bandalos (2001), relative sampling variance was used to determine the relative efficiency (RE) of parameter estimates based on the  $PCA_{AUX}$  relative to the AUX method where RE indicates the sample size increase need by the AUX method to reach the same efficiency as the  $PCA_{AUX}$  method (e.g.,  $RE = AUX / PCA_{AUX}$ ; see Table 13). For instance, a RE estimate of 1.100 indicates that the sample size for the AUX (i.e.,  $AUX_1 - AUX_8$ ) would need to increase by 10% to have the same efficiency as  $PCA_{AUX}$ .

While the sampling variability was smaller with larger sample sizes, RE was not influenced by sample size. As expected, RE estimates were influenced by the MCAR missing data rate where more missingness resulted in less efficiency. Table 13 also shows that across all conditions the  $PCA_{AUX}$  method was at least as efficient as and typically more efficient than the AUX method. The RE estimates were most pronounced at the  $\rho = .60$  association level among the auxiliary variables and the variable with missingness (i.e.,  $\rho_{Y,AUX}$ ) and at the 60% missing data rate. For instance, given the  $N = 1,000$  condition and the 60% missingness rate the AUX method would require approximately a 26% (i.e., 1.255) larger sample size relative to the  $PCA_{AUX}$  method (see Table 13). By comparison, the RE estimates were lowest when  $\rho = .10$  among the auxiliary variables and the variable with missingness and at the 10% missing data rate. In this situation, the AUX approach demonstrated the same level of precision reached by the  $PCA_{AUX}$  method.

### **Empirical Example Results**

After imputing all 100 data sets, a table with means, standard errors, fraction of missing information (FMI; see Bodner, 2008; Schafer & Olsen, 1998) and relative increase in variance (RIV; Enders & Bandalos, 2001; Graham et al., 2007) was created for each condition (see Table 14). Here, the FMI ranged from 0.43 to 0.80 and the RIV ranged from 0.74 to 5.01. Conversely, in Table 15 (which contains the results using PCAs), the FMI improved, ranging from less than 0.08 to 0.25, and the RIV improved, ranging from 0.11 to 0.34. Additionally, the standard errors

in Table 15 are much lower than the standard errors in Table 14. For example, the standard error of gestures at 9 months was .40 in Table 15 and was .21 in Table 16.

### **Discussion**

It follows from missing data theory that the inclusive auxiliary variable strategy cannot be ruled out yet there are important practical limitations to its implementation. Given convergence failure despite attempts to fix estimation errors, methodologists (e.g., Enders, 2010) typically recommend reducing the number of variables used in MI and FIML. However, this recommendation opposes current recommendations in favor of an inclusive strategy with regard to auxiliary variables. The methodological literature does not seem to provide clear guidance on how researchers should proceed when they are confronted with a large number of auxiliary variables; especially if they are also interested in incorporating non-linear information to better satisfy the MAR assumption. Currently, many methodologists seem to theoretically recommend an inclusive strategy but also note that a few highly correlated auxiliary variables are probably good enough and frequently end up recommending the more practical restricted auxiliary variable approach (see Collins, Schafer, & Kam, 2001; Enders, 2010; Graham, 2009; Graham, Cumsille, & Shevock, in press). Many of the recommendations seem to be based on limited empirical evidence in order to provide general guidance for applied researchers. While the basic idea for selecting a few good auxiliary variables may seem practical, some important limitations were discussed that should be given more attention, including more general knowledge of the role and behavior of auxiliary variables in FIML and MI (e.g., Kreuter and Olson, 2011).

Further, if one argues that any particular set of auxiliary variables are sufficient, then it follows that any other data analyst could locate yet another set of auxiliary variables and provide a rational justification for their choice. That is, it is reasonable to assume that different research questions would drive the selection of specific types of auxiliary variables. In this context, a

concern arises in the replication of research findings as the MAR assumption may be satisfied to varying degrees within the same data set depending on the auxiliary variables utilized for a particular analysis. In contrast, a true *all*-inclusive strategy (i.e., all possible linear and non-linear auxiliary variables are included) would provide the best approximation of the MAR assumption given the entirety of the observed data.

For the purpose of discussion, consider conducting multiple imputation on a large public survey dataset as originally proposed by Rubin (1972). Assume that various researchers will utilize this dataset to address many different hypotheses. Given that the analyst conducting MI must ensure that the imputed data is informed by the same level of specificity as the various analysis models, it is reasonable that an *all*-inclusive strategy be implemented. This idea and the current estimation limitations related to the incorporation of a large set of auxiliary variables has led some methodologists to generally recommend that the person conducting MI be the same person running the analysis model (e.g., Enders, 2010). Enders and Gottschall (2011) reiterated this idea noting, “special care must be exercised when using multiple imputation with multiple group models, as failing to preserve the interactive effects during the imputation phase can produce biased parameter estimates in the subsequent analysis phase, even when the data are missing completely at random or missing at random,” (p. 35).

The situation previously discussed presents a data analyst with a predicament as reducing the number of auxiliary variables used may yield biased estimates; yet, including too many auxiliary variables is likely to generate estimation problems and lead to the non-identification condition (i.e., more variables than cases) discussed previously. Further, while Enders (2009) suggested that it is beneficial to including auxiliary variables with missing data, consider the additional complication related to patterns of missingness and coverage when numerous auxiliary variables with missingness are incorporated into the model.

While there will always be uncertainty regarding the cause of missing data, it is a more theoretically reasonable approximation of the MAR assumption to include all potential causes of missingness by extracting variance information (linear and non-linear) that would otherwise remain hidden within the data. These relationships can be important and may lead researchers to more informed conclusions. As discussed, the history of research on missing data consistently provides a context where observed values provide indirect information about the likely values of missing data.

The prior discussion provided a historical account of research and theory on related to missing data discussed the context and rationale for including auxiliary variables in FIML and MI estimation procedures. Much of this work emphasized the importance of assumptions regarding why the data are missing as this can bias any inferences made from the data being studied. Given our current state of knowledge, a practical solution was proposed. As auxiliary variables are not of substantive interest, they are essentially analytic tools used to improve data recovery. As such, the researcher should not be concerned with the specific auxiliary variables selected. Instead, the research should consider that it is the variance information (linear and non-linear) that is needed. Methods to use PCA to consolidate large numbers of variables and incorporate non-linear information (such as interaction terms and power terms) were demonstrated in modern missing data handling procedures. More specifically, PCA was used to reduce the dimensionality of the auxiliary variables in a data set. A new smaller set of auxiliary variables are created (e.g., principal components) that contain all the useful information (both linear and non-linear) in the original data set were then used as auxiliary variables to inform the missing data handling procedure (i.e., FIML and MI)..

### **Convergence**

As suggested in the literature convergence failure was observed in the presence of a large number of auxiliary variables (i.e., an inclusive strategy). The smallest sample size ( $N = 50$ )

consistently resulted in convergence failures at a modest missing data rate of 30% and was likely motivated by non-identification (i.e., more variables than cases) related to the small sample size and high number of variables in the AUX condition. For example, while the sample size was 50 cases the number of variables was 52. Asparouhov and Muthén (2010) discussed convergence failures in this context and recommend removing variables until the model is properly identified. The literature suggests that an identified model will not contain inherent, "...linear dependencies that cause mathematical difficulties," (Enders, 2010, p. 255). When the PCA<sub>AUX</sub> condition was employed for the  $N = 50$  sample size, all models converged and this was the case across all software programs. This observed difference between AUX and PCA<sub>AUX</sub> was primarily due to the reduction in the number of auxiliary variables from 52 (the AUX method) to 10 (the PCA<sub>AUX</sub> method). Another explanation that likely influenced convergence across the other sample sizes was that the AUX approach included variables with high collinearity (by including AUX<sub>1</sub>, AUX<sub>2</sub> and AUX<sub>1</sub>\*AUX<sub>2</sub>) while the PCA<sub>AUX</sub> approach contained 10 variables that were uncorrelated to each other by definition (see PCA section above). This point highlights an important potential advantage of the PCA<sub>AUX</sub> method where it is possible to extract useful variance without the complication of a non-positive definite covariance matrix caused by multicollinearity.

The fact that the MSE estimates consistently decreased as sample size increased was not surprising given Equation 116 and prior research that suggests the quality of missing data recovery depends on available information (e.g., Asparouhov & Muthén, 2010; Collins, Schafer, & Kam, 2001; Enders, 2010). Asparouhov & Muthén reported relatively similar MSE estimates with sample sizes of  $N = 75, 100, 200,$  and  $1000$ . MSE estimates were consistently lower for the PCA<sub>AUX</sub> approach relative to the AUX method in this demonstration across sample sizes of  $N = 100,$  and  $500$  (note that  $N = 50$  only converged with the PCA<sub>AUX</sub> approach). These differences were also consistent across software packages. A possible explanation for this difference is that the PCA<sub>AUX</sub> method converged based on the default settings across all software packages, while

the AUX approach relied on the recommended fixes for non-convergence suggested by the literature (e.g., prior ridge, increased iterations, etc.). In summary, as was demonstrated across software programs the PCA<sub>AUX</sub> method converged in conditions where a typical inclusive auxiliary variable strategy failed. That is, as the PCA<sub>AUX</sub> approach used far fewer auxiliary variables and convergence was easier to obtain.

### **Parameter Bias**

The observed bias across conditions was consistent with prior research and missing data theory concerning auxiliary variables (e.g., Collins et al., 2001; Enders, 2010). Consistent with Collins et al. (2001), higher associations among the auxiliary variables and the analysis variables related to less bias under the MAR mechanism. As pointed out previously, this is perhaps the most common method for choosing a restrictive set of key auxiliary variables. In addition, higher sample sizes were associated with less bias and the inclusion of more auxiliary variables generally saw less bias than including fewer auxiliary variables. This finding supports the recommendations for using an inclusive strategy with auxiliary variables. That is, currently there are no known drawbacks for including as many auxiliary variables as possible other than increasing model complexity (Collins et al., 2001).

This study did not find differences across these sample sizes related to a uniform, low, or high homogeneity condition. It is possible that no difference was found because the specified ranges were not extreme enough to influence the estimation process or because each homogeneity condition averaged  $\rho = .30$ . More research is needed to address this topic.

While these findings suggest that the PCA<sub>AUX</sub> method may outperform the AUX approach, the standardized bias values did not exceed .01 for most conditions suggesting that the observed bias is not practically significant (see Collins et al., 2001; Enders & Gottschall, 2011). The fact that these techniques show comparable performance suggests that the PCA<sub>AUX</sub> approach extracted relevant variance information from the eight simulated auxiliary variables (AUX<sub>1</sub>–

AUX<sub>8</sub>). That is, using the PCA method to generate a set of auxiliary variables seems to be an effective approach for reducing the number of auxiliary variables without increasing bias.

Regarding the non-linear MAR condition, findings suggest that the PCA<sub>AUX</sub> approach is able to capture non-linear causes of missingness and thus produce more accurate parameter estimates relative to the typical auxiliary variable approach where the non-linear information is not included. These findings were expected because the PCA<sub>AUX</sub> approach contained variance from the interaction term (i.e., AUX<sub>1</sub>\*AUX<sub>2</sub>) and the linear components (AUX<sub>1</sub> and AUX<sub>2</sub>) unlike the AUX method that contained only the linear components. When the interaction term, which generated missing data, was included into the AUX approach bias was essentially eliminated from all but the smallest sample size of  $N = 50$ . These findings match the performance of the PCA<sub>AUX</sub> approach except that the sample size of  $N = 50$  was also unbiased. As before the PCA method appears to be an effective approach for reducing the number of auxiliary variables without increasing bias. In practice, the PCA<sub>AUX</sub> method is capable of capturing MAR related variance information that applied researchers may not have knowledge about. As the cause of missingness is typically unknown, these findings suggest that the PCA<sub>AUX</sub> approach may be a superior approach for including auxiliary variables in both FIML and MI.

### **Relative Efficiency**

The relative efficiency findings were in line with theoretical expectations. That is, the PCA<sub>AUX</sub> approach consistently yielded at least as efficient estimates and typically more efficient estimates than the AUX method. More specifically, the current results suggest that 6% efficiency gains are common given a 30% missing data rate. Gains of 20% - 26% are also possible when the associations among the auxiliary variables were high ( $\rho = .60$ ) and when the missing data percentages were high (60% missing). The magnitude of these efficiency gains are similar to those reported by Enders and Bandalos (2001) in an investigation of the relative efficiency of FIML versus pairwise deletion.

These results reflect the idea that  $PCA_{AUX}$  is most effective when the set of auxiliary variables share a large amount of variance as this information can be reduced into a more efficient set. While there were no theoretical expectations regarding sample size, the fact that sample size was not influential was somewhat surprising as large datasets have more information to inform the missing data handling procedure. The  $PCA_{AUX}$  method and the AUX method demonstrated equivalent efficiency when the associations among the auxiliary variables were low ( $\rho = .10$ ) and when the missing data percentages were also low (10% missing); however, these approaches became more pronounced as these two conditions increased (see Table 13). To summarize, these results suggest that regardless of sample size the  $PCA_{AUX}$  method is at worst as good as including all possible auxiliary variables and at best more than 25% more efficient.

### **Empirical Example**

The empirical data example findings were important to distinguish between an inclusive auxiliary variable strategy (AUX; where as many auxiliary variables are included as possible) and an all-inclusive auxiliary variable strategy ( $PCA_{AUX}$ ; where all auxiliary variables, their product terms, and powered terms are included via PCA auxiliary variables). It is important to distinguish between AUX and  $PCA_{AUX}$  in this context because the findings are based on a large dataset with thousands of potential auxiliary variables where the cause of missingness is beyond the researcher's control.

As discussed, recent research (Buhi, Goodson, & Neilands, 2008; Collins et al., 2001; Peugh & Enders, 2004; Schafer & Graham, 2002) suggests that the most appropriate approach for correcting missing data is either full-information maximum likelihood (FIML) or multiple imputation (MI). These techniques are based on sound theory, unlike many of the more traditional methods for addressing missing data (Collins et al.; Enders, 2010). Buhi et al. noted that FIML and MI are considered "state of the art" techniques and they often arrive at the same solution. One of the primary differences is that MI has been generally better able to incorporate

numerous auxiliary variables (Collins et al., 2001; Enders, 2010). That is, MI may generate better estimates due simply to the inclusion of more auxiliary variables.

The results indicate that the  $PCA_{AUX}$  approach improves parameter estimation compared to the absence of auxiliary variables and beyond the improvement of typical auxiliary variables. Also, because the number of auxiliary variables used may be significantly decreased without losing important information, the slight advantage that some methodologists attribute to MI over FIML in terms of auxiliary variables may no longer be relevant.

### **Limitations**

Although our findings provide strong support for the  $PCA_{AUX}$  approach, several limitations are noteworthy. First, the findings regarding the relative performance of the inclusive strategy with typical auxiliary variables (AUX) and a smaller set of auxiliary variables derived from principal component analysis ( $PCA_{AUX}$ ) were based on a relatively simple simulation model (see Figure 1). While this study compared the relative performance of AUX and  $PCA_{AUX}$  across various factors that impact the estimation of auxiliary variables (e.g., sample size, missing data rate, etc.), these results may not generalize to applied research settings. For instance, the current study generated data from a multivariate normal population with no model misspecifications (i.e., perfect model fit). The extent to which model misspecification may influence the results is unknown. Although the results support the effectiveness of the  $PCA_{AUX}$  approach relative to the AUX inclusive strategy with large data sets, further research is needed to investigate the relative performance of  $PCA_{AUX}$  and AUX across more complex models.

Questions also remain regarding the population values selected for the association among the auxiliary variables and the analysis variables as these relationships were clearly an important feature of this study. Some research suggests that the association direction across auxiliary variables (i.e., positively or negatively correlated) may influence the effectiveness of an inclusive auxiliary variable strategy (see Kreuter and Olson, 2011). Considering that the  $PCA_{AUX}$  approach

generates a set of auxiliary variables that are uncorrelated by definition, the degree of influence this may have on the current results is likely trivial. However, homogeneity of correlations across auxiliary variables (i.e., uniform, moderate homogeneity, low homogeneity) may influence the current findings and thus more research is needed to provide better guidance.

With regard to missingness, the current study imposed a univariate monotone missing data pattern on the analysis variable  $Y$ , however, it should be noted that in practice missing data are likely to occur across all variables, including the auxiliary variables used to generate PCA auxiliary variables. Like many other multivariate analytic techniques, PCA requires complete data. Thus research is needed to assess the effectiveness of intermediate steps to acquire a complete data set for the principal component analysis.

In addition, there is some uncertainty regarding the number of principal components (i.e., the percentage of the total variance) to retain. While the current findings suggest that the percent of useful variance may be quite small relative to the total variance, focused research is needed to provide better guidance.

### **Conclusions**

This paper compared the relative performance of the inclusive strategy with typical auxiliary variables (AUX) and a smaller set of auxiliary variables derived from principal component analysis ( $PCA_{AUX}$ ). We conclude that using principal component analysis (PCA) to obtain auxiliary variables for large data sets is a recommended alternative to a restrictive auxiliary variable strategy when the missing data handling technique fails to converge because the number of auxiliary variables is beyond a practical limit. A single  $PCA_{AUX}$  (which explained about 70% of the total variance) was as beneficial for parameter bias as the inclusive strategy (which included all auxiliary variables) with no obvious drawbacks. That is, the  $PCA_{AUX}$  approach performed at least as well as the implementation of a typical inclusive strategy and under certain conditions may perform much better. Our results suggest that the  $PCA_{AUX}$

approach allows for an “all-inclusive” auxiliary variable strategy that provides a more reasonable approximation of the MAR assumption by extracting all the useful information (both linear and non-linear) in the original data set.

### **Implementing the $PCA_{AUX}$ Strategy**

Previously, the similarities and differences related to AUX and  $PCA_{AUX}$  were discussed in their prospective historical contexts and in terms of their relative performance. However, researchers that are interested in utilizing the  $PCA_{AUX}$  approach will need to carefully consider its implementation. In order to provide guidance on these decisions, especially in relation to setting up the PCA procedure, an appendix is provided (see Appendix H).

## References

- Acock, A. C. (2008). *A gentle introduction to Stata*: Stata Press.
- Aldrich, J. (2007). *The Enigma of Karl Pearson and Bayesian Inference*. Paper presented at the Karl Pearson Sesquicentenary Conference, London, England.
- Allan, F., & Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *Journal of Agricultural Science*, 20(3), 399-406.
- Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, 17, 71-103.
- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications, Inc.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 545-557.
- Allison, P. D. (2006). Multiple imputation of categorical variables under the multivariate normal model. Retrieved from <http://www.statisticalhorizons.com/wp-content/uploads/2012/01/Allison.CatVarImp.pdf>
- Anderson, H. J. (1830). *On the motion of solids on surfaces, in the two hypotheses of perfect sliding and perfect rolling, with a particular examination of their small oscillatory motions* (Vol. 3). Philadelphia, PA: James Kay, Jun. & Co.
- Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278), 200-203.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In M. G.A. & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Asparouhov, T., & Muthén, B. (2010). Multiple Imputation with Mplus: Technical Report. Retrieved from [www.statmodel.com](http://www.statmodel.com).

- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37.
- Bartlett, M. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Supplement to the Journal of the Royal Statistical Society, 4*(2), 137-183.
- Beunckens, C., Molenberghs, G., Thij, H., & Verbeke, G. (2007). Incomplete hierarchical data. *Statistical methods in medical research, 16*, 457-492.
- Bock, D. R. (2007). Rethinking Thurstone. In R. Cudeck, & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 35-46). Mahwah, New Jersey: Lawrence Erlbaum.
- Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal, 15*(4), 651-675.
- Brandt, A. E. (1933). The Analysis of Variance in a "2 x s" Table with Disproportionate Frequencies. *Journal of the American Statistical Association, 28*(182), 164-173.
- Brown, B. (1932). A sampling test of the technique of analyzing variance in a 2xn table with disproportionate frequencies. Paper presented at the Iowa Academy of Science.
- Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(4), 287-316.
- Brown, T. A. (2006). *Confirmatory Factor Analysis for applied research*. New York: The Guilford Press.
- Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92.
- Buzhardt, J., Greenwood, C. R., Walker, D., Carta, J. J., Terry, B., & Garrett, M. (2010). Web-based tools to support the use of data-based early intervention decision making. *Topics in Early Childhood Special Education, 29*(4), 201-214.

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed., pp. 1-691). Mahwah, NJ: Lawrence Erlbaum Associates.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330-351.
- Conniffe, D. (1991). R. A. Fisher and the development of statistics - A view in his centenary year. *Journal of the Statistical and Social Inquiry Society of Ireland*, 26(3), 55-108.
- Cornish, E. A. (1944). The recovery of inter-block information in quasi-factorial designs with incomplete data. *Annals of Human Genetics*, 10(1), 137-143.
- Cowles, M. (2001). *Statistics in psychology: An historical perspective* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum.
- Cytel. (2009). StatXact User Manual (Version 8.0) [for Windows]. Cambridge, MA.
- Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1), 69-84.
- 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 (Methodological)*, 1-38.
- Dunteman, G. H. (1989). *Principal components analysis*. Newbury Park, CA: SAGE publications.
- Dodge, Y. (1985). *Analysis of experiments with missing data*. Wiley Series in Probability and Mathematical Statistics, New York: Wiley & Sons.
- Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1), 128-141.

- Enders, C. K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research*, 37(3), 359-377.
- Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, 8(3), 322-337.
- Enders, C. K. (2006). A Primer on the use of modern missing data methods in psychosomatic medicine research. *Psychosomatic Medicine*, 68(3), 427-736.
- Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, 15(3), 434 – 448.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430-457.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35-54.
- Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, 11(1), 1-19.
- Evans, M. J. (2005). *Minitab Manual for Moore and McCabe's Introduction to the Practice of Statistics* (5th ed.). New York: W.H. Freeman and Company.
- Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, 44(4), 409-420.
- Fisher, R., Immer, F., & Tedin, O. (1932). The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics*, 17(2), 107-124.

- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Great Britain: Oliver and Boyd.
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*(1), 80-100.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.
- Graham, J. W. (2012). Missing data: Analysis and design [PowerPoint slides]. Retrieved from Pennsylvania State University, The Methodology Center website:  
<http://methodology.psu.edu>
- Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2002). Methods for handling missing data. In J. A. Schinka, W. F. Velicer & I. B. Weiner (Eds.), *Handbook of Psychology, Research Methods in Psychology* (Vol. 2, pp. 87-114). New York: John Wiley & Sons.
- Graham, J. W., Cumsille, P. E., & Shevock, A. E. (in press). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research Methods in Psychology* (Vol. 3). New York: John Wiley & Sons.
- Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. *NIDA Research Monograph, 142*, 13-63.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213.
- Graham, J. W., & Schafer, J. L. (1999). Methods for handling missing data. In R. H. (Ed), *Statistical strategies for small sample research* (pp. 1-27). Thousand Oaks, CA: Sage Publications, Inc.

- Greenwood, C. R., Buzhardt, J., Walker, D., Howard, W. J., & Anderson, R. (2011). Program-Level Influences on the Measurement of Early Communication for Infants and Toddlers in Early Head Start. *Journal of Early Intervention, 33*(2), 110-134.
- Greenwood, C. R., & Walker, D. (2010). Development and validation of IGDIs. In J. J. Carta, C. R. Greenwood, D. Walker & J. Buzhardt (Eds.), *Using IGDIs: Monitoring progress and improving intervention for infants and young children*. Baltimore, MD: Brooks.
- Hald, A. (1999). On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science, 14*(2), 214-222.
- Hartley, H., & Hocking, R. (1971). The analysis of incomplete data. *Biometrics, 27*(4), 783-823.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth: Harcourt Brace College Publishers.
- Hawkins, T. (1974). *The theory of matrices in the 19th century*. Paper presented at the Proceedings of the International Congress of Mathematicians, Vancouver, Canada.
- Hawkins, T. (1975). Cauchy and the spectral theory of matrices. *Historia Mathematica, 2*(1), 1-29.
- Healy, M. (1995). Frank Yates, 1902-1994: the work of a statistician. *International Statistical Review/Revue Internationale de Statistique, 63*(3) 271-288.
- Herr, D. G. (1986). On the history of ANOVA in unbalanced, factorial designs: The first 30 years. *The American Statistician, 40*(4) 265-270.
- Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*(4), 229-232.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.
- Hox, J. J., & Roberts, K. J. (2010). *The handbook of advanced multilevel analysis* (series: European Association for Methodology Series) (Vol. viii). New York, NY: Routledge/Taylor & Francis Group.

- Hunt, D., & Triggs, C. (1989). Iterative missing value estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(2) 293-300.
- Iacobucci, D. (1995). The analysis of variance for unbalanced data. *Marketing theory and applications*, 6, 337-343.
- IBM SPSS Statistics 20 Core System User's Guide. (2011). 1-90. Retrieved from  
doi:ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM\_SPSS\_Missing\_Values.pdf
- Jackson, J. E. (1991). *A user's guide to principal components*. New York: John Wiley & Sons Inc.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.
- Jöreskog, K. G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika*, 36, 409-426.
- Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8 user's reference guide: Scientific Software.
- Kline, R. (2010). *Principles and practice of structural equation modeling*. New York: The Guilford Press.
- Kolman, B. (1996). *Elementary Linear Algebra* (6th ed.). Upper Saddle River, New Jersey: Prentice Hall
- Kreuter, F., & Olson, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods & Research*, 40(2), 311-332.
- Kutner, M., Nachtsheim, C., Li, W., & Neter, J. (2005). *Applied linear statistical models*. Boston: McGraw Hill.
- Little, R. J. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26(2), 162-174.

- Little, R. J., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2), 161-168.
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404) 1198-1202.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.
- Little, T. D. (2009, June). *Structural equation modeling foundations and extended applications: Missing data, power, and sample size*. Presentation given at the University of Kansas Summer Institutes in Statistics, Lawrence, Kansas.
- Luze, G. J., Linebarger, D. L., Greenwood, C. R., Carta, J. J., Walker, D., Leitschuh, C., et al. (2001). Developing a general outcome measure of growth in expressive communication of infants and toddlers. *School Psychology Review*, 30(3), 383-406.
- Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(1), 22-36.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Methodology in the social sciences. New York, NY, US: Guilford Press.
- McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies* (Vol. 26): John Wiley & Sons Inc.
- Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen.

- Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6(3), 328-362.
- Noruésis, M. (1990). SPSS base system user's guide: Chicago, Ill: SPSS Inc.
- Orchard, T., & Woodbury, M. A. (1972). *A missing information principle: theory and applications*. Paper presented at the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA.
- Pearson, E. S. (1936). Karl Pearson: An appreciation of some aspects of his life and work. *Biometrika*, 28(3/4), 193-257.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2), 559-572.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525-556.
- Porter, T. M. (2004). *Karl Pearson: The scientific life in a statistical age*. Princeton, New Jersey: Princeton University Press.
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13-43.
- Savalei, V., & Rhemtulla, M. (2012). On Obtaining Estimates of the Fraction of Missing Information From Full Information Maximum Likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3), 477-494. doi: 10.1080/10705511.2012.687669
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7(2), 147-177.

- Schoemann, A. M., Pornprasertmanit, S., & Miller, P. (July, 2012). *simsem: SIMulated Structural Equation Modeling in R*. Lawrence, KS: Retrieved from <http://cran.r-project.org/web/packages/simsem/index.html>.
- Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health, 25*, 99-117.
- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). A user's guide to MLwiN, v2. 10. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.
- Raudenbush, S. W. (2004). HLM 6: Hierarchical linear and nonlinear modeling: Scientific Software International.
- Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. New York: Routledge Taylor & Francis Group.
- Rubin, D. B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 21(2)* 136-141.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63(3)*, 581-592.
- Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72(359)* 538-543.
- Rubin, D. B. (1978a). Multiple imputation in sample surveys. A phenomenological Bayesian approach to nonresponse. *In Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce, Social Security Administration.
- Rubin, D. B. (1978b). A note on Bayesian, likelihood, and sampling distribution inferences. *Journal of Educational and Behavioral Statistics, 3(2)*, 189-201.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*: New York: John Wiley & Sons.

- Rubin, D. B., & Little, R. J. A. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.
- SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*: SAS Institute Inc.
- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 477-497.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall/CRC.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7(2), 147-177.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research. Special Issue: Innovative methods for prevention research*, 33(4), 545-571.
- Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for unbalanced data: an overview. *Ecology*, 74(6) 1638-1645.
- Shearer, P. (1973). Missing data in quantitative designs. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(2) 135-140.
- Smith, L. I. (2002). A tutorial on principal components analysis. Retrieved from [www.cs.otago.ac.nz/cosc453/student\\_tutorials/principal\\_components.pdf](http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf)
- Student. (1908). The probable error of a mean. *Biometrika*, 6(1) 1-25.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398) 528-540.
- Tanner, M. A., & Wong, W. H. (2010). From EM to Data Augmentation: The Emergence of MCMC Bayesian Computation in the 1980s. *Statistical Science*, 25(4), 506-516.

- Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064.
- Walker, D., & Buzhardt, J. (2010). IGDI Administration: Coding, scoring, and graphing. In J. J. Carta, C. R. Greenwood, D. Walker & J. Buzhardt (Eds.), *Using IGDI: Monitoring progress and improving intervention results for infants and young children*. Baltimore: Brookes.
- Walker, D., & Carta, J. J. (2010). The Communication IGDI: Early Communication Indicator. In J. J. Carta, C. R. Greenwood, D. Walker & J. Buzhardt (Eds.), *Using IGDI: Monitoring progress and improving intervention results for infants and young children*. Baltimore: Brookes.
- Watanabe, M., & Yamaguchi, K. (2004). *The EM algorithm and related statistical models*. New York, NY: Marcel Dekker Inc.
- Wickens, T. D. (1995). *The geometry of multivariate statistics*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 177-203). Mahwah, New Jersey: Lawrence Erlbaum Associates
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594-604.
- Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3), 163-195.
- Wothke, W. (1993). *Nonpositive definite matrices in structural modeling*: Newbury Park, CA: Sage Publications Inc.

- Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.). *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219-240). Mahwah, NJ: Erlbaum.
- Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, *1*(2), 129-142.
- Yates, F. (1936). Incomplete Latin squares. *Journal of Agricultural Science*, *26*, 301-315.
- Yates, F. (1968). Theory and practice in statistics. *Journal of the Royal Statistical Society. Series A (General)*, *131*(4), 463-477.
- Yates, F., & Hale, R. (1939). The analysis of Latin squares when two or more rows, columns, or treatments are missing. *Supplement to the Journal of the Royal Statistical Society*, *6*(1), 67-79.
- Yates, F., & Mather, K. (1963). Ronald Aylmer Fisher. 1890-1962. *Biographical Memoirs of Fellows of the Royal Society*, *9*, 91-129.
- Yoo, J. E. (2009). The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement*, *69*(6), 929-947.
- Yuan, Y. (2000). Multiple Imputation for Missing Values: Concepts and New Development. SUGI Proceedings. Retrieved from <http://support.sas.com/rnd/app/stat/papers/abstracts/multipleimputation.html>
- Yuan, K. H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology*, *30*(1), 165-200.
- Zhang, D. (2005). A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling. College Station, TX: Texas A&M University.

## Appendix A: Marginal Means with Unbalanced Data

When the data are unbalanced, the marginal means are non-orthogonal (Shaw & Mitchell-Olds, 1993). The following discussion demonstrates this point by reviewing the calculations of marginal means (see Tables 17 and 18). In this example (modified from Iacobucci, 1995), raw data from four groups are arranged in a 2 x 2 table, where the two rows represent experimental blocks  $B$  (or conditions;  $b_1, b_2$ ) and the two columns represent conditions  $A$  ( $a_1, a_2$ ). The average of each row and of each column is displayed in the margins of the table. The ANOVA model for this design can be written as:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \quad (109)$$

where  $\mu$  represents a grand mean,  $\alpha_j$  is the effect of the  $j^{\text{th}}$  level of  $A$ ,  $\beta_k$  is the effect of the  $k^{\text{th}}$  level of  $B$ , and  $(\alpha\beta)_{jk}$  is the interaction effect of level  $j$  of  $A$  and level  $k$  of  $B$ , and  $\varepsilon_{ijk}$  is the error term for each participant. In the current example, there are three possible hypotheses including the main effect of  $A$  ( $H_0: a_1 - a_2 = 0$ ), the main effect of  $B$  ( $H_0: b_1 - b_2 = 0$ ), and the interaction between  $A$  and  $B$  ( $H_0: ab = 0$ ). Note that each observed value in Table 11 is written in terms of the associated model parameters (excluding  $(\alpha\beta)_{jk}$  and  $\varepsilon_{ijk}$  for simplicity of the current illustration). That is, the first observed value in Table 17 Condition  $a_1$  and Block  $b_1$  is 13, a function of the grand mean ( $\mu$ ), and main effects of  $\alpha_1$  and  $\beta_1$ . To illustrate a simple example of how these hypotheses change in relation to missing data, consider the main effect of  $B$  in Table 17. Note that Table 17 contains complete (balanced) raw data and demonstrates the calculation of each marginal mean of  $B$ . Simplifying these expressions results in the following hypothesis test:

$$\begin{aligned} \bar{b}_1 - \bar{b}_2 &= \left[ (10\mu + 5\alpha_1 + 5\alpha_2 + 10\beta_1) / 10 \right] - \left[ (10\mu + 5\alpha_1 + 5\alpha_2 + 10\beta_2) / 10 \right] \\ &= (\beta_1 - \beta_2) \end{aligned} \quad (110)$$

Note that in the balanced situation the null hypothesis for the main effect of  $B$  simplifies to  $H_0: b_1 - b_2 = 0$  as it should. Now consider the main effect of  $B$  in Table 18. Notice that Table 18 contains incomplete (unbalanced) raw data. Specifically, there is a missing value (i.e.,  $12 = \mu + \alpha_2 + \beta_1$ ) in Condition  $a_2$  and Block  $b_1$ . The calculation of each marginal mean of  $B$  proceeds in the same manner as previously described. However, this time notice that simplifying the marginal mean expressions results in the following hypothesis test:

$$\begin{aligned}\bar{b}_1 - \bar{b}_2 &= [(9\mu + 5\alpha_1 + 4\alpha_2 + 9\beta_1)/9] - [(10\mu + 5\alpha_1 + 5\alpha_2 + 10\beta_2)/10] \\ &= (\beta_1 - \beta_2) + .04\alpha_1 + .06\alpha_2\end{aligned}\tag{111}$$

Here the null hypothesis for the main effect of  $B$  simplifies to  $H_0: b_1 - b_2 + .04a_1 + .06 + 0.06a_2 = 0$ . As demonstrated, the observed difference in marginal means for the two levels of  $B$  is a measure of the effect of  $B$  as well as a biasing effect of  $A$ . When the data are unbalanced, the marginal means are non-orthogonal and the sums of squares computed from these means are contaminated with functions of other parameters (Iacobucci, 1995; Shaw & Mitchell-Olds, 1993).

## Appendix B: Yates Formula Example

In order to highlight the parallels between Allan and Wishart (1930) and Yates, consider the following equation (see Yates, 1933) using the data in Table 1:

$$y = \frac{r(B) + t(T) - G}{(r-1)(t-1)} = \frac{8(174.37) + 9(132.98) - 1417.48}{(9-1)(8-1)} = 20.97 \quad (112)$$

where  $r$  is the number of blocks,  $t$  is the number of treatments,  $B$  is the total from the block that contains the missing value,  $T$  is the total from the treatment that contains the missing value and  $G$  is overall total. While Allan and Wishart (1930) focused on totals involving the blocks and treatments *without* a missing value, Yates (1933) used totals from the block and treatment *with* the missing value (see Formula 2). While conceptually, these approaches may seem different, the solutions are equivalent.

As previously described, the benefit of Yates' approach over that of Allan and Wishart (1930) relates to situations with more than one missing value. Yates' method can be used to impute several missing values so that standard statistical analytic methods can be applied (Dodge, 1987; Little & Rubin, 2002). To best illustrate Yates' 1933 method in relation to least-squares estimation, consider the following example. To begin with recall the simple regression equation reproduced below:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + e_i \quad (113)$$

where  $\beta_0$  is the intercept,  $\beta_1$  signifies a slope with a predictor  $X$  and an associated error term  $e$ .

Now let the equation above be expressed more compactly by utilizing a set of matrices (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (114)$$

where  $\mathbf{Y}$  is an  $(N \times 1)$  vector that contains the  $i$  dependent variable  $\hat{Y}_i$ , given that  $k$  represents the number of variables  $\mathbf{X}$  is an  $(N \times (k - 1))$  design matrix that contains a column of 1s and  $k - 1$

columns of predictors,  $\mathbf{b}$  is a  $(k \times 1)$  vector of beta weights, and  $\mathbf{e}$  is an  $(N \times 1)$  residual vector.

When illustrated this formula becomes:

$$\begin{aligned} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{N1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \\ \mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \\ (N \times 1) & \quad (N \times (k-1)) \quad (k \times 1) \quad (N \times 1) \end{aligned} \quad (115)$$

where  $\mathbf{e} \sim N_N(0, \sigma_e^2 \mathbf{I}_N)$  and the first case is  $Y_1 = \beta_0 + \beta_1 X_{11} + e_1$ . Given the previous matrix model, assume that the matrix  $\mathbf{Y}$  contains missing data. More specifically, let  $Y_1, \dots, Y_m$  represent observed values ( $Y_{\text{obs}}$ ), while  $Y_{m+1}, \dots, Y_p$  indicate missing values ( $Y_{\text{miss}}$ ; see Figure 1). Now, for demonstration purposes let  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  be rewritten to align  $Y_{\text{obs}}$  with  $X_{\text{obs}}$  and  $Y_{\text{miss}}$  with  $X_{\text{miss}}$  such that the formula resembles:

$$\begin{aligned} \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_p \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{m1} \\ 1 & X_{m+1,1} \\ \vdots & \vdots \\ 1 & X_{p1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_m \\ e_{m+1} \\ \vdots \\ e_p \end{bmatrix} \\ \mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \end{aligned} \quad (116)$$

In this way, the Yates method can be directly implemented to estimate  $Y_{\text{miss}}$  via the *least-squares criterion* by minimizing the residuals which can be expressed as (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\sum_{i=1}^N e^2 = \sum_{i=1}^N (Y_{\text{obs},i} - \mathbf{x}'_{\text{obs},i} \mathbf{b})^2 = (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b})' (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b}) = \mathbf{e}' \mathbf{e} \quad (117)$$

This equation can then be written in relation to  $\mathbf{b}$  that minimizes  $\mathbf{e}' \mathbf{e}$  as:

$$\mathbf{b} = (\mathbf{X}'_{\text{obs}} \mathbf{X}_{\text{obs}})^{-1} \mathbf{X}'_{\text{obs}} \mathbf{Y}_{\text{obs}} \quad (118)$$

where  $\mathbf{b}$  is a  $((k - 1) \times 1)$  vector of the least-squares beta estimates that minimizes the sum of squared residuals of a  $(m \times 1)$  vector of  $\mathbf{Y}_{obs}$  (note that this would be the mean of  $Y$  without predictors in the model) based on the associated  $(m \times 1)$  vector of  $\mathbf{X}_{obs}$ . Once  $\mathbf{b}$  is obtained it is then used to predict the missing  $Y_{m+1}, \dots, Y_p$  values from the observed  $X_{m+1}, \dots, X_p$  values. To illustrate this process, let example data from Blocks 4 and 5 of Table 1 be treated as  $\mathbf{Y}$  and  $\mathbf{X}$ , respectively, in the forthcoming example; see Figure 53). Given the observed data ( $\mathbf{Y}_{obs}, \mathbf{X}_{obs}$ ) in Figure 53, the next step is to apply the regression formula which can be demonstrated as:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \\
\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \left( \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \\
&= \begin{bmatrix} N & \sum x \\ \sum x & \sum x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} = \frac{1}{|\mathbf{X}'\mathbf{X}|} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & N \end{bmatrix} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \\
&= \frac{1}{N \sum x^2 - (\sum x)^2} \begin{bmatrix} \sum x^2 \sum y - \sum x \sum xy \\ N \sum xy - \sum x \sum y \end{bmatrix} = \begin{bmatrix} \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2} \\ \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\frac{1}{N^2} (\sum x^2 \sum y - \sum x \sum xy)}{\frac{1}{N^2} (N \sum x^2 - (\sum x)^2)} \\ \frac{\frac{1}{N} (\sum xy - \frac{1}{N} \sum x \sum y)}{\frac{1}{N} (\sum x^2 - \frac{1}{N} (\sum x)^2)} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ \frac{\text{cov}(x, y)}{\text{var}(x)} \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix} \\
&= \begin{bmatrix} 18.99 - 19.47 \frac{2.30}{4.12} \\ \frac{2.30}{4.12} \end{bmatrix} = \begin{bmatrix} 8.14 \\ 0.56 \end{bmatrix}
\end{aligned} \tag{119}$$

where  $\Sigma$  indicates summation,  $N$  is the observed sample size, cov denotes the covariance, var represents the variance,  $b_0$  is the intercept parameter and  $b_1$  is the slope parameter (see Kutner,

Nachtsheim, Neter, and Li, 2005). Note that calculations for the intercept and slope regression parameters are illustrated simultaneously in the above equation.

Given estimates of  $b_0$  and  $b_1$  the next step is to predict  $Y_{\text{miss}}$  (the missing value) from  $X_{\text{miss}}$  (i.e., 20.22; an observed value in  $X$  that corresponds to a missing value in  $Y$ ; see Figure 53). The subsequent equation can be illustrated as:

$$\begin{aligned} Y_{\text{miss}} &= b_0 + b_1 X_{\text{miss}} \\ Y_{\text{miss}} &= 8.1373 + .5577(20.22) \\ Y_{\text{miss}} &= 19.414 \end{aligned} \tag{120}$$

To highlight the least squares criterion in this imputation method, notice that imputing 19.414 for  $Y_{\text{miss}}$  and estimating  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  produces the following set of residuals (see  $\mathbf{e}$  vector below):

$$\begin{array}{c} \mathbf{Y} \\ \left[ \begin{array}{c} 21.15 \\ 19.41 \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{array} \right] \end{array} = \begin{array}{c} \mathbf{X} \\ \left[ \begin{array}{cc} 1 & 21.14 \\ 1 & 20.22 \\ 1 & 20.81 \\ 1 & 20.69 \\ 1 & 19.59 \\ 1 & 18.33 \\ 1 & 15.40 \\ 1 & 20.34 \end{array} \right] \end{array} \begin{array}{c} \mathbf{b} \\ \left[ \begin{array}{c} 8.14 \\ 0.56 \end{array} \right] \end{array} + \begin{array}{c} \mathbf{e} \\ \left[ \begin{array}{c} 1.22 \\ 0.00 \\ 0.66 \\ 0.28 \\ -1.85 \\ 0.93 \\ 0.18 \\ -1.42 \end{array} \right] \end{array} \tag{121}$$

where the imputed value ( $Y_{\text{miss}}$ ) has a residual of zero (see Appendix B for example R code).

Importantly, note that this least-squares regression technique generates imputed values that lie directly on a regression line constructed through the remaining  $n - 1$  data points, which decreases the natural variability present in the data (Shearer, 1973). Said differently, the least-squares regression technique outlined above fails to incorporate appropriate uncertainty about what the imputed value should be. Therefore, it does not capture the variability that would have been present had the data point not been missing.

### Appendix C: Hartley and Hocking (1971) Illustration

To demonstrate how Hartley and Hocking selected information for each of the missing data patterns, consider the data  $\mathbf{XY}$  mentioned previously with the missing data pattern illustrated in Table 11. Prior to maximum likelihood estimation the  $\mathbf{XY}$  matrix can be replaced by  $\mathbf{XY}^* = \mathbf{Z}(\mathbf{XY})$  where  $\mathbf{Z}$  is a selection matrix used to identify observed values within  $\mathbf{XY}$ . To illustrate a specific example consider the case of  $\text{FML}_n$  (i.e.,  $\mu_y$ ) where the column vector  $\mathbf{Y}$  is substituted by  $\mathbf{Y}^* = \mathbf{Z}(\mathbf{Y})$  as:

$$\begin{array}{c}
 \mathbf{Y}^* \\
 \left[ \begin{array}{c} y_1 \\ \vdots \\ y_i \\ y_{i+1} \\ \vdots \\ y_j \\ \vdots \\ y_j \end{array} \right] \\
 \text{FML}_n \text{ data}
 \end{array}
 =
 \begin{array}{c}
 \mathbf{Z} \\
 \left[ \begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]
 \end{array}
 \underbrace{\hspace{10em}}_{\text{Observed Data Selection Matrix}}
 \begin{array}{c}
 \mathbf{Y} \\
 \left[ \begin{array}{c} y_1 \\ \vdots \\ y_i \\ y_{i+1} \\ \vdots \\ y_j \\ - \\ \vdots \\ - \end{array} \right] \\
 \text{Original Data}
 \end{array}
 \quad (122)$$

where the first element in  $\mathbf{Y}^*$  is  $1(y_1)+, \dots, 0(y_i)+0(y_{i+1}), \dots, 0(y_j)+0(y_{j+1}), \dots, 0(y_k)$  and the last element in  $\mathbf{Y}^*$  is  $0(y_1)+, \dots, 0(y_i)+0(y_{i+1}), \dots, 1(y_j)+0(y_{j+1}), \dots, 0(y_k)$ . Notice that the column vector  $\mathbf{Y}^*$  is based on the identity matrix  $\mathbf{Z}$  with rows deleted that correspond to missing values (e.g.,  $y_{j+1}, \dots, y_k$ ). This was accomplished by altering the diagonal elements of the identity matrix  $\mathbf{Z}$ . Recall that pre- or post-multiplication of a matrix by an identity matrix does not change the matrix (i.e., nor does multiplying any scalar by 1). Therefore, given complete data  $\mathbf{ZY} = \mathbf{Y}^* = \mathbf{Y}$ . However, by changing the diagonal elements of the matrix of  $\mathbf{Z}$  from 1s to 0s the resulting pre-multiplication changes the resultant matrix by dropping the corresponding rows of

**Y**. In this way  $\mathbf{Y}^*$  is still considered a random sample drawn from a normal distribution and the typical likelihood equation can be applied. In this way Hartley and Hocking replaced the data matrix with the multiple of a modified identity matrix (coded to correspond to a particular missing data pattern) and the data matrix, which collapsed the missing elements from the likelihood function of the data matrix.

## Appendix D: Rubin's (1976) Non-Ignorable Missing Data

Consider a brief example to illustrate the concept. Standard inferential procedures seek to make probabilistic statements about population parameters from sample statistics. Let  $\hat{\theta}$  be an estimate of population parameter  $\theta$ . A null hypothesis states that  $\theta$  equals  $\theta_0$  in the population of interest, while an alternative hypothesis asserts that  $\theta$  does not equal  $\theta_0$ . A test statistic ( $T$ ) is a function of  $\theta$  and the sample data that includes a parameter estimate, a standard error for the parameter estimate, and a critical value on a reference distribution, which can be written as:

$$\frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} \sim T(1 - \alpha/2, df) \quad (123)$$

where  $SE(\hat{\theta})$  is the standard error for  $\hat{\theta}$  and  $T(1 - \alpha/2, df)$  is a reference distribution with an associated critical value under the assumption that  $\theta$  equals  $\theta_0$ . The null hypothesis is rejected for p-values less than  $\alpha$ , meaning that the observed distribution of the test statistic under the null hypothesis is unusual given the reference distribution. Similarly, a confidence interval for  $\theta$ , a range around  $\hat{\theta}$  where  $\theta$  lies with a given probability (i.e.,  $1 - \alpha$ ), can be written as

$$\hat{\theta} \pm T_{(1-\alpha/2, df)} * SE(\hat{\theta}) \quad (124)$$

which relays the probability that  $\theta$  falls within a specified interval. Since Frequentist statistical inference is based on a probability statement about the distribution of the population from which the sample was drawn, not a statement about the parameter (i.e., the parameter estimate is a fixed characteristic of the population), the values of the observed data must arise from a known probability distribution, to appropriately calculate  $\hat{\theta}$  and the associated  $SE$ .

## Appendix E: Demonstration of the EM Algorithm

To demonstrate the EM algorithm using the linear model equation, first recall the simple regression equation (for the  $i^{\text{th}}$  observation, prediction of  $Y_i$  by  $k$  variables in  $X_{ik}$ ):

$$Y_i = \beta_0 + \beta_1 X_{i1} + e_i \quad (125)$$

where  $\beta_0$  is the intercept,  $\beta_1$  signifies a slope with a predictor  $X$  and an associated error term  $e$ .

Now let the equation above be expressed more compactly by utilizing a set of matrices (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (126)$$

where  $\mathbf{Y}$  is an  $(N \times 1)$  vector that contains the dependent variable  $y$ ,  $\mathbf{X}$  is an  $(N \times (k - 1))$  design matrix that contains a column of 1s and  $k - 1$  columns of predictors,  $\mathbf{b}$  is a  $(k \times 1)$  vector of beta weights, and  $\mathbf{e}$  is an  $(N \times 1)$  residual vector. When illustrated this formula becomes:

$$\begin{aligned} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{N1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \\ \mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \\ (N \times 1) & \quad (N \times (k-1)) (k \times 1) (N \times 1) \end{aligned} \quad (127)$$

where  $\mathbf{e} \sim N_N(0, \sigma_e^2 \mathbf{I}_N)$  and the first case is  $Y_1 = \beta_0 + \beta_1 X_{11} + e_1$ . Given the previous matrix model, assume missing data on  $\mathbf{Y}$ . More specifically, let  $Y_1, \dots, Y_m$  represent observed values ( $Y_{\text{obs}}$ ), while  $Y_{m+1}, \dots, Y_p$  indicate missing values ( $Y_{\text{miss}}$ ; see Figure 53).

Next, consider that  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  can be rewritten to align  $Y_{\text{obs}}$  with  $X_{\text{obs}}$  and  $Y_{\text{miss}}$  with  $X_{\text{miss}}$  such that the formula resembles:

$$\begin{aligned}
\begin{bmatrix} Y_1 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_p \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ 1 & \vdots \\ 1 & X_{m1} \\ 1 & X_{m+1,1} \\ \vdots & \vdots \\ 1 & X_{p1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_m \\ e_{m+1} \\ \vdots \\ e_p \end{bmatrix} \\
\mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \quad (128)
\end{aligned}$$

Now, the E-step can be directly implemented to estimate  $Y_{\text{miss}}$  via the *least squares criterion* by minimizing (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\sum_{i=1}^N e^2 = \sum_{i=1}^N (Y_{\text{obs},i} - \mathbf{x}'_{\text{obs},i} \mathbf{b})^2 = (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b})' (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b}) = \mathbf{e}'\mathbf{e} \quad (129)$$

which can be written in relation to  $\mathbf{b}$  that minimizes  $\mathbf{e}'\mathbf{e}$  as:

$$\mathbf{b} = (\mathbf{X}'_{\text{obs}} \mathbf{X}_{\text{obs}})^{-1} \mathbf{X}'_{\text{obs}} \mathbf{Y}_{\text{obs}} \quad (130)$$

where  $\mathbf{b}$  is a  $((k - 1) \times 1)$  vector of the least-squares beta estimates that minimizes the sum of squared residuals of a  $(m \times 1)$  vector of  $\mathbf{Y}_{\text{obs}}$  (note that this is the mean of  $y$  without predictors in the model) based on the associated  $(m \times 1)$  vector of  $\mathbf{X}_{\text{obs}}$ . Once  $\mathbf{b}$  is obtained it is then used to predict the missing  $Y_{m+1}, \dots, Y_p$  values from the observed  $X_{m+1}, \dots, X_p$  values. To illustrate this process, let example data from Blocks 4 and 5 of Table 1 be treated as  $\mathbf{Y}$  and  $\mathbf{X}$ , respectively, in the forthcoming example; see Figure 53). Given the observed data  $(\mathbf{Y}_{\text{obs}}, \mathbf{X}_{\text{obs}})$  in Figure 53, the next step is to apply the regression formula (see Equation 10) which can be demonstrated as:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\
\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \left( \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \\
&= \begin{bmatrix} N & \sum x \\ \sum x & \sum x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} = \frac{1}{|\mathbf{X}'\mathbf{X}|} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & N \end{bmatrix} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \\
&= \frac{1}{N \sum x^2 - (\sum x)^2} \begin{bmatrix} \sum x^2 \sum y - \sum x \sum xy \\ N \sum xy - \sum x \sum y \end{bmatrix} = \begin{bmatrix} \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2} \\ \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\frac{1}{N^2} (\sum x^2 \sum y - \sum x \sum xy)}{\frac{1}{N^2} (N \sum x^2 - (\sum x)^2)} \\ \frac{\frac{1}{N} (\sum xy - \frac{1}{N} \sum x \sum y)}{\frac{1}{N} (\sum x^2 - \frac{1}{N} (\sum x)^2)} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ \frac{\text{cov}(x, y)}{\text{var}(x)} \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix} \\
&= \begin{bmatrix} 18.9971 - 19.5650 \frac{1.9687}{3.1499} \\ \frac{1.9687}{3.1499} \end{bmatrix} = \begin{bmatrix} 6.7689 \\ 0.6250 \end{bmatrix} \tag{131}
\end{aligned}$$

where  $\Sigma$  indicates summation,  $N$  is the observed sample size, cov denotes the covariance, var represents the variance,  $b_0$  is the intercept parameter and  $b_1$  is the slope parameter (see Kutner, Nachtsheim, Neter, and Li, 2005). Note that calculations for the intercept and slope regression parameters are illustrated simultaneously in the above equation.

The next step is to predict  $Y_{\text{miss}}$  (a missing value) from  $X_{\text{miss}}$  (i.e., 20.22; an observed value in  $X$  that corresponds to a missing value in  $Y$ ; see Figure 3) using the regression weights observed in  $\mathbf{b}$ . The subsequent equation can be illustrated as:

$$\begin{aligned}
Y_{miss} &= 6.7689 + .6250(20.22) \\
Y_{miss} &= 19.4065
\end{aligned}
\tag{132}$$

To highlight the least squares criterion in this imputation method, notice that imputing 19.4065 for  $Y_{miss}$  and estimating  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  produces the following set of residuals (see  $\mathbf{e}$  vector below):

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \\
\begin{bmatrix} 21.15 \\ 19.41 \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} &= \begin{bmatrix} 1 & 21.14 \\ 1 & 20.22 \\ 1 & 20.81 \\ 1 & 20.69 \\ 1 & 19.59 \\ 1 & 18.33 \\ 1 & 15.40 \\ 1 & 20.34 \end{bmatrix} \begin{bmatrix} 6.769 \\ 0.625 \end{bmatrix} + \begin{bmatrix} 1.22 \\ 0.00 \\ 0.66 \\ 0.28 \\ -1.85 \\ 0.93 \\ 0.18 \\ -1.42 \end{bmatrix}
\end{aligned}
\tag{133}$$

where the imputed value ( $Y_{miss}$ ) has a residual of zero (see Appendix QQQ for example R code).

The E-step uses  $Y_{miss} = 19.4065$  in the  $\sum y$  and  $\sum^{xy}$  calculations; however, the E-step does not literally fill in the missing value (e.g.,  $Y_{miss} = 19.4065$ ); rather this value is only referenced by the following M-step (Enders, 2010). The ensuing M-step uses the now complete data to generate estimates of the sufficient statistics as follows:

$$\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 3.1499 & 1.7561 \\ 1.7561 & 2.0238 \end{bmatrix} \\
\mathbf{\mu} &= \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 19.5650 \\ 19.0483 \end{bmatrix}
\end{aligned}
\tag{134}$$

where  $\mathbf{\mu}$  is a  $2 \times 1$  column vector of means for  $\mathbf{X}$  and  $\mathbf{Y}$ , and  $\mathbf{\Sigma}$  is a  $2 \times 2$  covariance matrix.

These parameter estimates are then carried to the next E-Step as follows:

$$\begin{aligned}
\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \begin{bmatrix} \frac{1}{N^2} (\sum x^2 \sum y - \sum x \sum xy) \\ \frac{1}{N^2} (N \sum x^2 - (\sum x)^2) \\ \frac{1}{N} (\sum xy - \frac{1}{N} \sum x \sum y) \\ \frac{1}{N} (\sum x^2 - \frac{1}{N} (\sum x)^2) \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ \frac{\text{cov}(x, y)}{\text{var}(x)} \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix} \\
&= \begin{bmatrix} 19.0483 - 19.5650 \frac{1.7561}{3.1498} \\ \frac{1.7561}{3.1498} \end{bmatrix} = \begin{bmatrix} 8.1404 \\ 0.5575 \end{bmatrix} \tag{135}
\end{aligned}$$

Again the procedure uses this information to predict  $Y_{\text{miss}}$  (a missing value) from  $X_{\text{miss}}$  (i.e., 20.22; an observed value in  $X$  that corresponds to a missing value in  $Y$ ; see Figure 53) using the regression weights observed in **b**. The subsequent equation can be illustrated as:

$$\begin{aligned}
Y_{\text{miss}} &= 8.1404 + .5575(20.22) \\
Y_{\text{miss}} &= 19.4135 \tag{136}
\end{aligned}$$

The following M-step again uses the estimate  $Y_{\text{miss}} = 19.4135$  in the calculation of another set of sufficient statistics and the process repeats until convergence (subsequent change in the MLE are negligible).

Recall that the value of the log-likelihood function is a measure of joint likelihood of an observed sample of data (represented by the individual data values) given the parameters in the population covariance matrix  $\Sigma$  and the population mean vector  $\mu$  where a smaller value indicates better fit of the data to the population parameters. To demonstrate convergence let, given  $\sigma_Y^2 = 2.0$  and  $\mu_Y = 19.5$  so the likelihood estimates can be found by:

$$\begin{aligned}
f(Y_1) &= \frac{1}{\sqrt{2\pi(2.0)}} e^{-.5 \frac{(21.15-19.5)^2}{2.0}} = 0.143 \\
&\vdots \\
f(Y_4) &= \frac{1}{\sqrt{2\pi(2.0)}} e^{-.5 \frac{(19.96-19.5)^2}{2.0}} = 0.268 \\
&\vdots \\
f(Y_8) &= \frac{1}{\sqrt{2\pi(2.0)}} e^{-.5 \frac{(18.06-19.5)^2}{2.0}} = 0.168
\end{aligned} \tag{137}$$

where  $Y_1$  is the first element (value) in the column matrix  $Y$ ,  $Y_2$  is the second element, etc.

Notice that the estimated likelihood value for  $Y_2$  (i.e.,  $Y_{\text{miss}}$  in Figure 3) would change depending on which imputed value was used (e.g., 19.4065, 19.4135, etc.). Recall that the maximum likelihood value is associated with the  $Y_i$  value that is closest to  $\mu_Y$  given the other values and  $\sigma_Y^2 = 2.0$ . Likewise, the smallest log-likelihood value is considered the best estimate. Now consider the sum of the log-likelihood (i.e., a measure of overall sample fit to  $\mu$ ; see appendix QQQ) given each estimate of  $Y_{\text{miss}}$  for the first six iterations:

$$\begin{aligned}
\Sigma(FML_{\text{iteration1}}) &= -14.5797815 \\
\Sigma(FML_{\text{iteration2}}) &= -14.5794679 \\
\Sigma(FML_{\text{iteration3}}) &= -14.5794253 \\
\Sigma(FML_{\text{iteration4}}) &= -14.5794193 \\
\Sigma(FML_{\text{iteration5}}) &= -14.5794184 \\
\Sigma(FML_{\text{iteration6}}) &= -14.5794184
\end{aligned} \tag{138}$$

Note that the largest change occurred between the first and second iteration and each subsequent iteration changed less until the change was negligible (e.g., beyond the seventh decimal place). In this illustration, the EM algorithm would have converged in the sixth iteration provided the convergence criteria were  $1.0e^{-7}$ .

## Appendix F: Discussion of FIML as a Direct ML Method

Consider how Finkelstein (1979) applied FIML to the factor analysis model. Note that the multivariate normal likelihood function can be expressed as  $L = \text{likelihood function} = \text{function}(\text{data}, \text{parameters})$  or more compactly as:

$$L = f(\mathbf{S}, \mathbf{\Sigma}) \quad (139)$$

where  $\mathbf{S}$  is the sample covariance and  $\mathbf{\Sigma}$  is the population covariance matrix. From this expression we can derive another equation that defines  $L$  as a function of the sample covariance matrix  $\mathbf{S}$  (as before) and the *model-implied* population covariance matrix  $\mathbf{\Sigma}$  generated as a function of the common factor model. That is, given that the common factor model for a specified number of factors holds in the population the model implied covariance matrix can be generated from the expression:  $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\psi}$  where  $\mathbf{\Lambda}$  represents factor loadings and  $\mathbf{D}_{\psi}$  indicates unique variances. Thus, the likelihood function can be re-expressed as:

$$L = f\left(\mathbf{S}, \left[\mathbf{\Lambda}, \mathbf{D}_{\psi}\right]\right) \quad (140)$$

That is, the value of the likelihood function is dependent on the sample covariance matrix (which is known) and the factor loadings and unique variances (which are not known). The goal is to find best estimates of these population parameters. So, maximum likelihood estimates,  $\mathbf{\Lambda}$  and  $\mathbf{D}_{\psi}$  are chosen so that the likelihood function  $L = f(\mathbf{S}, \mathbf{\Sigma})$  is maximized. Note that the mean vectors were excluded from the previous formula for ease of presentation; however, they can be easily incorporated.

The computational procedure for ML estimation works in general as follows: given  $\mathbf{S}$ , the parameters  $\mathbf{\Lambda}$  and  $\mathbf{D}_{\psi}$  are estimated. From these estimates, a re-constructed covariance matrix  $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\psi}$  is then generated. Afterwards, a value of the likelihood function,  $L = f(\mathbf{S}, \mathbf{\Sigma})$  is determined. Said differently, given  $\mathbf{S}$  the task is to find  $\mathbf{\Lambda}$  and  $\mathbf{D}_{\psi}$  so that the resulting  $\mathbf{\Sigma}$  yields

the largest possible value of  $L = f(\mathbf{S}, \mathbf{\Sigma})$ . Therefore, the MLE estimation routine is directed to locate values for  $\mathbf{\Lambda}$  and  $\mathbf{D}_{\psi}$  so that the resulting covariance matrix  $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\psi}$  yields a minimum value of MLE. When the optimal pattern of population estimates are found the iterative optimization algorithm (i.e., sequence of a re-constructed covariance matrix  $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{D}_{\psi}$ ) has found the best fit to the data (i.e., the highest log-likelihood; see Enders, 2010). In this way, the FIML procedure eliminates the need to then export the sufficient statistics for further analyses (unlike the EM algorithm). As demonstrated above, this is accomplished by building a model-implied covariance matrix that includes all the variables that are included in the factor analysis model.

## Appendix G: Discussion of Probability and Missing Data Mechanisms

To solidify the idea of probability distributions in this context, consider the diagram in Figure 16 where  $\mathbf{Y}$  represents the sampling space for four variables  $Y_1 - Y_4$ . That is, the variables  $Y_1 - Y_4$  each represent a possible event (occurrence) of  $\mathbf{Y}$  and form four partitions of the total sample space (represented by diagonal lines in Figure 16). For this example, let each patrician (variable) be sampled at random with a .25 probability of being selected. That is, the probability of  $Y_i$ , denoted  $p(Y_i)$ , can be thought of as the frequency of drawing  $Y_i$  (see Hays, 1994). Now, let  $\mathbf{R}$  be the occurrence of another event (i.e., the presence of missing data) which is depicted as a circle in Figure 16. Let 15% of each randomly selected variable have missing data, so that there is a .15 probability of the event “missing”.

Given the values of  $p(Y_1)$ ,  $p(Y_2)$ ,  $p(Y_3)$ ,  $p(Y_4)$ ,  $p(\text{missing}; \mathbf{R} = 1)$ , and  $p(\text{not missing}; \mathbf{R} = 0)$  it is possible to determine the conditional probability of observing missing data on the  $i^{\text{th}}$  variable  $Y_i$ ; however, the probability of each joint event (i.e., the occurrence of both  $Y_i$  and  $\mathbf{R}$ ) is dependent upon the independence of the events (Hays, 1994). Said differently, if the events (probability distributions) are independent then missingness (denoted  $\mathbf{R}$ ) has absolutely nothing to do with the variables  $Y_1 - Y_4$ . Otherwise,  $\mathbf{R}$  is not simply a random sample of  $\mathbf{Y}$ .

It is productive to consider an illustrative example of this idea. Consider Panel A of Figure 54 where the two conditions,  $Y_i$  and  $\mathbf{R}$ , are independent. That is, the probability  $p(Y_i \text{ and } [\mathbf{R}=1])$  is equal to the product of the two marginal probabilities (.25 and .15, respectively). Likewise, the probabilities for any of the other cells may be found by multiplying the associated marginal probabilities (see Hays, 1994). Because these events are independent the probability of missing data given any particular variable is as follows:

$$\begin{aligned}
p(R=1|Y_1) &= 0.15 \\
p(R=1|Y_2) &= 0.15 \\
p(R=1|Y_3) &= 0.15 \\
p(R=1|Y_4) &= 0.15
\end{aligned} \tag{141}$$

Hence, if the event  $Y_i$  is independent of the event “missing data” the probability of the joint event is found by  $p(Y_i \text{ and “missing data”}) = p(Y_i) \times p(R=1) = .25 \times .15 = .034$ . So based on these probabilities, the variable  $Y_i$  is drawn with missing data in about 34 of 1000 random selections (see Figure 54). Notice that the probability of missingness given  $Y_i$  is constant and this value is equal to the original probability of missingness, denoted  $p(R=1) = 0.15$  in Figure 54. Visually, the sampling space  $\mathbf{Y}$  in Figure 16 shrinks to  $\mathbf{R}$  as the variable  $Y_i$  is completely unrelated. That is, ignoring the diagonal partitions for  $Y_i$  in the circle  $\mathbf{R}$  leads to the probability of  $\mathbf{R}$  (see Figure 54) which can be expressed mathematically by:

$$p(\mathbf{R} | \mathbf{Y}) = p(\mathbf{R}) = 0.15 \tag{142}$$

where the probability of missingness is unrelated to the data. Now consider Panel B of Figure 54 where the two conditions  $Y_i$  and  $\mathbf{R}$  are *not* independent. That is, the probability  $p(Y_i \text{ and } [R=1])$  is not equal to the product of the two marginal probabilities such that:

$$\begin{aligned}
p(R=1|Y_1) &= 0.24 \\
p(R=1|Y_2) &= 0.08 \\
p(R=1|Y_3) &= 0.12 \\
p(R=1|Y_4) &= 0.16
\end{aligned} \tag{143}$$

Notice that the probability of missingness is dependent on  $Y_i$  and is not equal to the original probability of missingness. That is,  $\mathbf{R}$  is not a random sample of  $\mathbf{Y}$ , so the diagonal partitions for  $Y_i$  in the circle  $\mathbf{R}$  remain and directly relate to the probability of missingness (see Figure 54).

Therefore, the sample space for  $\mathbf{R}$  is associated with  $\mathbf{Y}$  such that:

$$p(\mathbf{R} | \mathbf{Y}) = p(\mathbf{R} | \mathbf{Y}) \tag{144}$$

Said differently, the probability of missingness is dependent on the data. The idea that the data provide information about the probability of missingness (i.e., causal relationships between variables and  $\mathbf{R}$ ) is a key concept behind Rubin's (1976) theory.

## Appendix H: Demonstration of data augmentation in MI

To demonstrate the data augmentation algorithm using the linear model equation, first consider a simple regression equation (for the  $i^{\text{th}}$  observation, prediction of  $Y_i$  by  $k$  variables in  $X_{ik}$ ) which can be written as:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + e_i \quad (145)$$

where  $\beta_0$  is the intercept,  $\beta_1$  signifies a slope with a predictor  $X$  and an associated error term  $e$ .

Now let the equation above be expressed more compactly by utilizing a set of matrices (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (146)$$

where  $\mathbf{Y}$  is an  $(N \times 1)$  vector that contains the dependent variable  $y$ ,  $\mathbf{X}$  is an  $(N \times (k - 1))$  design matrix that contains a column of 1s and  $k - 1$  columns of predictors,  $\mathbf{b}$  is a  $(k \times 1)$  vector of beta weights, and  $\mathbf{e}$  is an  $(N \times 1)$  residual vector. This formula can be explicitly illustrated as:

$$\begin{aligned} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{N1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \\ \mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e} \quad (147) \\ (N \times 1) & \quad (N \times (k-1))(k \times 1)(N \times 1) \end{aligned}$$

where  $\mathbf{e} \sim N_N(0, \sigma_e^2 \mathbf{I}_N)$  and the first case is  $Y_1 = \beta_0 + \beta_1 X_{11} + e_1$ . Given the previous matrix equation, assume missing data on  $\mathbf{Y}$ . More specifically, let  $Y_1, \dots, Y_m$  represent observed values ( $Y_{\text{obs}}$ ), while  $Y_{m+1}, \dots, Y_p$  indicate missing values ( $Y_{\text{miss}}$ ). Further, consider that  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  can be rewritten to align  $Y_{\text{obs}}$  with  $X_{\text{obs}}$  and  $Y_{\text{miss}}$  with  $X_{\text{miss}}$  such that the matrix expression resembles:

$$\begin{aligned}
\begin{bmatrix} Y_1 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_p \end{bmatrix} &= \begin{bmatrix} 1 & X_{11} \\ 1 & \vdots \\ 1 & X_{m1} \\ 1 & X_{m+1,1} \\ \vdots & \vdots \\ 1 & X_{p1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_m \\ e_{m+1} \\ \vdots \\ e_p \end{bmatrix} & (148) \\
\mathbf{Y} &= \mathbf{X} \mathbf{b} + \mathbf{e}
\end{aligned}$$

Now, the I-step can be directly implemented to estimate  $Y_{\text{miss}}$  via the least squares criterion by minimizing the sum of square error terms such that (see Kutner, Nachtsheim, Neter, and Li, 2005):

$$\sum_{i=1}^N e^2 = \sum_{i=1}^N (Y_{\text{obs},i} - \mathbf{x}'_{\text{obs},i} \mathbf{b})^2 = (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b})' (\mathbf{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \mathbf{b}) = \mathbf{e}' \mathbf{e} \quad (149)$$

which can be written in relation to  $\mathbf{b}$  that minimizes  $\mathbf{e}' \mathbf{e}$  as:

$$\mathbf{b} = (\mathbf{X}'_{\text{obs}} \mathbf{X}_{\text{obs}})^{-1} \mathbf{X}'_{\text{obs}} \mathbf{Y}_{\text{obs}} \quad (150)$$

where  $\mathbf{b}$  is a  $((k - 1) \times 1)$  vector of the least-squares beta estimates that minimizes the sum of squared residuals of a  $(m \times 1)$  vector of  $\mathbf{Y}_{\text{obs}}$  (note that this is the mean of  $y$  without predictors in the model) based on the associated  $(m \times 1)$  vector of  $\mathbf{X}_{\text{obs}}$ . Once  $\mathbf{b}$  is obtained it is then used to predict the missing  $Y_{m+1}, \dots, Y_p$  values from the observed  $X_{m+1}, \dots, X_p$  values. To illustrate this process, consider the example data in Figure 26 for  $\mathbf{Y}$  and  $\mathbf{X}$ , respectively. Given the observed data (i.e.,  $\mathbf{Y}_{\text{obs}}$ ,  $\mathbf{X}_{\text{obs}}$ ), the next step is to apply the regression formula, which can be demonstrated as:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\
\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \left( \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \\
&= \begin{bmatrix} N & \sum x \\ \sum x & \sum x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} = \frac{1}{|\mathbf{X}'\mathbf{X}|} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & N \end{bmatrix} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} \\
&= \frac{1}{N\sum x^2 - (\sum x)^2} \begin{bmatrix} \sum x^2 \sum y - \sum x \sum xy \\ N\sum xy - \sum x \sum y \end{bmatrix} = \begin{bmatrix} \frac{\sum x^2 \sum y - \sum x \sum xy}{N\sum x^2 - (\sum x)^2} \\ \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\frac{1}{N^2} (\sum x^2 \sum y - \sum x \sum xy)}{\frac{1}{N^2} (N\sum x^2 - (\sum x)^2)} \\ \frac{\frac{1}{N} (\sum xy - \frac{1}{N} \sum x \sum y)}{\frac{1}{N} (\sum x^2 - \frac{1}{N} (\sum x)^2)} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ \frac{\text{cov}(x, y)}{\text{var}(x)} \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix} \\
&= \begin{bmatrix} 18.9971 - 19.5650 \frac{1.9687}{3.1499} \\ \frac{1.9687}{3.1499} \end{bmatrix} = \begin{bmatrix} 6.7689 \\ 0.6250 \end{bmatrix} \tag{151}
\end{aligned}$$

where  $\Sigma$  indicates summation,  $N$  is the observed sample size, “cov” denotes the covariance, “var” represents the variance,  $b_0$  is the intercept parameter and  $b_1$  is the slope parameter (see Kutner, Nachtsheim, Neter, and Li, 2005). Note that calculations for the intercept and slope regression parameters are illustrated simultaneously in the above equation.

The next step is to predict  $Y_{\text{miss}}$  (a missing value) from  $X_{\text{miss}}$  (i.e., 20.22; an observed value in  $X$  that corresponds to a missing value in  $Y$ ; see Figure 26) using the regression weights observed in the vector  $\mathbf{b}$ . The subsequent equation can be illustrated as:

$$\begin{aligned}
Y_{\text{miss}} &= 6.7689 + .6250(20.22) \\
Y_{\text{miss}} &= 19.4065 \tag{152}
\end{aligned}$$

To highlight the least squares criterion in this imputation method, notice that imputing 19.4065 for  $Y_{\text{miss}}$  and estimating  $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$  produces the following set of residuals (see  $\mathbf{e}$  vector below):

$$\mathbf{Y} = \mathbf{X} \mathbf{b} + \mathbf{e} \quad (153)$$

$$\begin{bmatrix} 21.15 \\ 19.41 \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} = \begin{bmatrix} 1 & 21.14 \\ 1 & 20.22 \\ 1 & 20.81 \\ 1 & 20.69 \\ 1 & 19.59 \\ 1 & 18.33 \\ 1 & 15.40 \\ 1 & 20.34 \end{bmatrix} \begin{bmatrix} 6.769 \\ 0.625 \end{bmatrix} + \begin{bmatrix} 1.22 \\ 0.00 \\ 0.66 \\ 0.28 \\ -1.85 \\ 0.93 \\ 0.18 \\ -1.42 \end{bmatrix}$$

where the imputed value ( $Y_{\text{miss}}$ ) has a residual of zero. This residual indicates that the imputed value lies directly on a regression line (i.e., there is no variability associated with the predicted value). Enders (2010) notes that the I-step is needed to add variability by infusing a normally distributed residual term ( $z_i$ ; random noise) into the predicted value. Importantly, the I-step uses  $Y_{\text{miss}} = (19.4065 + z_i)$  in the  $\sum y$  and  $\sum xy$  calculations; however, the I-step does not literally fill in the missing value; rather this value is only referenced by the following P-step (see Enders, 2010). The ensuing P-step uses the imputed data from the prior I-step to generate a posterior distribution of the mean vector and covariance matrix as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 3.1499 & 1.7561 \\ 1.7561 & 2.0238 \end{bmatrix} \quad (154)$$

$$\mathbf{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} = \begin{bmatrix} 19.5650 \\ 19.0483 \end{bmatrix}$$

where  $\mathbf{\mu}$  is a  $2 \times 1$  column vector of means for  $\mathbf{X}$  and  $\mathbf{Y}$ , and  $\mathbf{\Sigma}$  is a  $2 \times 2$  covariance matrix. Note that this illustration does not actually add random variability  $z_i$  to the estimated value  $Y_{\text{miss}}$  for the purpose of clarity in moving through this example. That is, in reality the values reported in the mean vector and covariance matrix above would slightly differ (due to  $z_i$ ) from the values demonstrated.

During the next step, Monte Carlo simulation is used to generate new parameter values for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  that differ randomly from the parameters used in the initial I-step (see Enders, 2010).

These parameter estimates are then carried to the next I-Step as follows:

$$\begin{aligned}
 \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \begin{bmatrix} \frac{1}{N^2} (\sum x^2 \sum y - \sum x \sum xy) \\ \frac{1}{N^2} (N \sum x^2 - (\sum x)^2) \\ \frac{1}{N} (\sum xy - \frac{1}{N} \sum x \sum y) \\ \frac{1}{N} (\sum x^2 - \frac{1}{N} (\sum x)^2) \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\text{cov}(x, y)}{\text{var}(x)} \\ \frac{\text{cov}(x, y)}{\text{var}(x)} \end{bmatrix} = \begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix} \\
 &= \begin{bmatrix} 19.0483 - 19.5650 \frac{1.7561}{3.1498} \\ \frac{1.7561}{3.1498} \end{bmatrix} = \begin{bmatrix} 8.1404 \\ 0.5575 \end{bmatrix} \quad (155)
 \end{aligned}$$

Again the procedure uses this information to predict  $Y_{\text{miss}}$  (a missing value) from  $X_{\text{miss}}$  (i.e., 20.22; an observed value in  $X$  that corresponds to a missing value in  $Y$ ) using the regression weights observed in  $\mathbf{b}$ . The subsequent equation can be illustrated as:

$$\begin{aligned}
 Y_{\text{miss}} &= 8.1404 + .5575(20.22) \\
 Y_{\text{miss}} &= 19.4135 \quad (156)
 \end{aligned}$$

The following M-step again uses the estimate  $Y_{\text{miss}} = (19.4135 + z_i)$  in the calculation of another set of sufficient statistics and the process repeats a large number of times. Unlike the EM algorithm, this procedure does not converge on a set of most likely values (as each of these iterations involves a random perturbation). Rather, each  $m$  imputed dataset is saved after a given set of I-step to P-step cycles. For example, suppose that the  $m = 1$  imputed data set is the imputed data at the 200<sup>th</sup> P-step, and the  $m = 2$  dataset is the imputed data at the 300<sup>th</sup> P-step, etc. (i.e., the default criteria in SAS, for example, is 200 burn-in iterations before the first imputation and 100 iterations between imputations; see SAS Institute Inc., 2011 for details).

## Appendix I: Syntax Guide to Using PCA Auxiliary Variables

### Preliminaries

This appendix outlines a method of using principal component analysis (PCA) to obtain auxiliary variables for *large data* sets. Typically, auxiliary variables are not the focus of an analysis but are instead used to inform the missing data handling procedure (e.g., FIML, MI) by adding information about why a particular variable has missing data or describing the probability that a particular case has missing data. Therefore, auxiliary variables support the missing at random assumption and are used to reduce bias and decrease standard errors (Collins, Schafer, & Kam, 2001; Enders & Peugh, 2004; Graham, 2003).

Researchers suggest an “inclusive strategy” where as many auxiliary variables are included as possible (see Enders, 2010). However, it is often difficult to include large numbers of auxiliary variables without having estimation problems; especially when the auxiliary variables also have missing data. Large data projects can present a challenge because it is possible to have hundreds of potential auxiliary variables to choose from.

The current method uses PCA to reduce the dimensionality of the data set. A new set of uncorrelated variables are created (e.g., principal components) that contain all the useful information in the original data set. These principal components are then used to inform the missing data handling procedure. This guide illustrates the application of these PCA auxiliary variables to both FIML and MI methods.

### Step One: Obtain principal components from the data set

#### *1) Perform a single imputation for the entire data set*

*The SAS Proc MI syntax will look something like this:*

```
Proc mi data=sample out=outmi nimpute=1 seed=461981;  
em maxiter=1000;  
mcmc chain=multiple initial=em (maxiter = 1000);  
run;
```

\*For Proc mi options see: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>;

*At this step the data are all imputed a single time.*

a) Note. Like other multivariate methods, PCA requires complete data.

## 2) *Impute at the level of the aggregate scales.*

*The SAS Proc MI syntax will look something like this:*

```
Proc mi data=sample out=outmi nimpute=1 seed=761253;  
em maxiter=1000;  
mcmc chain=multiple initial=em (maxiter = 1000);  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ;  
run;
```

\*For Proc mi options see: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>;

*At this step the scales now have NO missing data but the items still contain missing data*

## 3) *Use the imputed scales as anchors to impute missing data in the items in a sequential process*

*Now the key is to put the right variables into the variable list for estimating the missing data:*

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI  
j1 j2 j3 j4 j5 j6 j7 j8 j9;
```

\* At this step ScaleJ is *not* in the list but the *items* for ScaleJ are now imputed;

\* ScaleJ needs to be removed because it is a linear combination of its items;

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleJ  
i1 i2 i3 i4 i5 i6 i7 i8;
```

\* At this step ScaleI is *not* in the list but the *items* for ScaleI are now imputed;

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleI ScaleJ  
h1 h2 h3 h4 h5 h6 h7 h8 h9 h10;
```

\* At this step ScaleH is *not* in the list but the *items* for ScaleH are now imputed;

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleH ScaleI ScaleJ  
g1 g2 g3 g4 g5 g6 g7 g8 g9 g10;
```

\* At this step ScaleG is *not* in the list but the *items* for ScaleG are now imputed;

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleG ScaleH ScaleI ScaleJ  
f1 f2 f3 f4 f5 f6 f7 f8 f9 f10;
```

\* At this step ScaleF is *not* in the list but the *items* for ScaleF are now imputed;

```
Proc mi data=outmi out=outmi.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleF ScaleG ScaleH ScaleI ScaleJ  
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10;
```

\* At this step ScaleE is *not* in the list but the *items* for ScaleE are now imputed;

```
Proc mi data=outmi out=outmi.....;
var ScaleA ScaleB ScaleC    ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ
    d1 d2 d3 d4 d5 d6 d7 d8 d9 d10;
* At this step ScaleD is not in the list but the items for ScaleD are now imputed;
```

```
Proc mi data=outmi out=outmi.....;
var ScaleA ScaleB    ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ
    c1 c2 c3 c4 c5 c6 c7 c8 c9 c10;
* At this step ScaleC is not in the list but the items for ScaleC are now imputed;
```

```
Proc mi data=outmi out=outmi.....;
var ScaleA    ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ
    b1 b2 b3 b4 b5 b6 b7 b8 b9 b10;
* At this step ScaleB is not in the list but the items for ScaleB are now imputed;
```

```
Proc mi data=outmi out=outmi.....;
var    ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ
    a1 a2 a3 a4 a5 a6 a7 a8 a9 a10;
* At this step ScaleA is not in the list but the items for ScaleA are now imputed;
    At this point ALL of the items are imputed.
```

## Method Two:

### *1) Do a principal components analysis of the items on the dataset and output the meaningful components as component scores.*

*The SAS Proc Princomp syntax ( Proc Factor may also be used) will look something like this:*

```
Proc princomp data=sample out=Prin;
var a1-a10 b1-b10 c1-c10 d1-d10 e1-e10 f1-f10 g1-g10 h1-h10 i1-i8 j1-j9;
run;
*For Proc princomp options: http://support.sas.com/onlinedoc/913/docMainpage.jsp;
```

In this example, let's assume that the first 12 components accounted for a meaningful amount of variance and any more components would not yield much more reliable information.

### *2) Estimate any missing data on the component scores*

*The SAS Proc MI syntax will look something like this:*

```
Proc mi data=Prin out=outmi nimpute=1 seed=761253;
em maxiter=1000;
mcmc chain=multiple initial=em (maxiter = 1000);
var Component1-Component12;
run;
*For Proc mi options see: http://support.sas.com/onlinedoc/913/docMainpage.jsp;
```

*At this step, the **component scores** now have **NO** missing data but the **items** still contain*

*missingness.*

**3) Use the estimated component scores as anchors to estimate the missing data in the items on the dataset.**

**Proc mi** data=outmi out=outmi.....;  
var Component1-Component12 a1-a10 b1-b10;

\* At this step we plug approximately 20 items into the variable list to estimate their missing values;

**Proc mi** data=outmi out=outmi.....;  
var Component1-Component12 c1-c10 d1-d10;

\* At this step we plug in approximately 20 more items into the list to estimate their missing values;

**Proc mi** data=outmi out=outmi.....;  
var Component1-Component12 e1-e10 f1-f10;

\* At this step we plug in approximately 20 more items into the list to estimate their missing values;

**Proc mi** data=outmi out=outmi.....;  
var Component1-Component12 g1-g10 h1-h10;

\* At this step we plug in approximately 20 more items into the list to estimate their missing values;

**Proc mi** data=outmi out=outmi.....;  
var Component1-Component12 i1-i8 j1-j9;

\* At this step we plug the remaining items to estimate their missing values;

*At this point ALL of the items are imputed.*

**What about Multiple Imputations:**

**1) Change the number of imputations option (nimpute= ) in the Proc MI syntax**

*The SAS Proc MI syntax will look something like this:*

**Proc mi** data=outmi out=outmi nimpute=**20**.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI  
j1 j2 j3 j4 j5 j6 j7 j8 j9;

\* At this step ScaleJ is *not* in the list but the *items* for ScaleJ are now imputed;

**Proc mi** data=outmi out=outmi nimpute=**20**.....;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleJ  
i1 i2 i3 i4 i5 i6 i7 i8;

\* At this step ScaleI is *not* in the list but the *items* for ScaleI are now imputed;

*Notes.* Multiple imputations should be carried out during **Step 3** of the previous two

examples. It is not suggested to impute the initial aggregate scales or the initial component scores as in Step 2. Therefore, multiple imputations are created *only* for the imputed items.

The input and output datasets in the SAS syntax do *not* change with each imputation step; this stacks the imputed datasets on top of each other 1 – 20. If the original dataset contained 200 observations then the new data set with 20 imputations will contain 4000 observations with the 201<sup>st</sup> observation being the first observation in the 2<sup>nd</sup> imputation. SAS will create a new variable called “\_imputation\_” with a value of 1 for the first imputation, 2 for the second imputation, and so on to a value of 20 for the last imputation.

*At this point ALL of the **items** are imputed 20 times.*

For additional information regarding further analysis of a dataset with multiple imputations in LISREL see the Rubin’s Rules Pooler for LISREL Estimates:

<http://www.quant.ku.edu/resources/tools.html>

### **Helpful SAS code for the aggregate scale method:**

#### *a) Creating a scale based on item means.*

*The SAS Proc Standard syntax will look something like this:*

```
Data sampleNew ; set sample ;  
ScaleA = mean(of a1 a2 a3 a4 a5 a6 a7 a8 a9 a10);  
ScaleB = mean(of b1 b2 b3 b4 b5 b6 b7 b8 b9 b10);  
... ;  
run;
```

#### *b) Standardize before imputing (when a scale is created from item SUM).*

*The SAS Proc Standard syntax will look something like this:*

```
Proc standard data=sample mean=0 std=1 out=zsampl;  
var ScaleA ScaleB ScaleC ScaleD ScaleE ScaleF ScaleG ScaleH ScaleI ScaleJ;  
run;
```

Table 1

*Data from an exemplar randomized block design taken from Allan and Wishart (1930).*

Block	Treatments									Totals
	1	2	3	4	5	6	7	8	9	
A	15.96	16.80	19.08	21.15	21.14	23.02	20.60	26.34	25.32	189.41
B	21.96	24.44	26.33	—	20.22	20.87	19.21	20.04	21.30	174.37
C	18.78	18.36	19.40	20.40	20.81	23.77	17.24	18.00	20.46	177.22
D	20.18	20.52	20.42	19.96	20.69	20.90	21.68	21.90	20.68	186.93
E	16.20	16.90	19.81	17.21	19.59	21.22	18.72	20.04	19.02	168.71
F	18.64	20.33	22.71	19.29	18.33	19.06	20.04	21.31	22.19	181.90
G	18.89	23.60	22.42	16.91	15.40	17.00	16.71	18.02	19.84	168.79
H	19.22	20.47	20.98	18.06	20.34	21.38	16.83	15.66	17.21	170.15
Total	149.83	161.42	171.15	132.98	156.52	167.22	151.03	161.31	166.02	1417.48
Mean	18.73	20.18	21.39	19.00	19.57	20.90	18.88	20.16	20.75	19.96

*Note.* “—” denotes a missing value. The data in this example represent protein percentage in peas across nine treatments and eight replicates.

Table 2

*Iterative solutions for a set of linear equations.*

---

Iteration Number	Estimates		
	X	Y	Z
1	1.147	0.798	0.357
2	1.004	0.645	0.635
3	1.006	0.759	0.612
4	0.989	0.745	0.645
5	0.989	0.758	0.643
6	0.987	0.756	0.647
7	0.987	0.758	0.647
8	0.986	0.758	0.647
9	0.986	0.758	0.647
10	0.986	0.758	0.647

---

Table 3

*Exemplar data illustrating MCAR, MAR and MNAR missing data mechanisms.*

Case	Complete Data		MCAR Data		MAR Data		MNAR Data	
	$y_1$	$y_2$	$y_1$	$y_2$	$y_1$	$y_2$	$y_1$	$y_2$
1	-1.36	-0.39	-1.36	—	-1.36	—	-1.36	—
2	0.97	0.91	0.97	—	0.97	0.91	0.97	0.91
3	-0.58	-0.68	-0.58	-0.68	-0.58	—	-0.58	—
4	-2.20	-1.28	-2.20	—	-2.20	—	-2.20	—
5	-1.72	-1.34	-1.72	—	-1.72	—	-1.72	—
6	-1.52	-0.06	-1.52	—	-1.52	—	-1.52	—
7	-0.11	-0.11	-0.11	-0.11	-0.11	—	-0.11	—
8	-0.51	0.11	-0.51	—	-0.51	—	-0.51	0.11
9	-0.18	0.22	-0.18	—	-0.18	—	-0.18	0.22
10	0.51	0.67	0.51	—	0.51	0.67	0.51	0.67
11	0.74	-1.13	0.74	-1.13	0.74	-1.13	0.74	—
12	0.27	-0.34	0.27	—	0.27	-0.34	0.27	—
13	0.44	1.04	0.44	1.04	0.44	1.04	0.44	1.04
14	1.34	1.58	1.34	1.58	1.34	1.58	1.34	1.58
15	-0.02	0.18	-0.02	—	-0.02	—	-0.02	0.18
16	0.92	0.60	0.92	0.60	0.92	0.60	0.92	0.60
17	-1.73	-1.57	-1.73	-1.57	-1.73	—	-1.73	—
18	-0.14	0.02	-0.14	0.02	-0.14	—	-0.14	—
19	0.39	1.64	0.39	—	0.39	1.64	0.39	1.64
20	-0.33	-1.57	-0.33	-1.57	-0.33	—	-0.33	—
21	0.60	0.07	0.60	—	0.60	0.07	0.60	—
22	0.24	0.64	0.24	—	0.24	0.64	0.24	0.64
23	0.69	-0.33	0.69	—	0.69	-0.33	0.69	—
24	0.24	0.64	0.24	0.64	0.24	0.64	0.24	0.64
25	2.60	0.44	2.60	0.44	2.60	0.44	2.60	0.44
26	2.54	1.13	2.54	1.13	2.54	1.13	2.54	1.13
27	-1.07	-1.13	-1.07	—	-1.07	—	-1.07	—
28	0.07	0.07	0.07	0.07	0.07	—	0.07	0.07
29	-0.67	-0.77	-0.67	-0.77	-0.67	—	-0.67	—
30	-0.07	-0.35	-0.07	—	-0.07	—	-0.07	—
Mean ( $\bar{y}_2$ ) =	-0.04		-0.02		0.54		0.71	
SD ( $SD_{y_2}$ ) =	0.89		1.00		0.77		0.51	
Corr ( $r_{y_1y_2}$ ) =	0.67		0.69		0.19		0.40	

*Note.* The mean ( $\bar{y}_2$ ), standard deviation ( $SD_{y_2}$ ), and correlation ( $r_{y_1y_2}$ ) are  $M = -0.04$ ,  $SD = .89$ ,  $r = .67$  among the complete data example; are  $M = -0.02$ ,  $SD = 1.00$ ,  $r = .69$  among the MCAR

data example; are  $M = 0.54$ ,  $SD = .77$ ,  $r = .19$  among the MAR data example; and are  $M = 0.71$ ,  $SD = .51$ ,  $r = .40$  among the MNAR data example.

Table 4

*Simulation results showing the impact of omitting a cause or correlate of missingness.*

<u>Average parameter estimates</u>			
$\rho_{A,Y}$	$\rho_{X,Y}$	$\mu_Y$	$\sigma^2_Y$
0.10	0.298	0.007	0.989
0.20	0.284	0.056	0.973
0.30	0.261	0.104	0.943
0.40	0.246	0.161	0.914
0.50	0.230	0.207	0.858
0.60	0.219	0.262	0.807
0.70	0.200	0.316	0.747
0.80	0.185	0.368	0.672
True Values	0.300	0.000	1.000

*Note.*  $\rho_{A,Y}$  denotes the population correlation between  $Y$  and the omitted auxiliary variable;  $\rho_{X,Y}$  denotes the correlation between  $X$  and  $Y$ ;  $\mu_Y$  is the mean of  $Y$  and  $\sigma^2_Y$  is the variance of  $Y$ .

Table 5

*Simulation results showing the influence of auxiliary variables to improve parameter estimation in an MNAR situation.*

Average parameter estimates			
$\rho_{A,Y}$	$\rho_{X,Y}$	$\mu_Y$	$\sigma_Y^2$
0.10	0.156	0.473	0.490
0.20	0.160	0.467	0.492
0.30	0.161	0.452	0.496
0.40	0.162	0.434	0.498
0.50	0.169	0.412	0.505
0.60	0.185	0.376	0.521
0.70	0.191	0.354	0.546
0.80	0.220	0.312	0.567
True Values	0.300	0.000	1.000

*Note.*  $\rho_{A,Y}$  denotes the population correlation between  $Y$  and the omitted auxiliary variable;  $\rho_{X,Y}$  denotes the correlation between  $X$  and  $Y$ ;  $\mu_Y$  is the mean of  $Y$  and  $\sigma_Y^2$  is the variance of  $Y$ .

Table 6

*Simulation results showing the bias associated with ignoring a non-linear cause of missingness.*

$\rho_{A,Y}$	$\rho_{X,Y}$	Raw Bias	Std. Bias
<u>Linear only (i.e., Aux<sub>1</sub> and Aux<sub>2</sub>)</u>			
0.10	0.240	0.060	0.042
0.20	0.257	0.043	0.039
0.30	0.259	0.041	0.037
0.40	0.262	0.038	0.035
0.50	0.268	0.032	0.031
0.60	0.270	0.030	0.030
0.70	0.273	0.027	0.028
0.80	0.275	0.025	0.027
<u>Non-linear only (i.e., Aux<sub>1,2</sub>)</u>			
0.10	0.262	0.038	0.036
0.20	0.265	0.035	0.032
0.30	0.268	0.032	0.027
0.40	0.274	0.026	0.021
0.50	0.283	0.018	0.017
0.60	0.289	0.011	0.013
0.70	0.294	0.006	0.007
0.80	0.299	0.001	0.004
True Values	0.300	0.000	1.000

*Note.*  $\rho_{A,Y}$  denotes the population correlation between  $Y$  and the omitted auxiliary variable;  $\rho_{X,Y}$  denotes the correlation between  $X$  and  $Y$ ;  $\mu_Y$  is the mean of  $Y$  and  $\sigma_Y^2$  is the variance of  $Y$ .

Table 7

*Simulated data for PCA examples.*

Obs	x	y
1	0.82	0.48
2	0.99	1.26
3	1.20	0.58
4	0.61	1.00
5	-0.67	-0.48
6	-0.12	-0.12
7	-0.84	-0.83
8	1.26	0.73
9	1.16	0.94
10	0.69	0.61
11	0.16	-0.64
12	1.26	0.84
13	-1.73	-1.27
14	0.34	0.49
15	-0.28	0.13
16	-1.11	-0.59
17	-1.40	-1.09
18	-1.92	-1.55
19	0.25	0.46
20	-0.39	-1.03
21	0.79	0.37
22	0.36	-0.33
23	-0.23	0.18
24	1.13	1.26
25	-0.88	-0.32

Table 8

*Monte Carlo simulation studies published in the social sciences on the topic of missing data handling.*

Authors	Sample Size	Percent Missing	Reps	Mechanism	Evaluation Criteria
Collins, Shafer, & Kam (2001)	500	25, 50	1000	MCAR, MAR	parameter & SE bias, RMSE, CI coverage, power
Enders & Bandalos (2001)	100, 250, 500, 750	2, 5, 10, 15, 25	250	MCAR, MAR	convergence failure, parameter bias, parameter estimate efficiency, model fit
Enders & Gottschall (2011)	50, 250	5, 15, 25	1000	MCAR, MAR	parameter & SE bias
Enders & Peugh (2004)	200, 400, 600	5, 15, 25	1000	MCAR	CI coverage, model rejection rate
Enders (2001)	250, 500, 750	0, 5, 15, 25	250	MCAR, MAR	parameter & SE bias, CI coverage, model rejection rate
Enders (2002)	100, 250, 500	10, 20	500	MCAR	model rejection rate
Enders (2003)	100, 300, 500	15, 30	1000	MCAR, MAR, MNAR	parameter bias - RMSE, CI coverage
Enders(2001b)	100, 250, 400	5, 15, 25, 35	250	MCAR, MAR, MNAR	parameter bias
Gold & Bentler (2000)	100, 500	4, 8, 12, 16	100	MCAR	convergence failure, parameter bias
Graham, Olchowski & Gilreath (2007)	889, 1143, 1600, 2667, 8000	10 to 90 by 20	8000	MCAR	parameter & SE bias, power
Hedden, Woolson & Malcolm (2008)	100	10, 40	2000	MCAR, MAR, MNAR, (combo)	model rejection rate, power
Littvay (2009)	1350	0, 33, 67	1000	MCAR	parameter & SE bias, power
McDonald, Thurston, & Nelson (2000)	100, 200	20, 40	500	MCAR	parameter bias - RMSE, AE, AB
Newman (2003)	440	20, 50, 75	100	MCAR, MAR, MNAR	parameter standard error - AE, convergence failure
Roth, Switzer, & Switzer (1999)	400	20	100	MCAR	parameter standard error - AE, RMSE
Schlomer, Bauman, & Card (2010)	60	0, 10, 20, 50	N/A	MCAR, MAR	parameter bias, parameter standard error
Shin, Davidson, & Long (2009)	100, 200, 500, 1000, 2000	0, 10, 19, 27.1	1000	MCAR, MAR, MNAR	convergence failure, parameter & SE bias, CI coverage, model fit
Van Buuren, et al. (2006)	400	50	1000	MCAR, MAR (fraction missing)	parameter & SE bias, CI coverage
Wothke & Arbuckle (1996)	145, 500	0, 5, 10, 20, 30	200	MCAR, MAR	parameter & SE bias
Wothke (2000)	500	0 to 40 by 1	400	MCAR, MAR	convergence failure, parameter & SE bias
Yeh (2007)	50	35, 50, 70	500	MCAR, MAR, MNAR	parameter bias
Yoo (2009)	200, 500	10, 20	1000	MCAR, MAR, MNAR, (combo)	convergence failure, parameter & SE bias, CI coverage
Yuan (2009)	50, 100, 300, 500	5, 10, 25, 50	500	MAR, MNAR	convergence failure, parameter & SE bias, model rejection rate, power
Zhang & Walker (2008)	500, 1000, 2000	15, 30, 50	100	MCAR	parameter bias - RMSE, power

Table 9

Raw ECI key skills data in 3-month intervals from 9 – 36 months.

Age	Gestures	Vocalizations	Single Words	Multiple Words
	Mean (SE) <i>n</i>	Mean (SE) <i>n</i>	Mean (SE) <i>n</i>	Mean (SE) <i>n</i>
9	9.49 (.34) <i>n</i> = 287	12.79 (.56) <i>n</i> = 285	0.09 (.02) <i>n</i> = 285	--- <i>n</i> = 285
12	10.25 (.38) <i>n</i> = 285	22.34 (.79) <i>n</i> = 285	0.8 (.11) <i>n</i> = 285	0.03 (.02) <i>n</i> = 79
15	12.39 (.42) <i>n</i> = 295	22.74 (.77) <i>n</i> = 294	2.8 (.26) <i>n</i> = 289	0.11 (.03) <i>n</i> = 293
18	10.92 (.42) <i>n</i> = 295	23.3 (.64) <i>n</i> = 296	6.66 (.47) <i>n</i> = 289	0.56 (.08) <i>n</i> = 288
21	10.16 (.39) <i>n</i> = 275	21.55 (.77) <i>n</i> = 276	12.92 (.69) <i>n</i> = 276	2.22 (.24) <i>n</i> = 269
24	10.17 (.42) <i>n</i> = 258	19.14 (.75) <i>n</i> = 257	17.87 (.80) <i>n</i> = 256	5.91 (.49) <i>n</i> = 250
27	9.71 (.43) <i>n</i> = 256	16.74 (.77) <i>n</i> = 258	20.51 (.78) <i>n</i> = 258	13.57 (.86) <i>n</i> = 259
30	8.86 (.42) <i>n</i> = 224	15.19 (.82) <i>n</i> = 222	20.7 (.74) <i>n</i> = 223	18.18 (.99) <i>n</i> = 225
33	8.75 (.44) <i>n</i> = 224	12.82 (.71) <i>n</i> = 223	20.67 (.72) <i>n</i> = 225	20.99 (1.03) <i>n</i> = 226
36	7.4 (.53) <i>n</i> = 127	11.78 (.77) <i>n</i> = 127	19.59 (.94) <i>n</i> = 127	25.67 (1.54) <i>n</i> = 128

Note. Age reflects average months since birth; *SE* represents the standard error; *n* denotes the sample size associated with each mean and standard deviation. Multiple Words were not assessed at 9 months due to developmental limitations.

Table 10

Monte Carlo simulation results showing the impact of  $PCA_{AUX}$  and  $AUX$  on parameter estimation raw bias, standardized bias, and percent bias across various sample sizes when missingness is related to a 60% MAR missing data rate and a population association of  $\rho = .60$  among the auxiliary variables.

Mechanism Parameter	No Auxiliary Variables			All Possible Auxiliary Variables			PCA Auxiliary Variable						
	Est.	Raw Bias	Std. Bias	Est.	Raw Bias	Std. Bias	Est.	Raw Bias	Std. Bias	Percent Bias			
<i>N = 75</i>													
MAR													
Y with X ( $\rho_{X,Y}$ )	0.300	0.182	-0.118	-0.085	-39.335%	0.297	-0.003	-0.003	-1.010%	0.300	0.000	0.000	-0.161%
Mean Y ( $\mu_Y$ )	0.000	0.522	0.522	0.381		0.006	0.006	0.005		0.005	0.005	0.005	
Variance Y ( $\sigma^2_Y$ )	1.000	0.725	-0.275	-0.168	-27.477%	0.996	-0.004	-0.002	-0.383%	0.981	-0.002	-0.012	-0.188%
<i>N = 200</i>													
MAR													
Y with X ( $\rho_{X,Y}$ )	0.300	0.186	-0.114	-0.081	-38.001%	0.297	-0.003	-0.003	-0.975%	0.300	0.000	0.000	-0.059%
Mean Y ( $\mu_Y$ )	0.000	0.522	0.522	0.372		-0.001	-0.001	0.000		0.005	0.005	0.005	
Variance Y ( $\sigma^2_Y$ )	1.000	0.747	-0.253	-0.151	-25.278%	0.999	-0.001	0.000	-0.064%	0.981	-0.002	-0.012	-0.168%
<i>N = 800</i>													
MAR													
Y with X ( $\rho_{X,Y}$ )	0.300	0.188	-0.112	-0.080	-37.387%	0.299	-0.001	-0.001	-0.309%	0.300	0.000	0.000	-0.026%
Mean Y ( $\mu_Y$ )	0.000	0.525	0.525	0.372		0.003	0.003	0.002		0.003	0.003	0.002	
Variance Y ( $\sigma^2_Y$ )	1.000	0.752	-0.248	-0.147	-24.787%	0.996	-0.004	-0.003	-0.445%	0.996	-0.004	-0.003	-0.413%
<i>N = 1000</i>													
MAR													
Y with X ( $\rho_{X,Y}$ )	0.300	0.188	-0.112	-0.080	-37.337%	0.299	-0.001	-0.001	-0.234%	0.300	0.000	0.000	-0.001%
Mean Y ( $\mu_Y$ )	0.000	0.524	0.524	0.372		0.002	0.002	0.002		0.000	0.000	0.000	
Variance Y ( $\sigma^2_Y$ )	1.000	0.753	-0.247	-0.146	-24.721%	0.996	-0.004	-0.002	-0.356%	1.000	0.000	0.000	-0.004%

Table 11

Monte Carlo simulation results showing the impact of  $PCA_{AUX}$  and AUX on parameter estimation raw bias, standardized bias, and percent bias across various sample sizes when missingness is related to a non-linear process with a 60% MAR missing data rate and a population association of  $\rho = .60$  among the auxiliary variables.

Mechanism Parameter	No Auxiliary Variables						All Possible Auxiliary Variables						PCA Auxiliary Variable												
	Est.		Raw Bias		Std. Bias		Percent Bias		Est.		Raw Bias		Std. Bias		Percent Bias		Est.		Raw Bias		Std. Bias		Percent Bias		
<i>N = 75</i>																									
Non-linear MAR																									
Y with X ( $\rho_{X,Y}$ )	0.300	0.523	0.223	0.123	0.123	74.274%	0.334	0.034	0.024	0.024	11.486%	0.295	-0.005	-0.005	-0.005	-1.695%	0.000	0.003	0.003	0.003	-0.001	-0.001	-0.001	-0.001	-0.001
Mean Y ( $\mu_Y$ )	1.000	1.549	0.549	0.163	0.163	54.853%	1.218	0.218	0.099	0.099	21.843%	0.992	-0.008	-0.008	-0.005	-0.794%	0.000	0.003	0.003	0.003	-0.001	-0.001	-0.001	-0.001	-0.001
Variance Y ( $\sigma^2_Y$ )	1.000	1.549	0.549	0.163	0.163	54.853%	1.218	0.218	0.099	0.099	21.843%	0.992	-0.008	-0.008	-0.005	-0.794%	0.000	0.003	0.003	0.003	-0.001	-0.001	-0.001	-0.001	-0.001
<i>N = 200</i>																									
Non-linear MAR																									
Y with X ( $\rho_{X,Y}$ )	0.300	0.521	0.221	0.121	0.121	73.584%	0.324	0.024	0.017	0.017	7.998%	0.301	0.001	0.001	0.001	0.395%	0.000	0.002	0.002	0.002	0.000	0.000	0.000	0.000	0.000
Mean Y ( $\mu_Y$ )	1.000	1.549	0.549	0.163	0.163	54.853%	1.169	0.169	0.080	0.080	16.907%	0.996	-0.004	-0.004	-0.003	-0.392%	0.000	0.002	0.002	0.002	0.000	0.000	0.000	0.000	0.000
Variance Y ( $\sigma^2_Y$ )	1.000	1.549	0.549	0.163	0.163	54.853%	1.169	0.169	0.080	0.080	16.907%	0.996	-0.004	-0.004	-0.003	-0.392%	0.000	0.002	0.002	0.002	0.000	0.000	0.000	0.000	0.000
<i>N = 800</i>																									
Non-linear MAR																									
Y with X ( $\rho_{X,Y}$ )	0.300	0.520	0.220	0.121	0.121	73.296%	0.308	0.008	0.006	0.006	2.728%	0.300	0.000	0.000	0.000	-0.118%	0.000	-0.002	-0.002	-0.001	0.002	0.002	0.002	0.002	0.002
Mean Y ( $\mu_Y$ )	1.000	1.529	0.529	0.159	0.159	52.910%	0.966	-0.034	-0.016	-0.016	-3.397%	0.999	-0.001	-0.001	-0.001	-0.091%	0.000	-0.002	-0.002	-0.001	0.002	0.002	0.002	0.002	0.002
Variance Y ( $\sigma^2_Y$ )	1.000	1.529	0.529	0.159	0.159	52.910%	0.966	-0.034	-0.016	-0.016	-3.397%	0.999	-0.001	-0.001	-0.001	-0.091%	0.000	-0.002	-0.002	-0.001	0.002	0.002	0.002	0.002	0.002
<i>N = 1000</i>																									
Non-linear MAR																									
Y with X ( $\rho_{X,Y}$ )	0.300	0.505	0.205	0.115	0.115	68.390%	0.301	0.001	0.001	0.001	0.368%	0.300	0.000	0.000	0.000	0.072%	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
Mean Y ( $\mu_Y$ )	1.000	1.496	0.496	0.152	0.152	49.562%	0.989	-0.011	-0.006	-0.006	-1.147%	0.999	-0.001	-0.001	-0.001	-0.115%	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
Variance Y ( $\sigma^2_Y$ )	1.000	1.496	0.496	0.152	0.152	49.562%	0.989	-0.011	-0.006	-0.006	-1.147%	0.999	-0.001	-0.001	-0.001	-0.115%	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000

Table 12

Monte Carlo simulation results showing the impact of  $PCA_{AUX}$  and  $AUX$  (with non-linear information) on parameter estimation raw bias, standardized bias, and percent bias across various sample sizes when missingness is related to a non-linear process with a 60% MAR missing data rate and a population association of  $\rho = .60$  among the auxiliary variables.

Mechanism Parameter	PCA Auxiliary Variable				All Possible Auxiliary Variables (INT)				
	Est.	Raw Bias	Std. Bias	Percent Bias	Est.	Raw Bias	Std. Bias	Percent Bias	
<hr/> <i>N = 75</i> <hr/>									
Non-linear MAR									
Y with X ( $\rho_{X,Y}$ )	0.300	0.295	-0.005	-0.005	-1.695%	0.290	0.010	0.080	3.367%
Mean Y ( $\mu_Y$ )	0.000	-0.001	-0.001	-0.001		-0.006	0.006	0.045	
Variance Y ( $\sigma^2_Y$ )	1.000	0.992	-0.008	-0.005	-0.794%	1.013	-0.013	-0.073	-1.340%
<hr/> <i>N = 200</i> <hr/>									
Non-linear MAR									
Y with X ( $\rho_{X,Y}$ )	0.300	0.301	0.001	0.001	0.395%	0.299	0.001	0.025	0.233%
Mean Y ( $\mu_Y$ )	0.000	0.000	0.000	0.000		-0.002	0.002	0.024	
Variance Y ( $\sigma^2_Y$ )	1.000	0.996	-0.004	-0.003	-0.392%	0.997	0.003	0.030	0.320%
<hr/> <i>N = 800</i> <hr/>									
Non-linear MAR									
Y with X ( $\rho_{X,Y}$ )	0.300	0.300	0.000	0.000	-0.118%	0.300	0.000	-0.005	-0.133%
Mean Y ( $\mu_Y$ )	0.000	0.002	0.002	0.002		0.001	-0.001	-0.020	
Variance Y ( $\sigma^2_Y$ )	1.000	0.999	-0.001	-0.001	-0.091%	0.998	0.002	0.034	0.180%
<hr/> <i>N = 1000</i> <hr/>									
Non-linear MAR									
Y with X ( $\rho_{X,Y}$ )	0.300	0.300	0.000	0.000	0.072%	0.300	0.001	0.015	0.167%
Mean Y ( $\mu_Y$ )	0.000	0.000	0.000	0.000		0.000	0.000	-0.008	
Variance Y ( $\sigma^2_Y$ )	1.000	0.999	-0.001	-0.001	-0.115%	0.998	0.002	0.043	0.200%

Table 13

*Relative Efficiency of correlation parameter estimate between X and Y by sample size, MCAR missing data rate and association strength between the variable with missingness and the auxiliary variables.*

N	Percent Missing							
	10%	20%	30%	40%	50%	60%	70%	80%
$\rho_{X,Y} = .10$								
100	1.002	1.005	1.008	1.008	1.015	1.022	1.026	1.028
400	1.001	1.001	1.001	1.000	0.999	0.999	0.999	0.999
1000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
$\rho_{X,Y} = .30$								
100	1.002	1.007	1.015	1.018	1.027	1.031	1.033	1.038
400	1.002	1.004	1.007	1.009	1.012	1.019	1.022	1.023
1000	1.001	1.004	1.006	1.008	1.011	1.016	1.019	1.021
$\rho_{X,Y} = .60$								
100	1.014	1.034	1.065	1.100	1.145	1.222	1.263	1.289
400	1.016	1.034	1.060	1.097	1.143	1.210	1.231	1.240
1000	1.015	1.026	1.037	1.096	1.117	1.155	1.193	1.206

*Note.* Values greater than 1.0 indicate that  $PCA_{AUX}$  estimates are more efficient.

Table 14

*Multiple Imputation of ECI key skills data using no auxiliary variables.*

Test Age (in months)	Gestures			Vocalizations			Single Words			Multiple Words		
	M	SE	FMI RIV	M	SE	FMI RIV	M	SE	FMI RIV	M	SE	FMI RIV
9	9.68	0.40	0.55 1.20	12.81	0.68	0.58 1.32	0.10	0.02	0.56 1.25	0.00	0.00	0.00 0.00
12	10.45	0.43	0.50 0.99	22.57	0.92	0.52 1.08	0.84	0.12	0.54 1.16	0.03	0.02	0.84 5.01
15	12.37	0.42	0.43 0.75	22.64	0.85	0.50 0.97	2.88	0.31	0.54 1.16	0.11	0.03	0.51 1.02
18	11.21	0.49	0.56 1.24	23.23	0.74	0.55 1.18	6.74	0.51	0.45 0.81	0.59	0.09	0.57 1.30
21	10.30	0.37	0.35 0.53	21.38	0.84	0.52 1.07	12.87	0.73	0.48 0.90	2.64	0.27	0.43 0.74
24	10.22	0.43	0.49 0.93	19.61	0.72	0.42 0.71	17.93	0.84	0.51 1.03	6.37	0.52	0.44 0.78
27	10.15	0.46	0.51 1.01	17.41	0.94	0.62 1.57	20.22	0.82	0.50 0.98	13.10	0.86	0.49 0.94
30	9.06	0.44	0.56 1.25	16.62	0.88	0.56 1.24	20.85	0.87	0.62 1.56	17.19	1.06	0.55 1.22
33	8.84	0.54	0.68 2.08	13.03	0.91	0.72 2.45	21.05	0.76	0.55 1.19	20.67	1.07	0.59 1.43
36	7.72	0.70	0.79 3.70	11.83	1.00	0.80 3.86	20.68	1.20	0.78 3.48	26.27	1.57	0.70 2.24

Table 15

*Multiple Imputation of ECI key skills data using the AUX method (all possible auxiliary variables).*

Test Age (in months)	Gestures				Vocalizations				Single Words				Multiple Words			
	M	SE	FMI	RIV	M	SE	FMI	RIV	M	SE	FMI	RIV	M	SE	FMI	RIV
9	9.45	0.35	0.43	0.74	12.87	0.60	0.49	0.93	0.09	0.02	0.44	0.77	0.00	0.00	0.00	0.00
12	10.32	0.39	0.41	0.67	22.36	0.78	0.38	0.60	0.76	0.10	0.33	0.49	0.04	0.02	0.79	3.59
15	12.27	0.40	0.33	0.50	22.40	0.73	0.31	0.45	2.81	0.25	0.33	0.48	0.11	0.03	0.45	0.82
18	11.12	0.42	0.39	0.62	23.19	0.63	0.38	0.60	6.75	0.44	0.34	0.51	0.59	0.08	0.37	0.59
21	10.27	0.40	0.43	0.75	21.41	0.74	0.39	0.62	13.23	0.72	0.44	0.78	2.45	0.24	0.38	0.60
24	10.23	0.41	0.41	0.69	19.72	0.77	0.49	0.94	18.13	0.74	0.38	0.60	5.88	0.47	0.43	0.74
27	10.02	0.44	0.45	0.80	16.99	0.79	0.47	0.89	20.11	0.76	0.44	0.78	13.04	0.81	0.40	0.66
30	8.90	0.39	0.44	0.76	15.77	0.84	0.52	1.06	20.53	0.77	0.54	1.15	18.04	0.90	0.40	0.67
33	8.76	0.42	0.46	0.85	12.88	0.66	0.46	0.83	20.58	0.75	0.55	1.20	20.89	0.90	0.41	0.67
36	7.91	0.54	0.71	2.38	11.48	0.79	0.70	2.30	20.22	0.87	0.64	1.72	27.15	1.47	0.67	2.02

Table 16

*Multiple Imputation of ECI key skills data using the PCA auxiliary variables.*

Test Age (in months)	Gestures				Vocalizations				Single Words				Multiple Words Words			
	M	SE	FMI	RIV	M	SE	FMI	RIV	M	SE	FMI	RIV	M	SE	FMI	RIV
9	9.51	0.21	0.18	0.22	12.63	0.33	0.10	0.11	0.10	0.01	0.14	0.16	0.00	0.00	0.00	0.00
12	10.32	0.22	0.16	0.18	22.59	0.45	0.13	0.15	0.73	0.06	0.18	0.21	0.06	0.01	0.15	0.18
15	12.13	0.24	0.18	0.21	22.73	0.45	0.18	0.22	2.76	0.15	0.16	0.19	0.09	0.02	0.16	0.18
18	11.18	0.24	0.16	0.19	23.67	0.37	0.14	0.16	6.85	0.25	0.09	0.10	0.60	0.05	0.15	0.17
21	10.26	0.22	0.17	0.21	21.26	0.42	0.16	0.18	13.07	0.39	0.17	0.20	2.52	0.14	0.13	0.15
24	10.16	0.23	0.20	0.25	19.84	0.42	0.19	0.23	18.02	0.43	0.16	0.19	5.88	0.26	0.15	0.18
27	10.07	0.25	0.22	0.29	17.10	0.43	0.18	0.22	19.99	0.42	0.15	0.17	12.82	0.46	0.15	0.17
30	8.77	0.22	0.18	0.22	16.02	0.43	0.20	0.25	20.33	0.39	0.14	0.16	17.89	0.49	0.08	0.09
33	8.97	0.24	0.21	0.26	13.02	0.38	0.23	0.29	20.51	0.38	0.16	0.19	20.49	0.50	0.12	0.13
36	8.11	0.23	0.20	0.25	11.91	0.36	0.16	0.19	20.10	0.41	0.25	0.34	26.83	0.63	0.16	0.19

Table 17

*A Demonstration of marginal means estimation for the main effect of B with complete data.*

		<b>Conditions (Factor A)</b>			
<b>Blocks</b>		$a_1$	$a_2$		
$b_1$	<b>13</b> = $\mu + \alpha_1 + \beta_1$	<b>30</b> = $\mu + \alpha_2 + \beta_1$	$\bar{b}_1 = [(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_2 + \beta_1)] / 10$		
	<b>10</b> = $\mu + \alpha_1 + \beta_1$	<b>16</b> = $\mu + \alpha_2 + \beta_1$			
	<b>9</b> = $\mu + \alpha_1 + \beta_1$	<b>12</b> = $\mu + \alpha_2 + \beta_1$			
	<b>6</b> = $\mu + \alpha_1 + \beta_1$	<b>18</b> = $\mu + \alpha_2 + \beta_1$			
	<b>12</b> = $\mu + \alpha_1 + \beta_1$	<b>24</b> = $\mu + \alpha_2 + \beta_1$			
$b_2$	<b>26</b> = $\mu + \alpha_1 + \beta_2$	<b>15</b> = $\mu + \alpha_2 + \beta_2$		$\bar{b}_2 = [(\mu + \alpha_1 + \beta_2) + (\mu + \alpha_2 + \beta_2)] / 10$	
	<b>20</b> = $\mu + \alpha_1 + \beta_2$	<b>8</b> = $\mu + \alpha_2 + \beta_2$			
	<b>18</b> = $\mu + \alpha_1 + \beta_2$	<b>6</b> = $\mu + \alpha_2 + \beta_2$			
	<b>12</b> = $\mu + \alpha_1 + \beta_2$	<b>9</b> = $\mu + \alpha_2 + \beta_2$			
	<b>24</b> = $\mu + \alpha_1 + \beta_2$	<b>12</b> = $\mu + \alpha_2 + \beta_2$			
$\bar{a}_1 = 15$		$\bar{a}_2 = 15$			

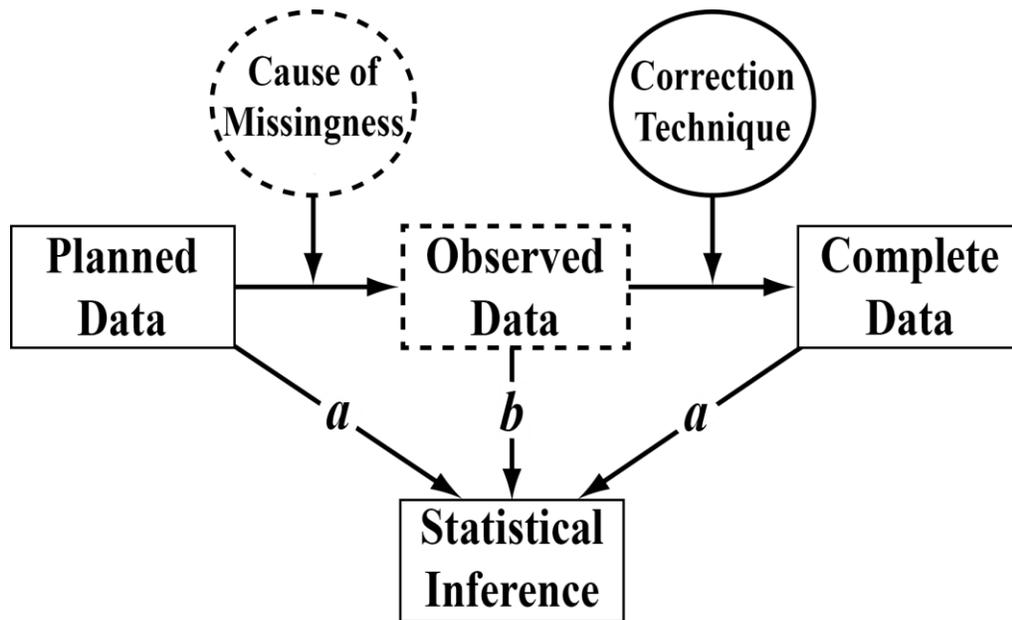
*Note.* Each observed value is written in terms of the ANOVA model parameters where  $\mu$  represents a grand mean,  $\alpha_j$  is the effect of the  $j^{\text{th}}$  level of A,  $\beta_k$  is the effect of the  $k^{\text{th}}$  level of B (excluding  $(\alpha\beta)_{jk}$  and  $\epsilon_{ijk}$  for simplicity).

Table 18

*A Demonstration of marginal means estimation for the main effect of B with incomplete data.*

		<b>Conditions (Factor A)</b>		
<b>Blocks</b>		$a_1$	$a_2$	
$b_1$	<b>13</b> = $\mu + \alpha_1 + \beta_1$	<b>30</b> = $\mu + \alpha_2 + \beta_1$	$\bar{b}_1 = [(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_2 + \beta_1)] / 9$	
	<b>10</b> = $\mu + \alpha_1 + \beta_1$	<b>16</b> = $\mu + \alpha_2 + \beta_1$		
	<b>9</b> = $\mu + \alpha_1 + \beta_1$			
	<b>6</b> = $\mu + \alpha_1 + \beta_1$	<b>18</b> = $\mu + \alpha_2 + \beta_1$		
	<b>12</b> = $\mu + \alpha_1 + \beta_1$	<b>24</b> = $\mu + \alpha_2 + \beta_1$		
$b_2$	<b>26</b> = $\mu + \alpha_1 + \beta_2$	<b>15</b> = $\mu + \alpha_2 + \beta_2$	$\bar{b}_2 = [(\mu + \alpha_1 + \beta_2) + (\mu + \alpha_2 + \beta_2)] / 10$	
	<b>20</b> = $\mu + \alpha_1 + \beta_2$	<b>8</b> = $\mu + \alpha_2 + \beta_2$		
	<b>18</b> = $\mu + \alpha_1 + \beta_2$	<b>6</b> = $\mu + \alpha_2 + \beta_2$		
	<b>12</b> = $\mu + \alpha_1 + \beta_2$	<b>9</b> = $\mu + \alpha_2 + \beta_2$		
	<b>24</b> = $\mu + \alpha_1 + \beta_2$	<b>12</b> = $\mu + \alpha_2 + \beta_2$		
	$\bar{a}_1 = 15$	$\bar{a}_2 = 16$		

*Note.* Each observed value is written in terms of the ANOVA model parameters where  $\mu$  represents a grand mean,  $\alpha_j$  is the effect of the  $j^{\text{th}}$  level of A,  $\beta_k$  is the effect of the  $k^{\text{th}}$  level of B (excluding  $(\alpha\beta)_{jk}$  and  $\varepsilon_{ijk}$  for simplicity).

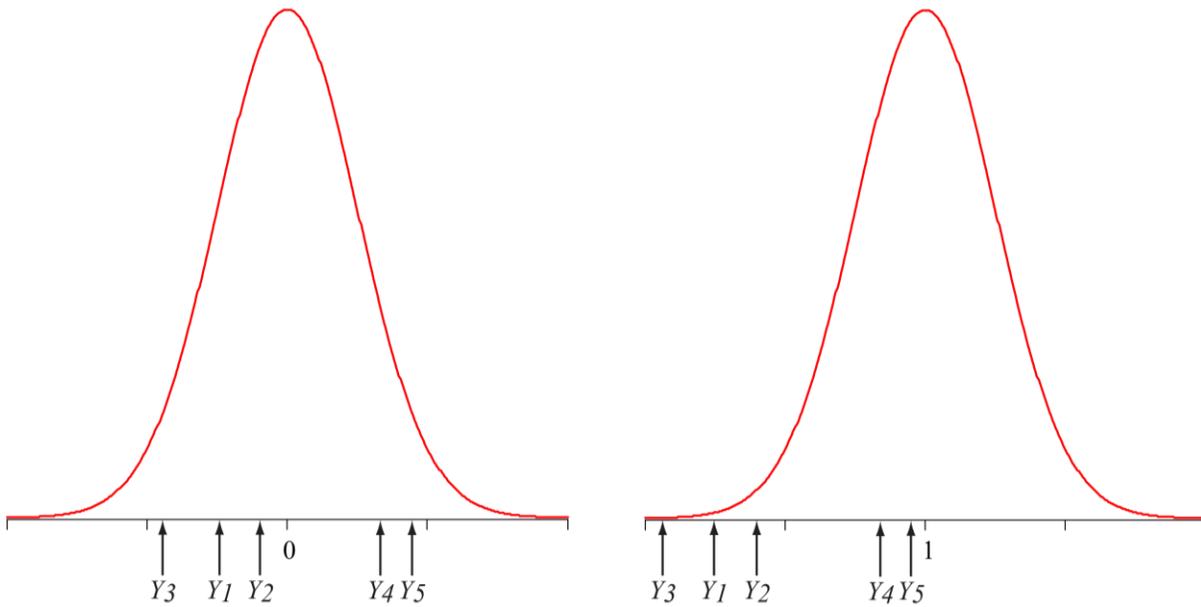


*Figure 1.* The image provides a conceptual demonstration of research that leads to valid statistical inference in the presence of missing data. This diagram shows a research scenario in which an investigator plans to collect data (denoted by the rectangle labeled “Planned Data”); however, some situation (typically unknown to the investigator) causes missing data. This “Cause of Missingness” is demonstrated as a dashed circle. Following the progression from “Planned Data” to “Observed Data” (shown as a predictive arrow pointing from “Planned Data” to “Observed Data”), notice that the “Cause of Missingness” can be thought of as an interaction (displayed as a predictive arrow pointing from “Cause of Missingness” to another predictive arrow). That is, the relationship between the “Planned Data” and the “Observed Data” depends on the “Cause of Missingness”. The “Observed Data” represents data with missingness and is shown as a dashed rectangle to indicate that it is missing some portion of data (i.e., not all of the planned data are observed). Moving from “Observed Data” to “Complete Data” (represented by a predictive arrow between these two rectangles), notice the interaction of “Correction Technique” denoted by a solid circle. This interaction represents the notion that moving from the “Observed Data” that contains missingness to “Complete Data” is dependent upon the missing

data handling technique utilized by the investigator. Conceptually, the “Complete Data” rectangle in this demonstration consists of data that would have been obtained had there been no missing values (i.e., the original “Planned Data”). Said differently, the “Correction Technique” represents a missing data handling technique used to infer the complete data for subsequent statistical modeling. Now, consider that the statistical inference implied from the “Planned Data” (represented by a predictive arrow pointing from “Planned Data” to “Statistical Inference”) should be equivalent to that implied from the complete data (denoted “ $a$ ”). Additionally, the statistical inference implied from deletion methods (i.e., failure to progress to complete data) is likely to be biased (denoted “ $b$ ”).

(A) Distribution with  $\mu_Y = 0, \sigma_Y^2 = 1.5$

(B) Distribution with  $\mu_Y = 1.0, \sigma_Y^2 = 1.5$



*Figure 2.* A graphical representation of the relationship between a sampling density ( $n = 5$ ) and two possible values of  $\mu$ . The figure depicts that the population estimate  $\mu = 0.0$  is more reasonable than population estimate  $\mu = 1.0$  given the observed sample from a normal distribution.

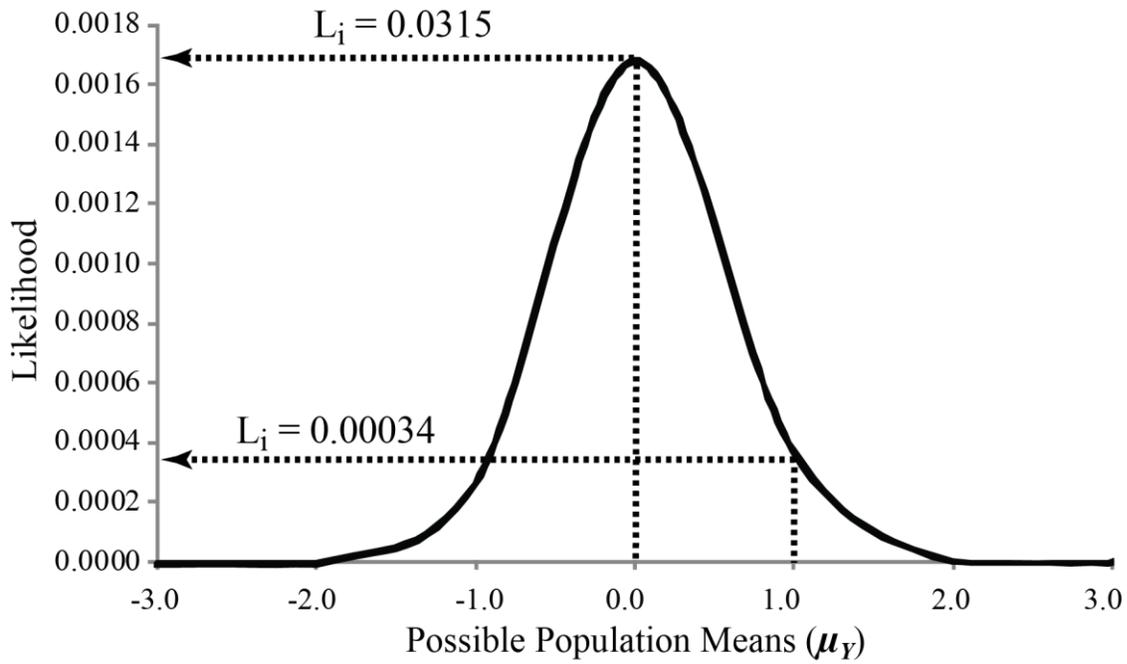


Figure 3. A plot of the likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$ . The image illustrates that the population estimate  $\mu = 0.0$  is most likely (i.e., maximum likelihood) given the observed sample.

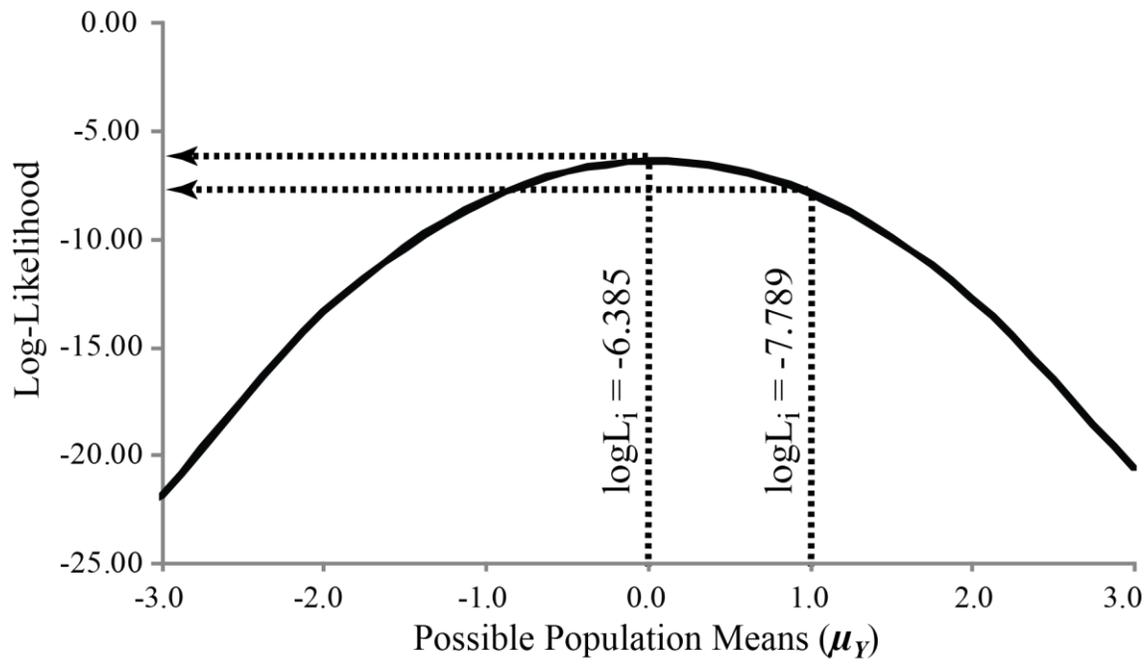


Figure 4. A plot of the log-likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$ . The image illustrates that the population estimate  $\mu = 0.0$  is most likely (i.e., minimum log-likelihood) given the observed sample.

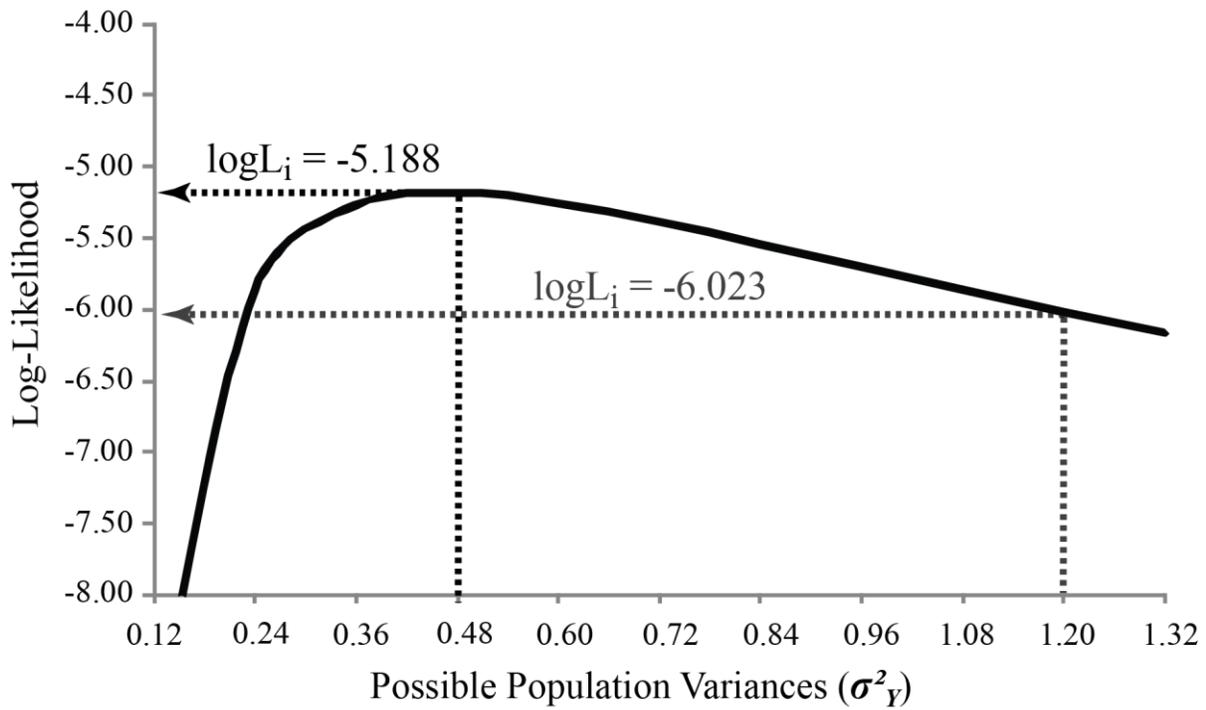
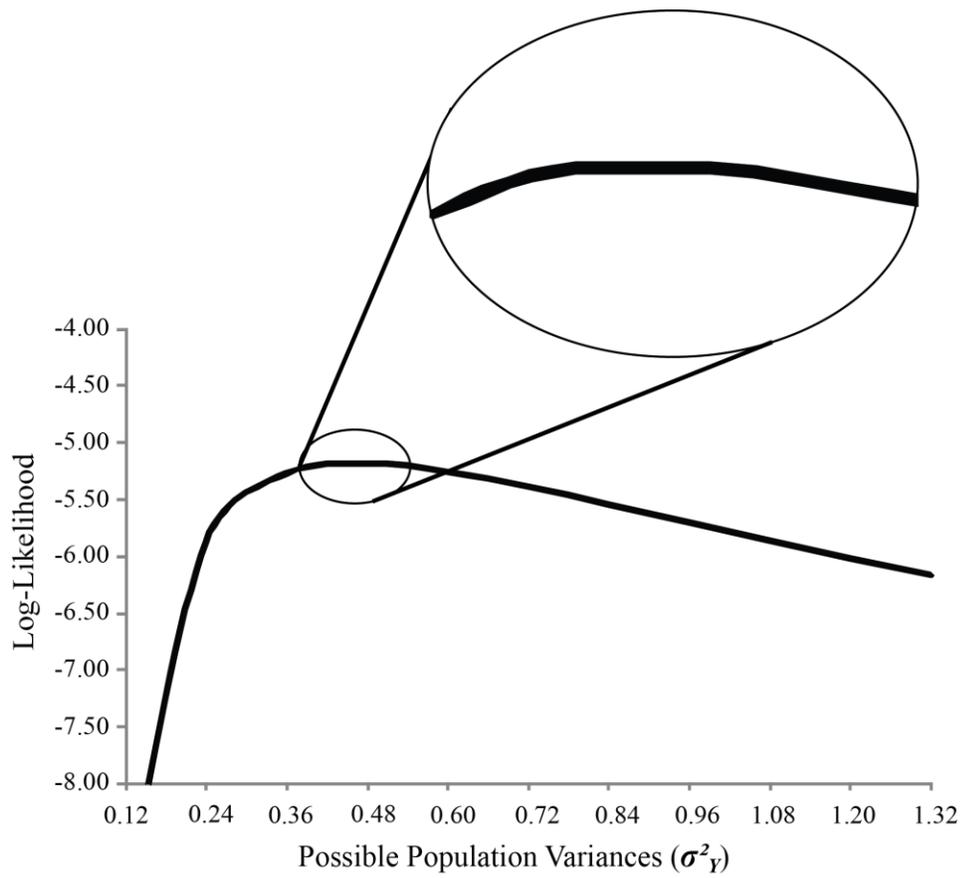


Figure 5. A plot of the log-likelihood values as a function of various population variances ranging from  $\sigma_Y^2 = 0.12$  to  $\sigma_Y^2 = 1.32$ . The image illustrates that the population estimate  $\sigma_Y^2 = 0.48$  is most likely (i.e., minimum log-likelihood) given the observed sample.



*Figure 6.* A close up view of the peak of the likelihood function used to demonstrate MLE precision where the range of  $\sigma^2_\gamma = 0.36 - 0.60$  seem nearly as likely as  $\sigma^2_\gamma = 0.48$ , the maximum likelihood variance estimate.

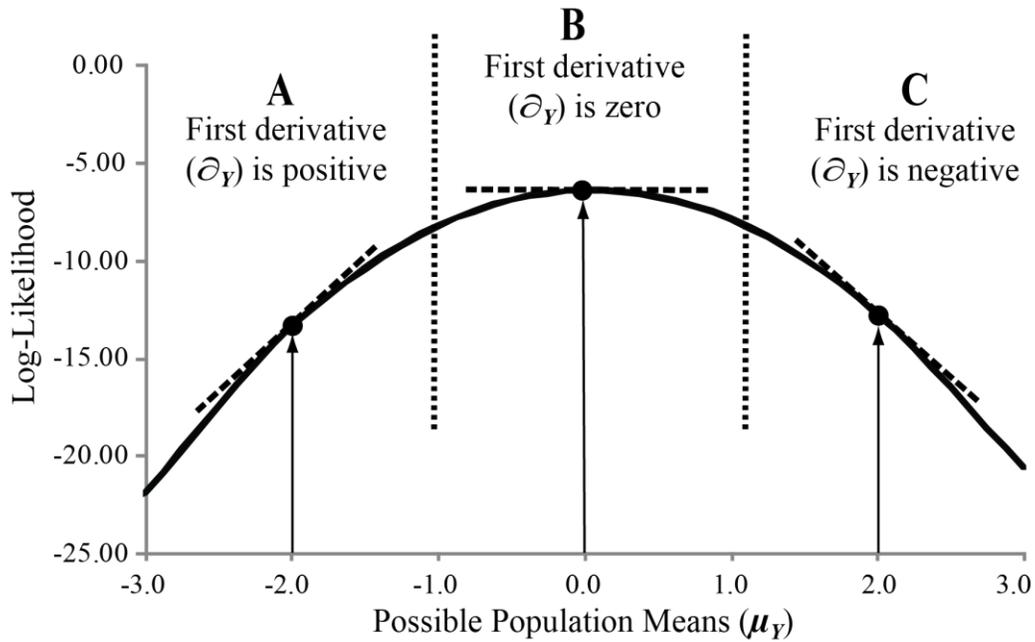


Figure 7. A plot of the log-likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$ . The image illustrates three sections of the likelihood function where tangent lines are imposed on a point of the curve. Section A represents a region where all tangent lines (first derivatives) are positive. Here the slopes are increasing (e.g.,  $\mu_Y = -2.0$ ). Likewise, Section C symbolizes a region in which all tangent lines are negative. Here the slopes are decreasing (e.g.,  $\mu_Y = 2.0$ ). Section B denotes an area that contains the maximum likelihood value (associated with  $\mu_Y = 0.0$ ) where the tangent line is flat and the slope is zero.

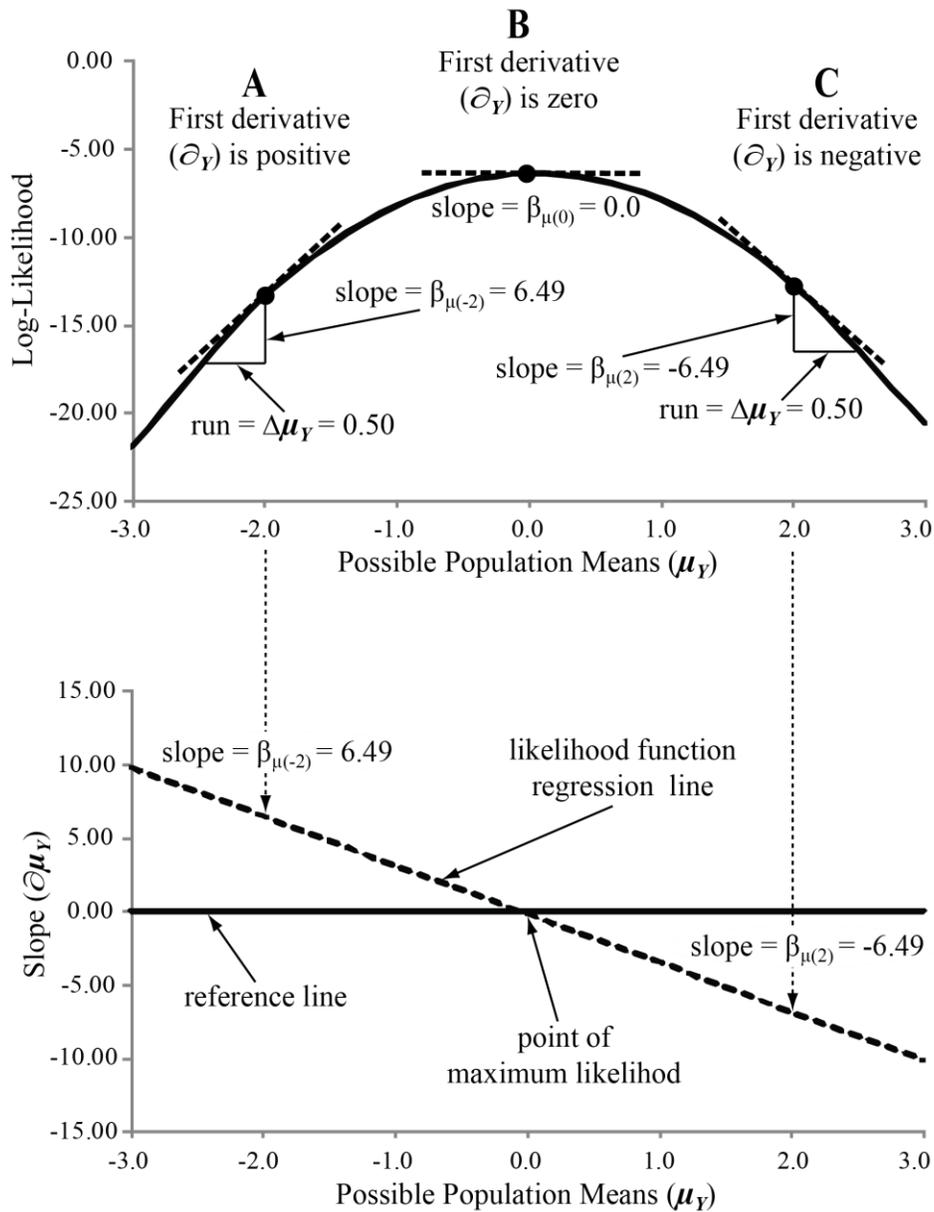
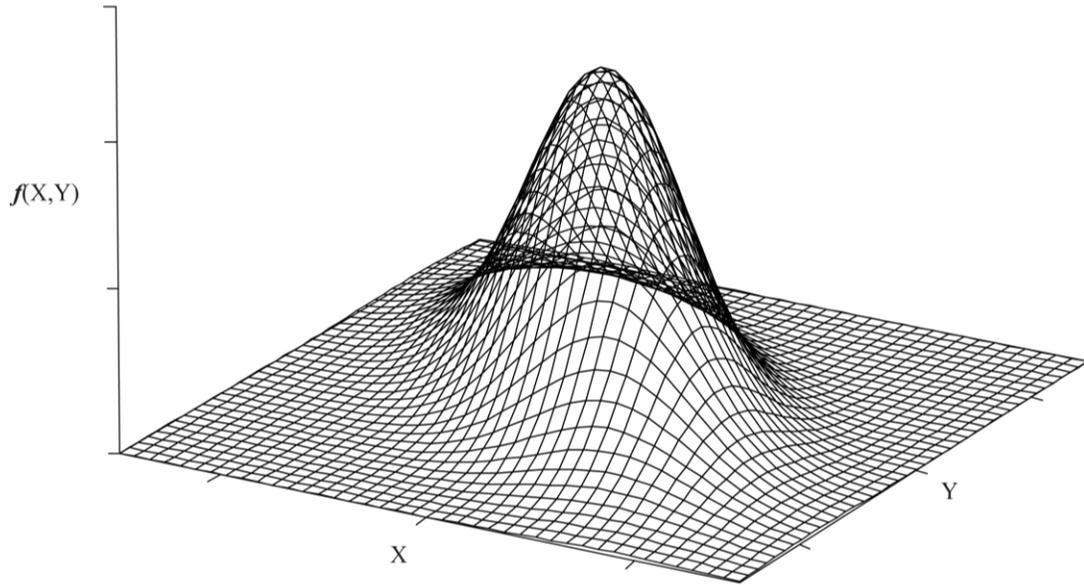
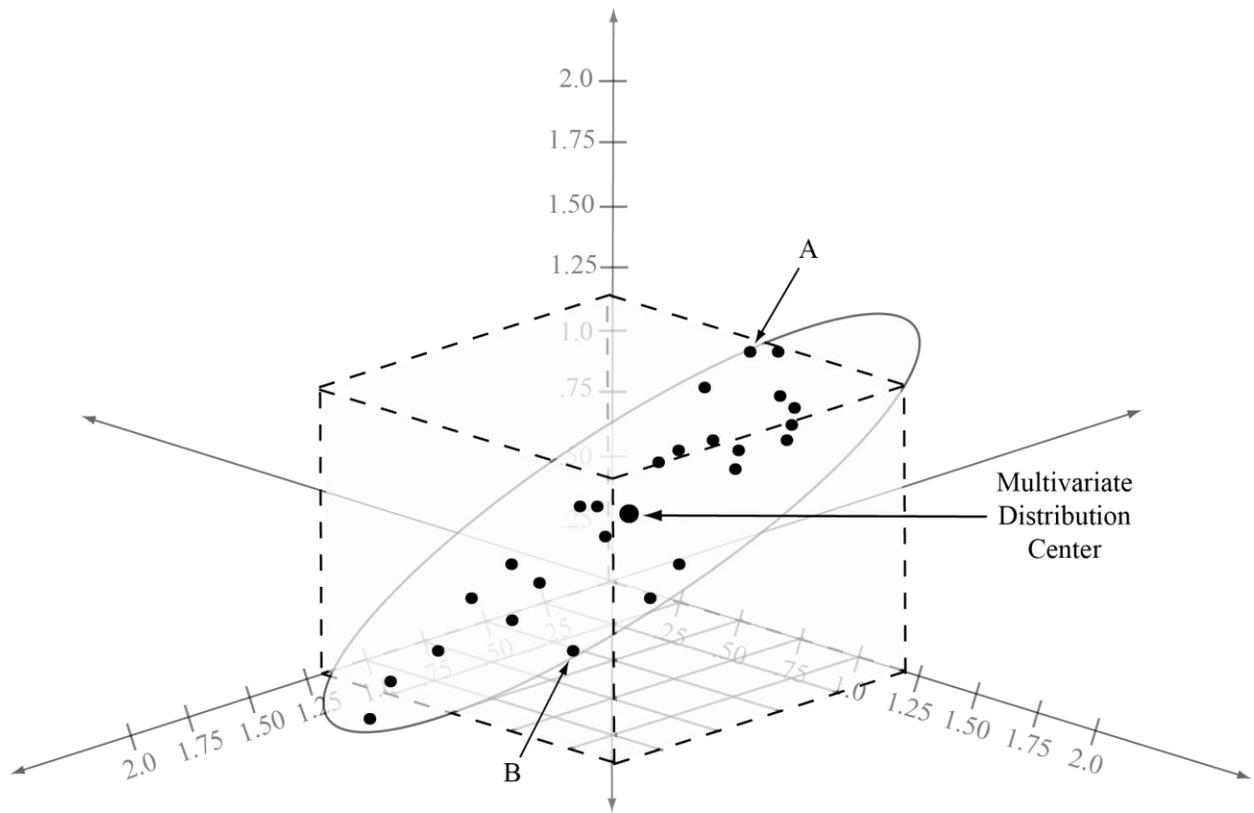


Figure 8. A plot of the log-likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$  (top). This image explicitly illustrates the first derivatives of  $\mu_Y = -2.0$ ,  $\mu_Y = 0.0$ , and  $\mu_Y = 2.0$  in relation to the likelihood function (calculated from equation 19). A specific plot of the derivative regression line related to the likelihood function (bottom). The image illustrates the relationship between the likelihood function and the first derivatives. Note that the regression line transitions from positive values to negative values as it crosses the point of maximum likelihood.



*Figure 9.* A three-dimensional plot of a bivariate normal probability distribution with  $X$  and  $Y$  where the density function  $f(X,Y)$  represents the height of the probability surface.



*Figure 10.* Illustration of Mahalanobis distance (denoted by the correlation ellipse) in a three-dimensional scatter plot where point A and point B are the same distance from point C, the multivariate centroid.

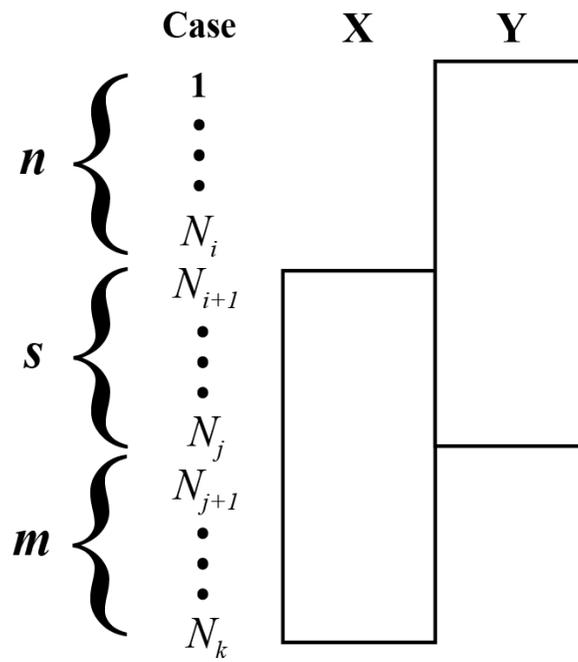
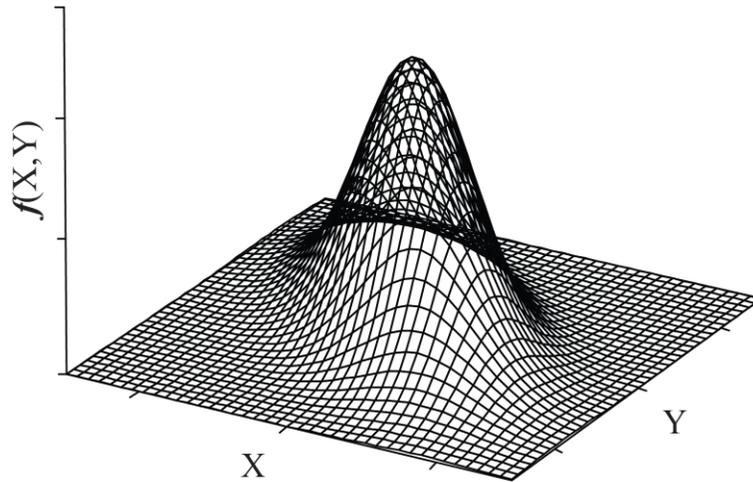
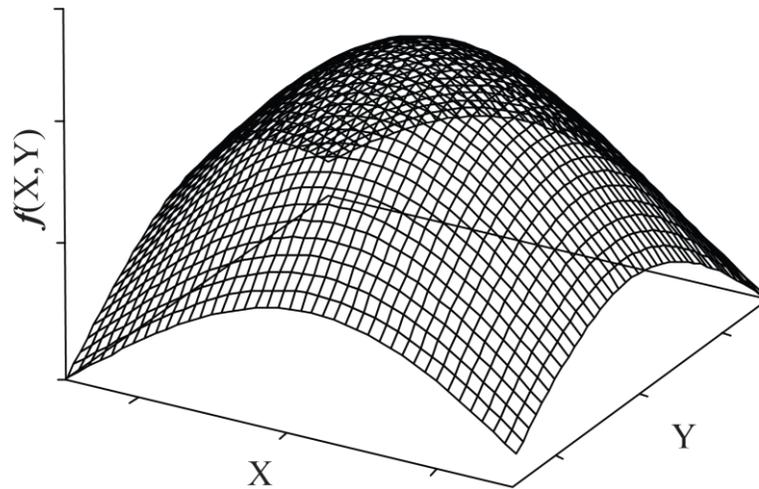


Figure 11. Illustration of missing data patterns where  $n$  is the number of  $i$  cases observed on  $\mathbf{Y}$  but not on  $\mathbf{X}$ ,  $s$  represents  $i + 1$  to  $j$  cases that are observed on both  $\mathbf{X}$  and  $\mathbf{Y}$ , and  $m$  signifies  $j + 1$  to  $k$  cases observed on  $\mathbf{X}$  but not on  $\mathbf{Y}$ .

**(A) Likelihood Function (less variance - more information)**

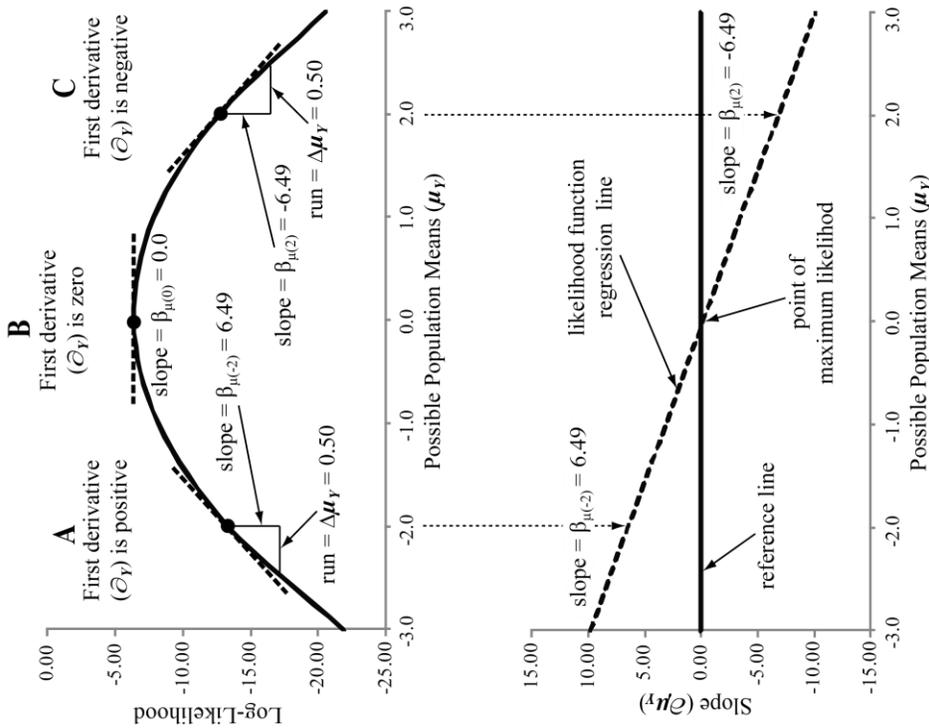


**(B) Likelihood Function (more variance - less information)**



*Figure 12.* A depiction of two three-dimensional plots of a bivariate normal probability distribution with variables  $X$  and  $Y$  where the density function  $f(X,Y)$  represents the height of the probability surface. Panel A relates to high information and small standard errors and Panel B relates to low information and large standard errors.

(A) Likelihood Function (less variance - more information)



(B) Likelihood Function (more variance - less information)

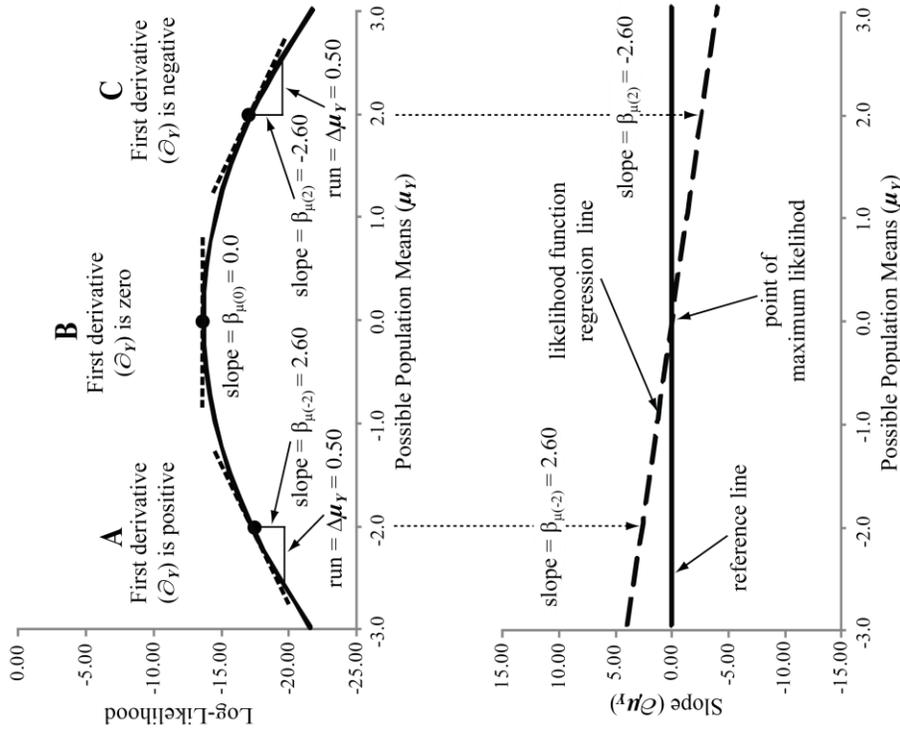


Figure 13. A graphical depiction of the log-likelihood values as a function of various population means ranging from  $\mu_Y = -3.0$  to  $\mu_Y = 3.0$  (top of panels A and B). This image explicitly illustrates the first derivatives of  $\mu_Y = -2.0$ ,  $\mu_Y = 0.0$ , and  $\mu_Y = 2.0$  in relation to the likelihood function (calculated from equation 19) for two random samples with the same mean; however, Panel B corresponds to a sample with a variance that is 2.5 times larger than the sample data illustrated in Panel A. The image illustrates two likelihood functions that estimate the same population mean (e.g.,  $\mu_Y = 0.0$ ) but vary in the precision of the parameter estimate (rate of change of the first derivatives). Notice that the slopes demonstrated in Panel B are 2.5 times smaller than the corresponding slopes in Panel A.

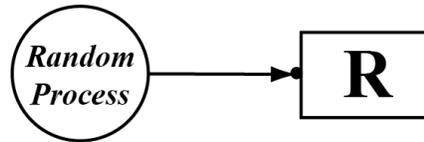
$$\begin{array}{c}
\text{Incomplete Data Matrix} \\
\mathbf{Y} = \left[ \begin{array}{cccccccccc}
y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & - & - & y_{18} & \cdots & y_{1p} \\
y_{21} & y_{22} & y_{23} & y_{24} & - & y_{26} & y_{27} & - & \cdots & y_{2p} \\
y_{31} & y_{32} & y_{33} & - & - & y_{36} & - & y_{38} & \cdots & y_{3p} \\
\vdots & \ddots & \vdots \\
y_{i1} & y_{i2} & y_{i3} & y_{i4} & y_{i5} & y_{i6} & y_{i7} & y_{i8} & \cdots & y_{ip}
\end{array} \right] \left. \vphantom{\begin{array}{c} \mathbf{Y} \\ \mathbf{R} \end{array}} \right\} N \text{ participants} \\
\underbrace{\hspace{10em}}_{p \text{ variables}} \\
\downarrow \\
\text{Missingness Indicator Matrix} \\
\mathbf{R} = \left[ \begin{array}{cccccccccc}
1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & \cdots & r_{1p} \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & \cdots & r_{2p} \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & \cdots & r_{3p} \\
\vdots & \ddots & \vdots \\
r_{i1} & r_{i2} & r_{i3} & r_{i4} & r_{i5} & r_{i6} & r_{i7} & r_{i8} & \cdots & r_{ip}
\end{array} \right] \left. \vphantom{\begin{array}{c} \mathbf{Y} \\ \mathbf{R} \end{array}} \right\} N \text{ participants} \\
\underbrace{\hspace{10em}}_{p \text{ variables}}
\end{array}$$

Figure 14. Illustration of an incomplete data matrix  $\mathbf{Y}$  where missingness is denoted by “-” and a missingness indicator matrix  $\mathbf{R}$  where missingness is denoted by “0”. Note that the matrix  $\mathbf{R}$  has the same dimensions as the matrix  $\mathbf{Y}$ .

		<b>Observed Variables</b>	
		$Y_p$ and $\mathbf{R}$ are independent: $p(\mathbf{R})$	$Y_p$ and $\mathbf{R}$ are dependent: $p(\mathbf{R}   Y_{obs})$
<b>Unobserved Variables</b>	$Y_p$ and $\mathbf{R}$ are independent: $p(\mathbf{R})$	<b>MCAR</b> <ul style="list-style-type: none"> <li>• Partially to Fully Recoverable</li> <li>• Completely Unbiased</li> </ul>	<b>MAR</b> <ul style="list-style-type: none"> <li>• Partially to Fully Recoverable</li> <li>• Less Biased to Unbiased</li> </ul>
	$Y_p$ and $\mathbf{R}$ are dependent: $p(\mathbf{R}   Y_{mis})$	<b>MNAR</b> <ul style="list-style-type: none"> <li>• Unrecoverable</li> <li>• Biased</li> </ul>	<b>MAR / MNAR</b> <ul style="list-style-type: none"> <li>• Partially Recoverable</li> <li>• Biased to Unbiased</li> </ul>

Figure 15. A modified version of Little's (2009) schematic representation of the missing data mechanisms and the associated frameworks, together with measured ( $Y_{obs}$ ) and unmeasured ( $Y_{mis}$ ) variable influences.

### (A) MCAR Conceptually



### (B) Representation of MCAR

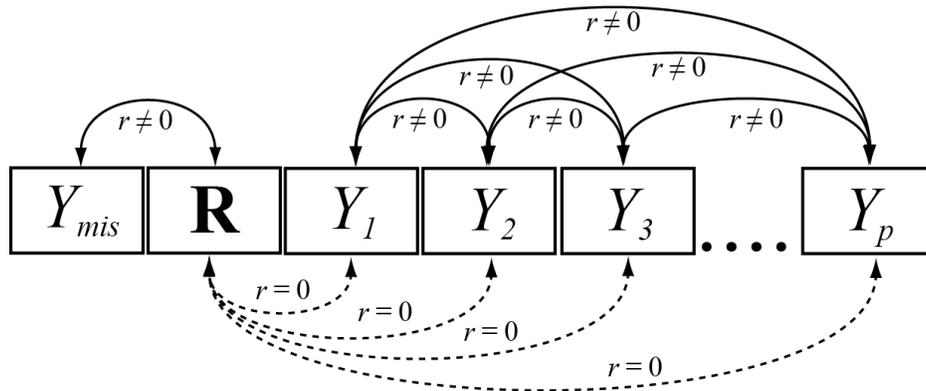
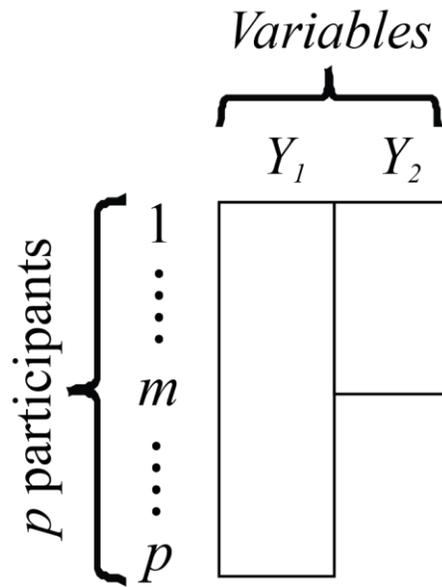


Figure 16. A graphical representation of the MCAR missing data mechanism. Panel A provides a conceptual diagram of MCAR where missingness (denoted  $R$ ) is regressed on a random process. Panel B provides a more detailed representation of MCAR that depicts  $p$  observed variables ( $Y_1 - Y_p$ ) as boxes that are related (correlated) to some degree represented by solid reciprocal paths where  $r \neq 0$ . Additionally, there is no relationship between  $R$  and  $Y_1 - Y_p$  which is represented by dotted reciprocal paths where  $r = 0$ . Visually, imagine that information can flow within the diagram in Panel B through pathways are  $r \neq 0$ . Note that the dot on the tip of the arrow point in Panel A indicates a random process.



*Figure 17.* Univariate missing data pattern with missing on  $Y_2$  but not on  $Y_1$ .

### (A) MAR Conceptually



### (B) Representation of MAR

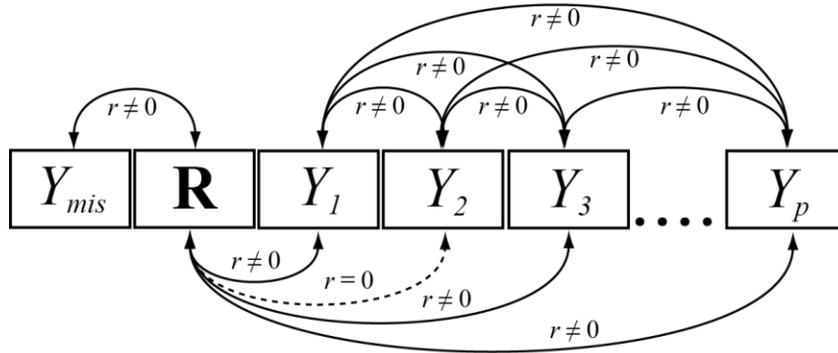
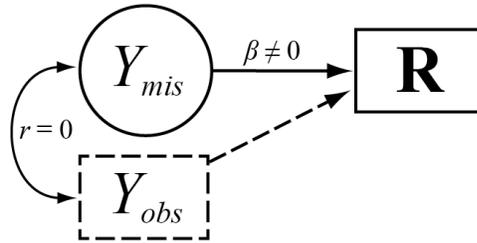


Figure 18. Diagram of the MAR missing data mechanism. Panel A offers a conceptual diagram MAR that shows observed variables ( $Y_{obs}$ ) predicting the missing data ( $\mathbf{R}$ ). Panel B shows the MAR mechanism in relation to  $p$  variables where  $Y_2$  contains missing data. Visually, imagine that information can flow within the diagram in Panel B through pathways are  $r \neq 0$ .

### (A) MNAR Conceptually



### (B) Representation of MNAR

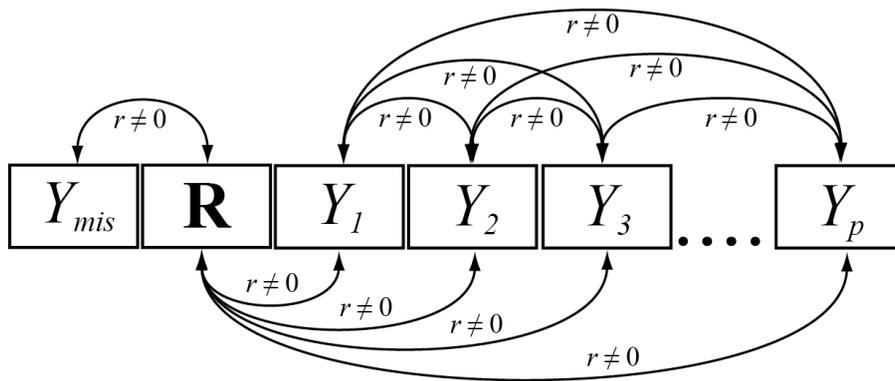


Figure 19. Diagram of the MNAR missing data mechanism. Panel A of Figure 21 provides a theoretical explanation of MNAR. Here unmeasured variables (denoted  $Y_{mis}$ ) predict (i.e., cause) missingness. Panel B of Figure 21 shows the MNAR mechanism in relation to  $p$  variables where  $Y_2$  contains missing data.

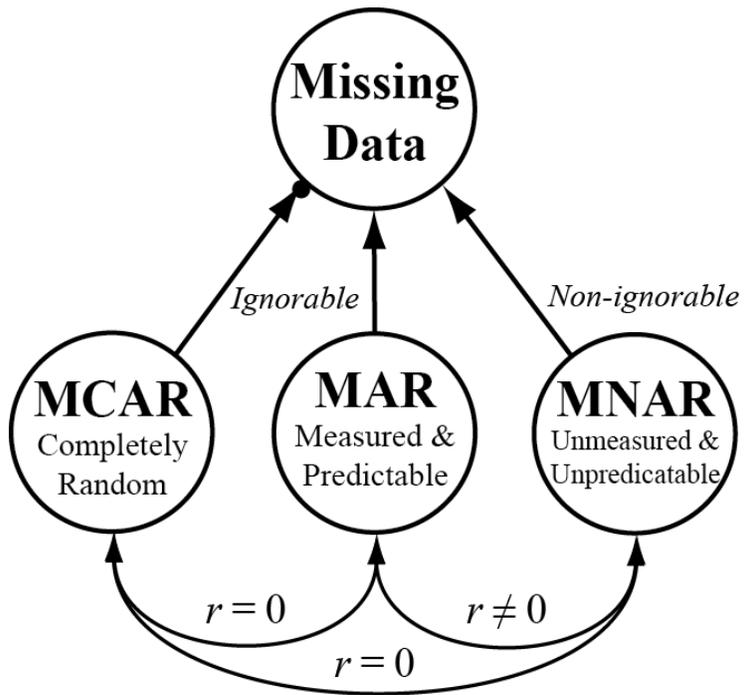


Figure 20. Diagram illustrating a conceptual overview of the missing data mechanisms where each contributes to the observed missing data in a given study adopted from Little (2002).



*Figure 21.* Diagram illustrating the relationship between MNAR and MAR as a continuum adopted from Graham (2012).

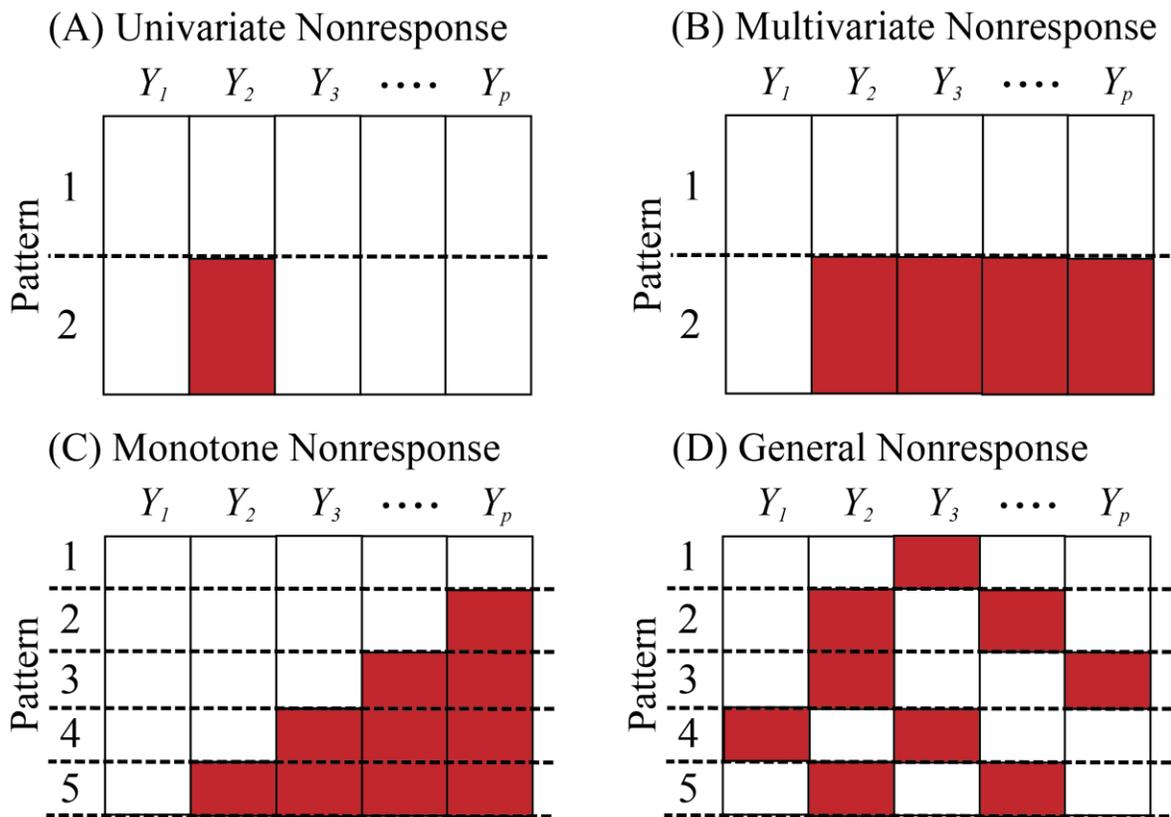
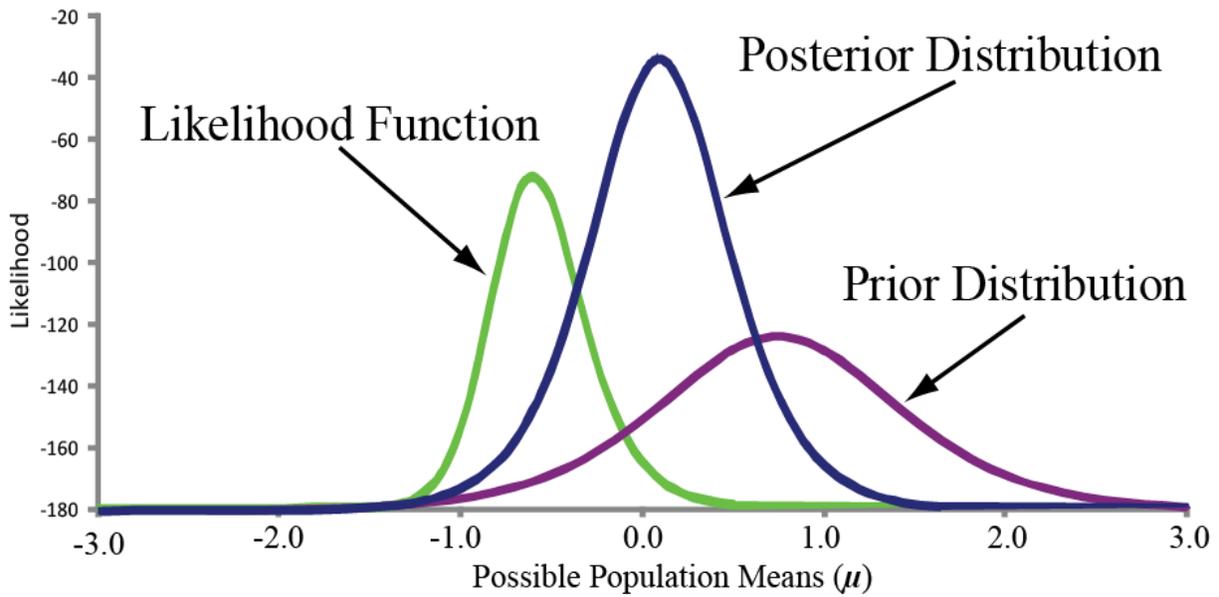


Figure 22. An illustration of common missing data patterns. Within each Panel, variables ( $Y_1 - Y_p$ ) are displayed across the top of the image while individual patterns are labeled on the left of the image. Note that shaded blocks represent missing data.



*Figure 23.* A graphical demonstration of the Bayesian approach used by Rubin (1977) to address missing data. Note that the multiplication of the prior distribution by the likelihood function generates the posterior distribution. In this image the relative shapes of each distribution provide information about how informative they are. That is, the height and width of each distribution reflects the probability density for that distribution. For instance, the prior distribution in this example is not very informative as it is relatively flat. An uninformative prior distribution could be represented as a uniform distribution.

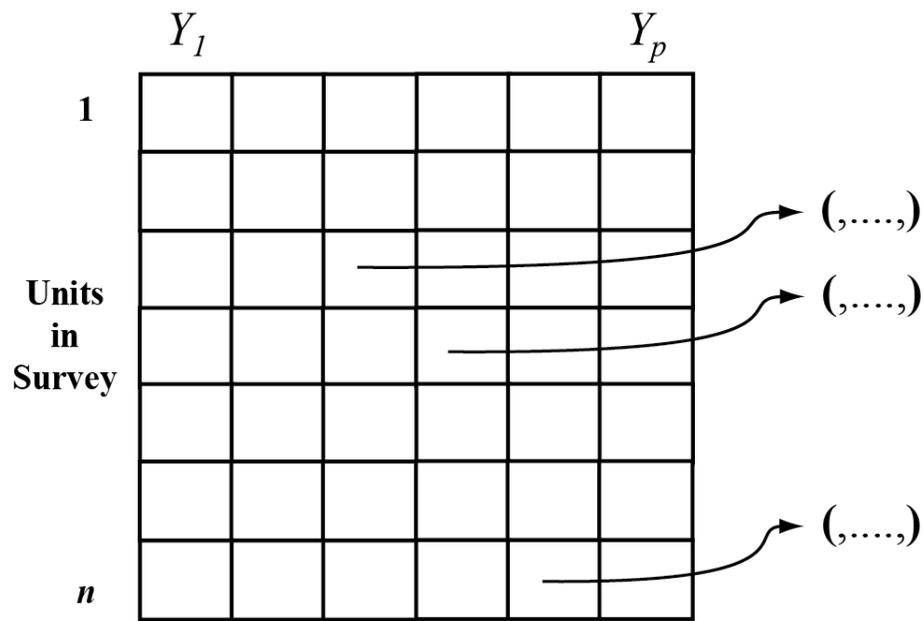


Figure 24. A graphical depiction of Rubin's (1978a) multiple imputation procedure modified from his original paper. Here each missing data point (shown as the start point of an arrow) is replaced by a vector (denoted  $(, \dots, )$ ) of plausible values.

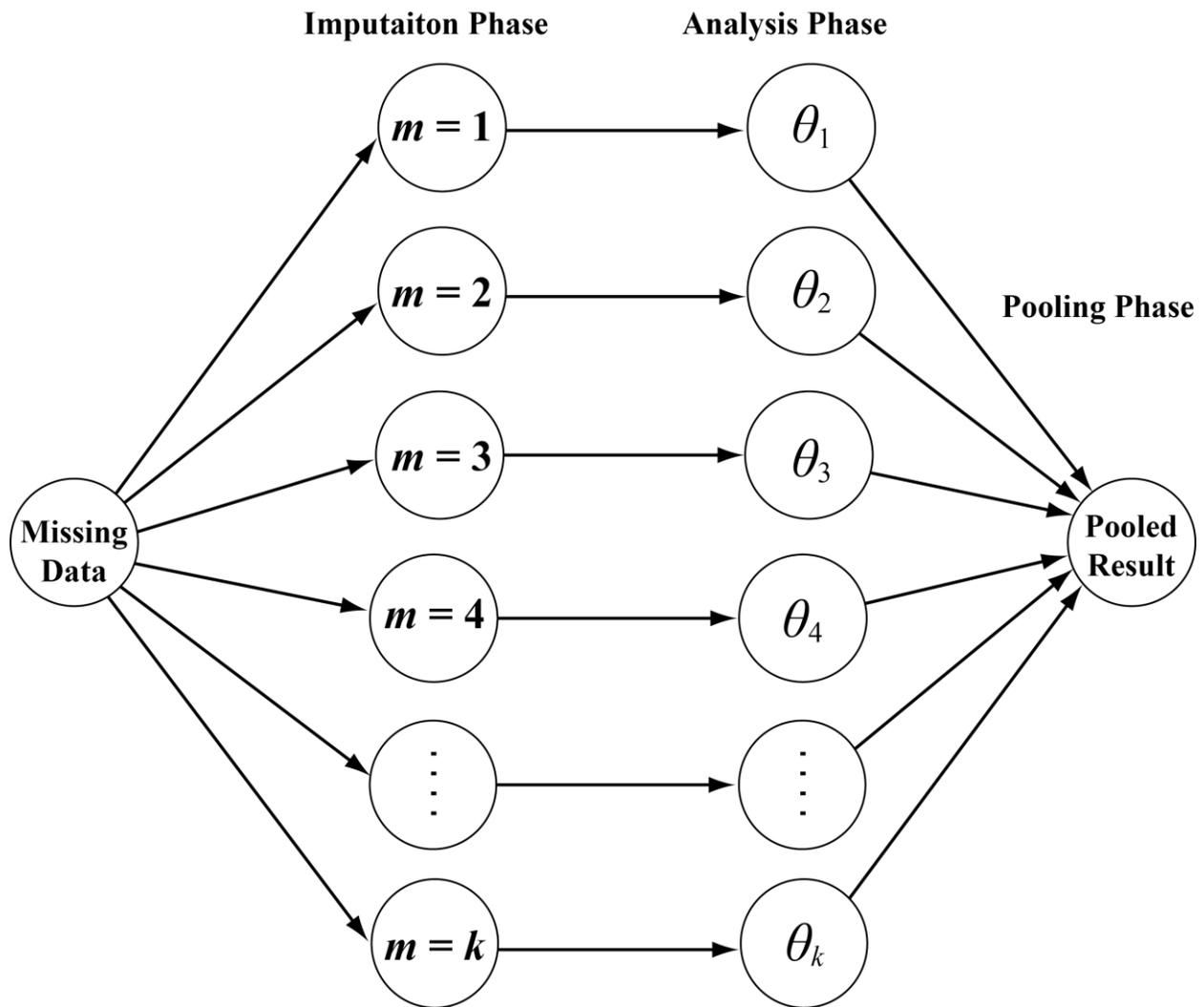
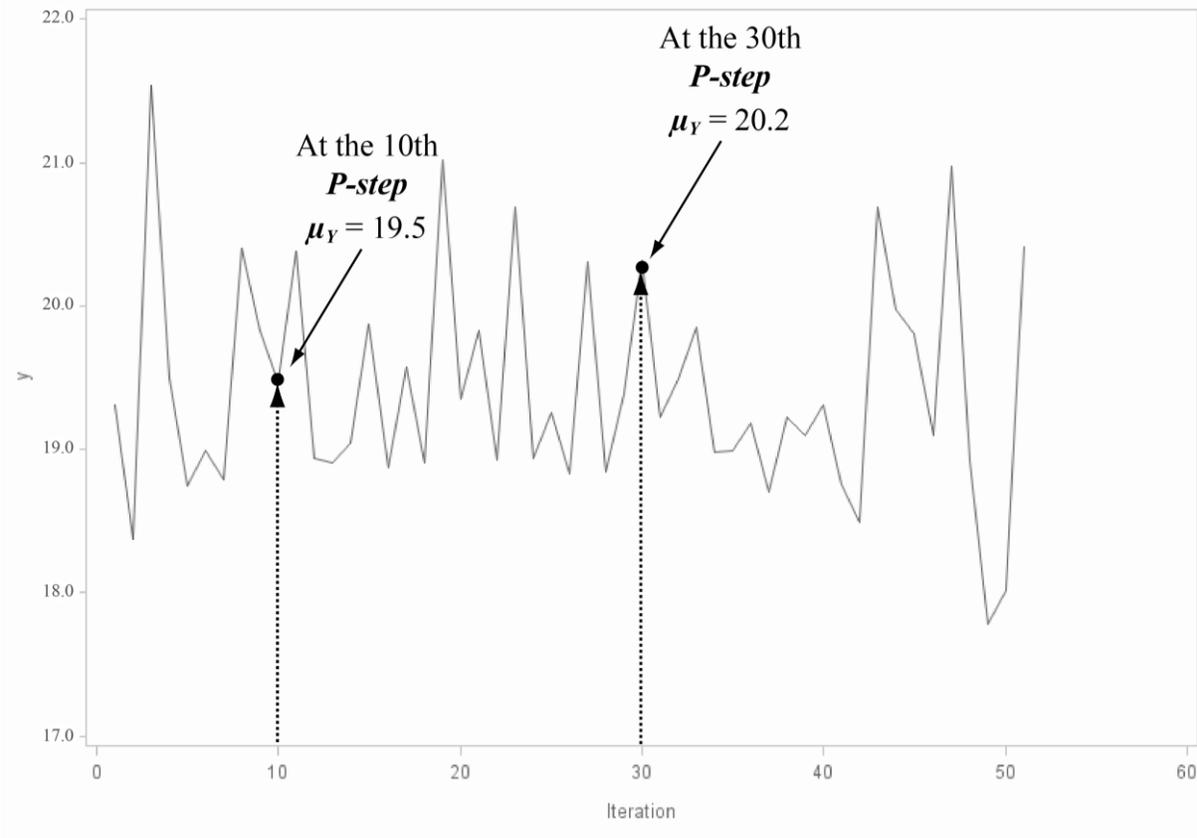


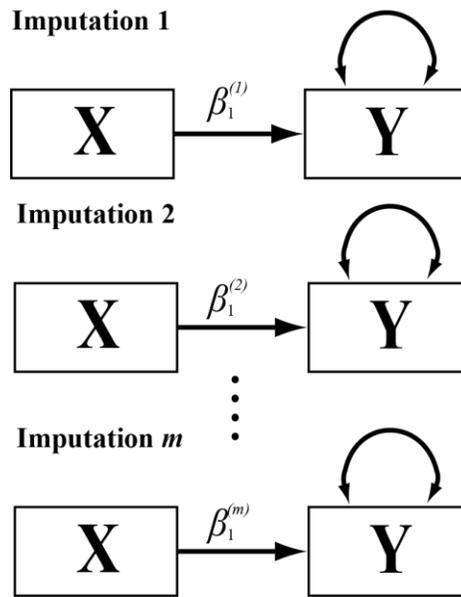
Figure 25. An illustration of Rubin’s (1987) multiple imputation procedure modified from Enders (2010). The circle labeled “Missing Data” denotes a data set with incomplete data. The arrows pointing from the “Missing Data” circle to the circles labeled “ $m = 1$ ” through “ $m = k$ ” represent each imputation (as demonstrated in Figure 25). The arrows pointing from each  $m$  imputation to the circles labeled “ $\theta_1$ ” through “ $\theta_k$ ” illustrate a separate analysis of each imputed data set, where the  $k^{\text{th}}$   $\theta$  represents a parameter estimate based on data from a particular imputation. Finally, the arrows pointing from each of the  $k^{\text{th}}$   $\theta$  circles to the circle labeled “Pooled Result” represent an process of combining (i.e., pooling) each of the imputed results. The circle labeled “Pooled Result” represents the analytic solution informed by the MI process.

$$\mathbf{Y} = \begin{bmatrix} 21.15 \\ Y_{miss} \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 21.14 \\ 20.22 \\ 20.81 \\ 20.69 \\ 19.59 \\ 18.33 \\ 15.40 \\ 20.34 \end{bmatrix} \rightarrow \mathbf{Y}_{obs} = \begin{bmatrix} 21.15 \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} \quad \mathbf{X}_{obs} = \begin{bmatrix} 21.14 \\ 20.81 \\ 20.69 \\ 19.59 \\ 18.33 \\ 15.40 \\ 20.34 \end{bmatrix}$$

Figure 26. The Y and X vectors from the previous Yates method example. Notice that  $\mathbf{Y}_{obs}$ ,  $\mathbf{X}_{obs}$  contain only associated values that are observed. That is,  $\mathbf{Y}_{obs}$  does not contain missing values and  $\mathbf{X}_{obs}$  does not contain the value 20.22 as it is associated with the  $Y_{miss}$  value.



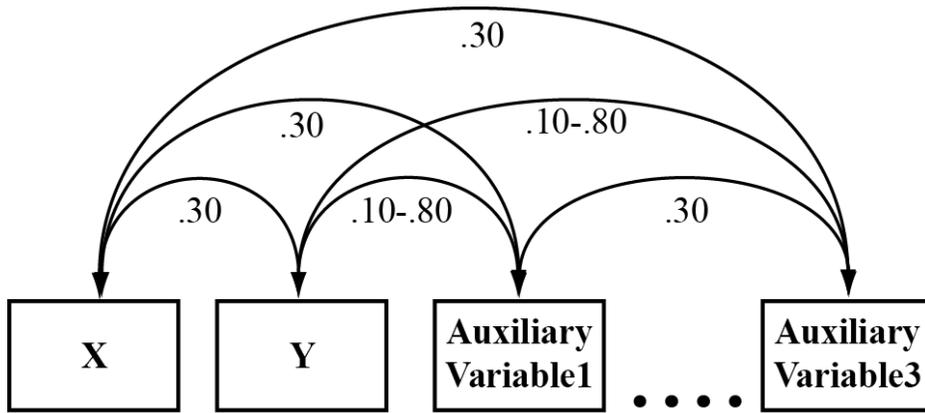
*Figure 27.* Trace plot for the simulated mean of  $Y$ . The  $Y$ -axis contains a range of mean estimates for  $Y$  and the  $X$ -axis displays the number of data augmentation cycles. This image demonstrates iterative variation associated with each imputation. Notice that at the 10<sup>th</sup> iteration (i.e., data augmentation cycle) the  $Y$  parameter was 19.5 and at the 30<sup>th</sup> iteration was 20.2, which are close to the estimates derived above (e.g., 19.065 and 19.413). Also observe that an arbitrarily small number of iterations were chosen (i.e., max P-step = 50) for clarity of presentation.



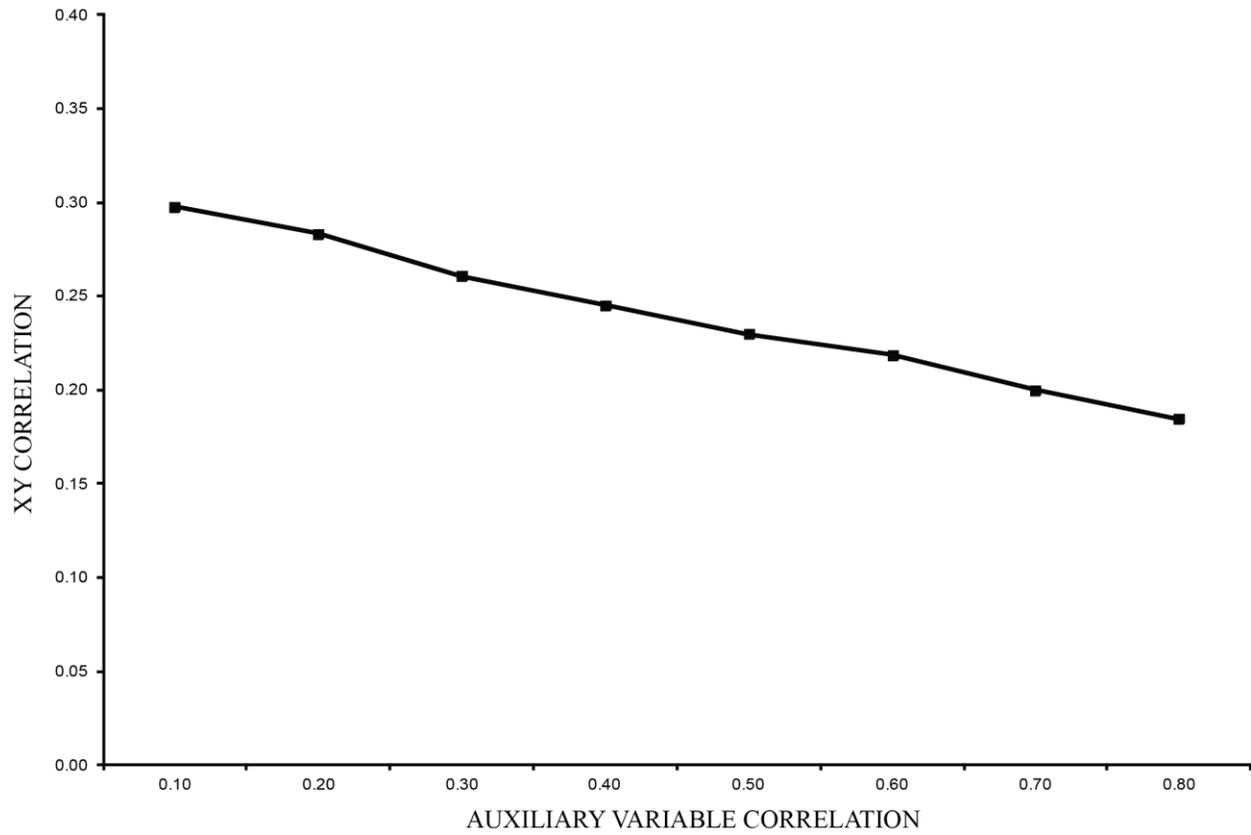
*Figure 28.* Illustration of a bivariate regression in the context of MI. Note that the superscript in parentheses indicates the imputation number associated with each of the  $m$  beta weights.

Case	X	Y
1		
•		
•		
•		
<i>m</i>		
•		
•		
•		
•		
<i>p</i>		

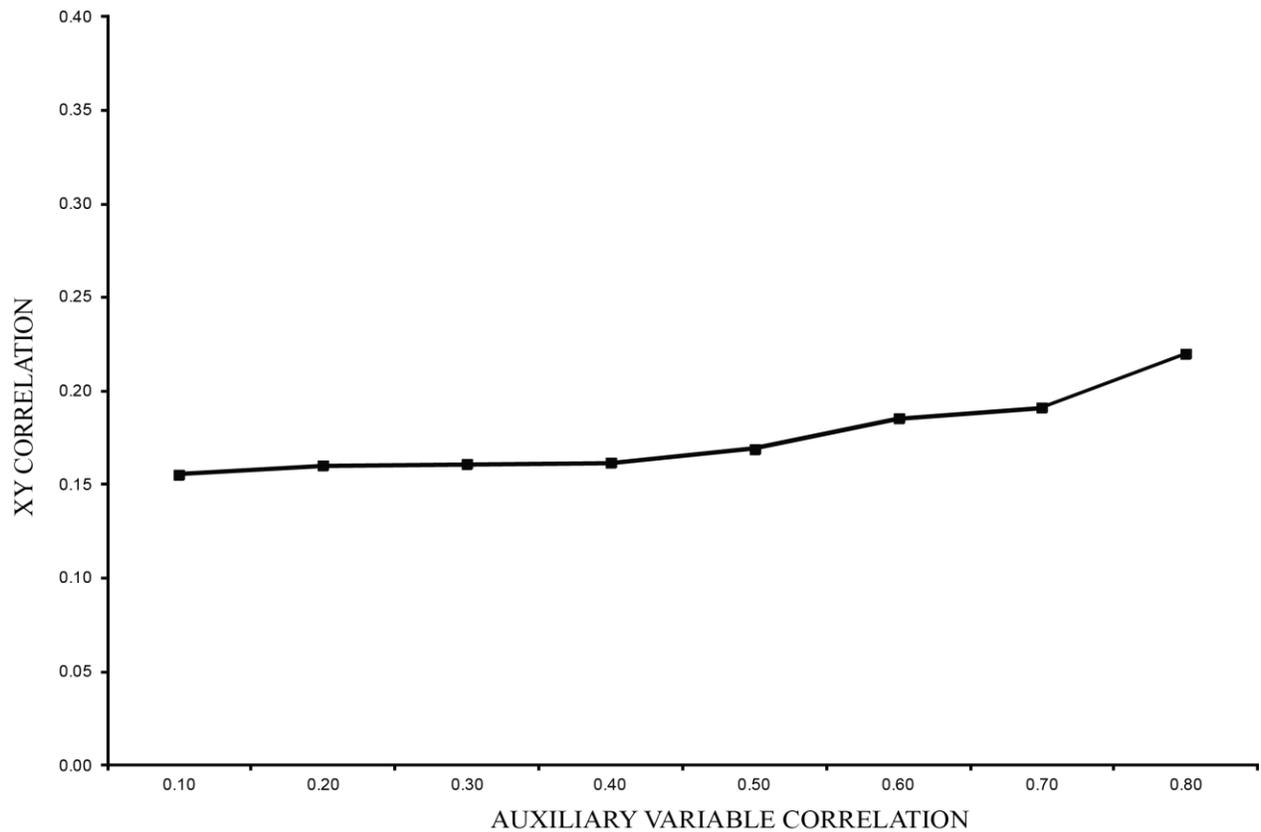
Figure 29. Illustration of a missing data pattern with missing on Y but not on X.



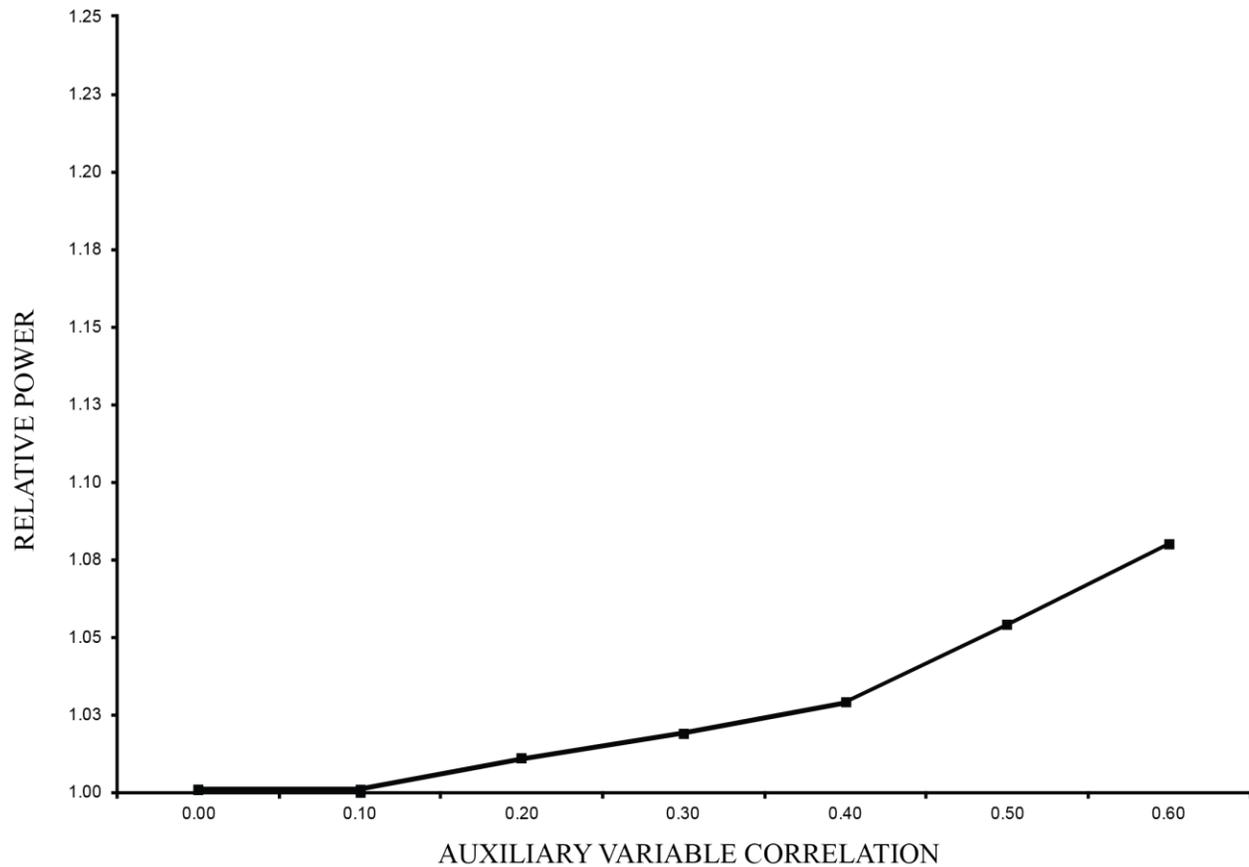
*Figure 30.* A path diagram illustrating the range of population correlations ( $\rho$ ) relative to the analysis variables and the associated auxiliary variables.



*Figure 31.* A plot of simulation results showing bias associated with the exclusion of a cause or correlate of missingness. Note that “AUXILIARY VARIABLE CORRELATION” denotes the population correlation between  $Y$  and the omitted auxiliary variable ( $AUX_I$ ) and “XY CORRELATION” refers to the population association among the variables  $X$  and  $Y$  ( $\rho_{XY} = .30$ ).



*Figure 32.* Simulation results showing the bias reduction associated with including auxiliary variables in a MNAR situation. Note that “AUXILIARY VARIABLE CORRELATION” denotes the population correlation between  $Y$  and the omitted auxiliary variable ( $AUX_I$ ) and “XY CORRELATION” refers to the population association among the variables  $X$  and  $Y$  ( $\rho_{XY} = .30$ ).



*Figure 33.* Simulation results showing the relative power associated with including auxiliary variables in a MCAR Situation. Note that “AUXILIARY VARIABLE CORRELATION” denotes the population correlation between  $Y$  and the omitted auxiliary variable ( $AUX_I$ ) and “RELATIVE POWER” refers to the estimated increase in power relative to a model with no auxiliary variables.

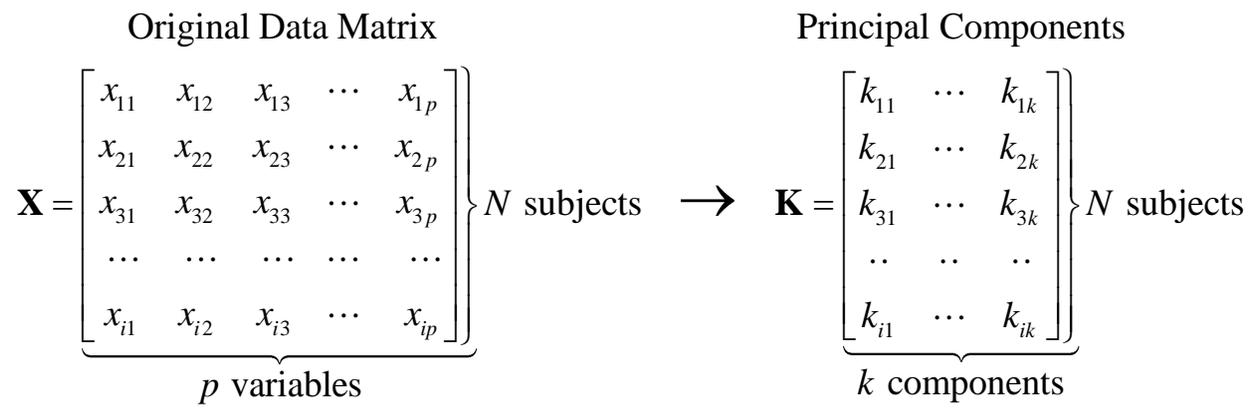


Figure 34. Illustration of data reduction using PCA.

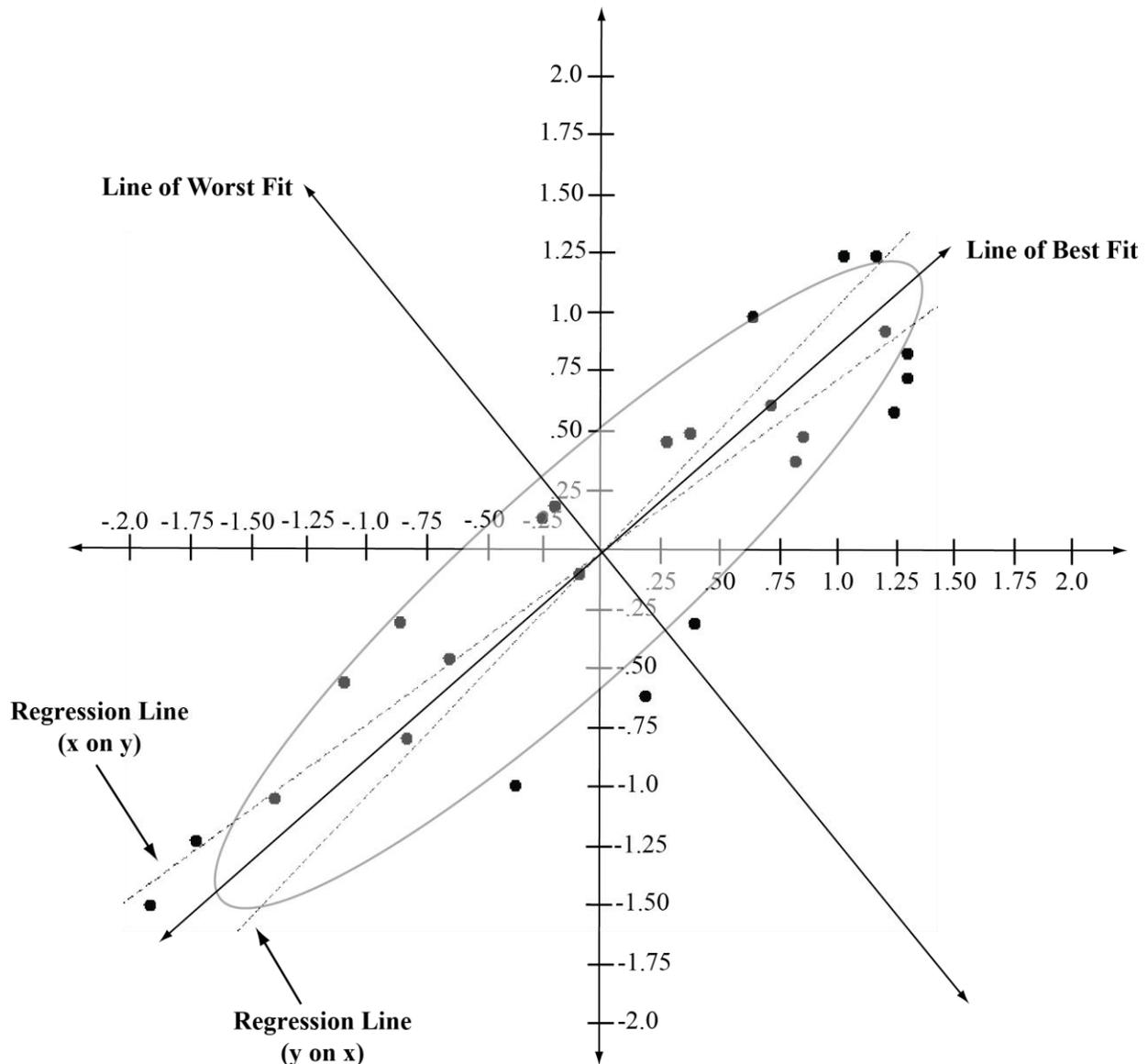


Figure 35. Scatterplot of the data in Table 10 ( $N = 25$ ) with a correlation ellipse and regression lines (see dotted lines) that reflect the arbitrary division between  $y$  regressed on  $x$  (where  $Y_i = \beta_0 + \beta_1 X_i + e_i$ ) and of  $x$  regressed on  $y$  (where  $X_i = \beta_0 + \beta_1 Y_i + e_i$ ). Note  $\beta_1$ , the unstandardized regression coefficient, represents the average change in  $Y$  for each unit change in  $X$  or vice versa. Likewise, the error term  $e_i$  represents the error associated with the prediction of  $Y$  from  $X$  or vice versa (see Hays, 1994). The “Line of Best Fit” (which is the first principal component) minimizes the sum of squared errors from each point irrespective of the

specification of a dependent and independent variable. The “Line of Worst Fit” (which is the second principal component) minimizes the sum of squared errors not already explained by the “Line of Best Fit”. In terms of the correlation ellipse, the “Line of Best Fit” corresponds to the most expanded portion of the ellipse. Similarly, the “Line of Worst Fir” maps onto the most compressed portion of the ellipse.

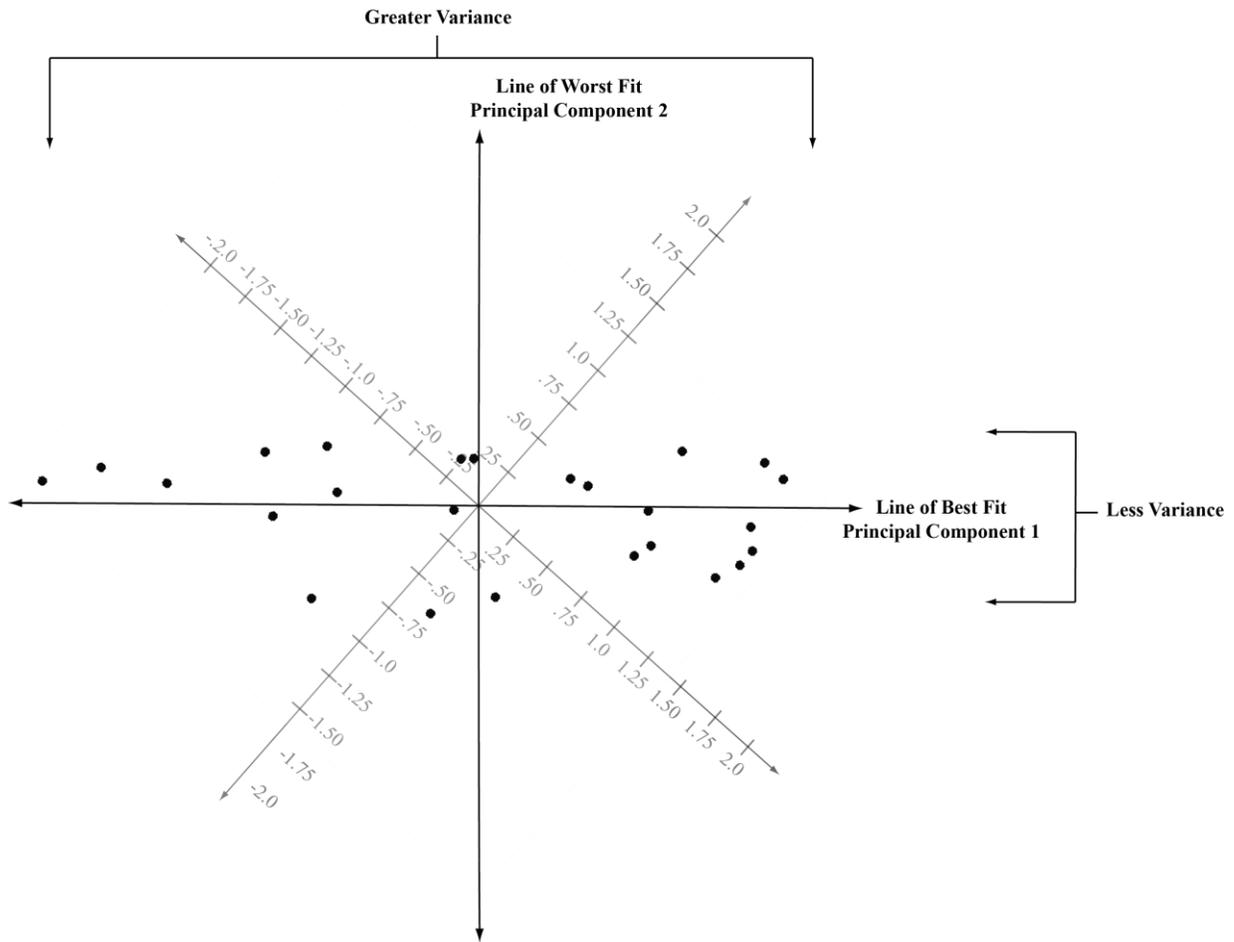


Figure 36. Illustration of the rotated scatterplot from Figure 37. Note that the “Line of Best Fit” and “Line of Worst Fit” now occupy the original coordinate position of variables x and y. Through principal axis rotation the locations of the new axes define the new variables (i.e., the principal components). Note that the new variable called “Principal component 1”, explains most of the variability in the scatterplot. Principal component 2 captures the remaining variance.

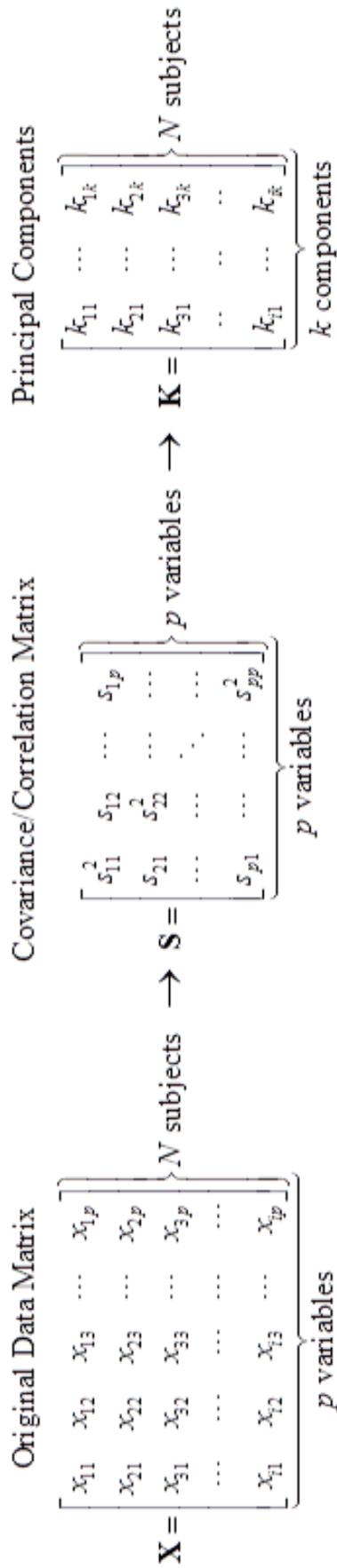
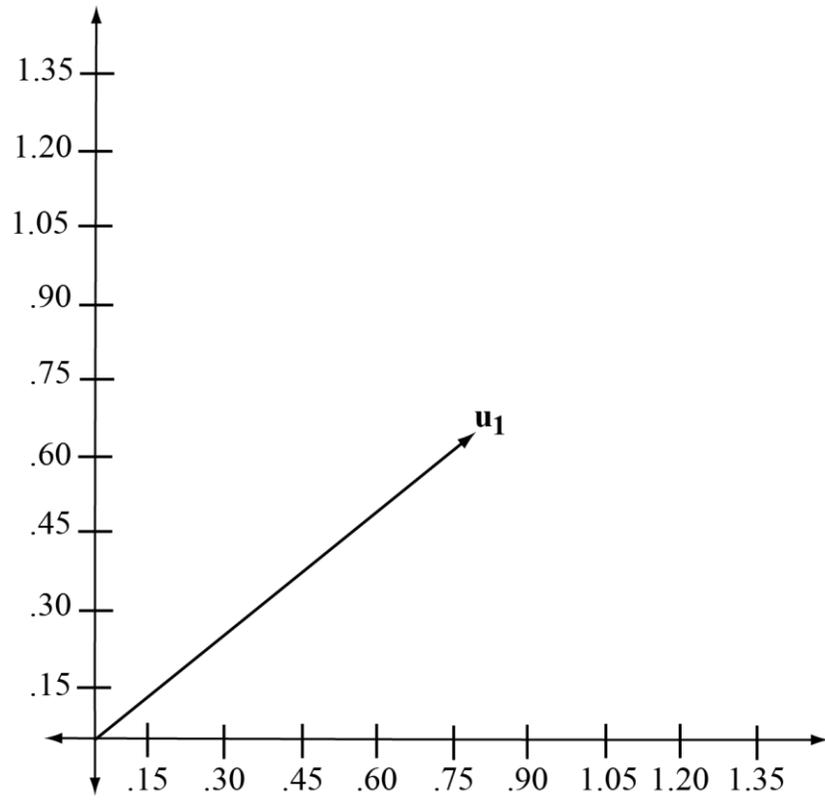
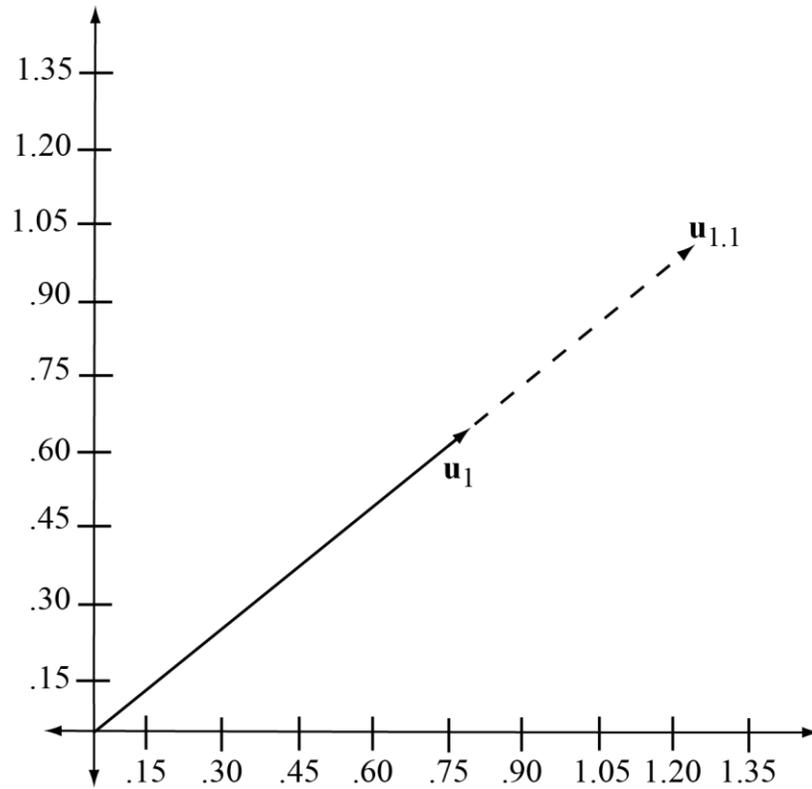


Figure 37. Illustration of data transformation process from the original data matrix to a covariance/correlation matrix and then to a set of new variables called principal components.



*Figure 38.* Illustration of vector  $\mathbf{u}_1$  as a trajectory in 2 dimensional space.



*Figure 39.* Illustration of scalar multiplication of vector  $\mathbf{u}_1$  that results in vector  $\mathbf{u}_{1.1}$ , a change in the length of vector  $\mathbf{u}_1$ .

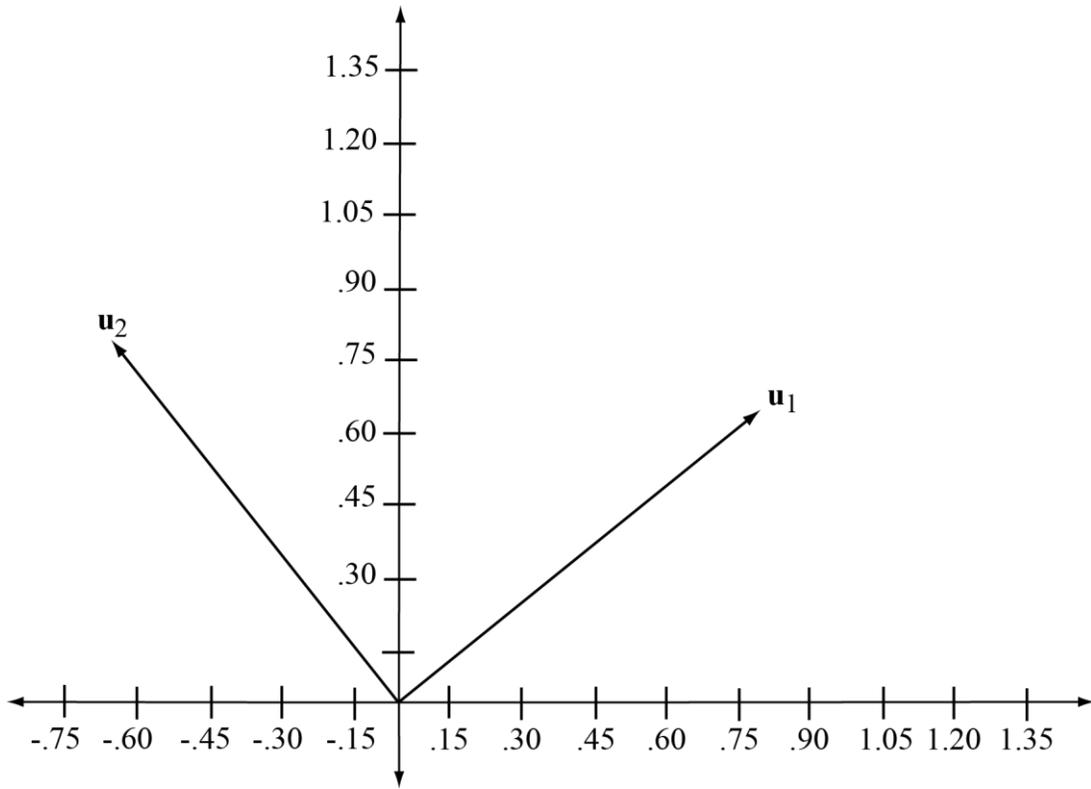
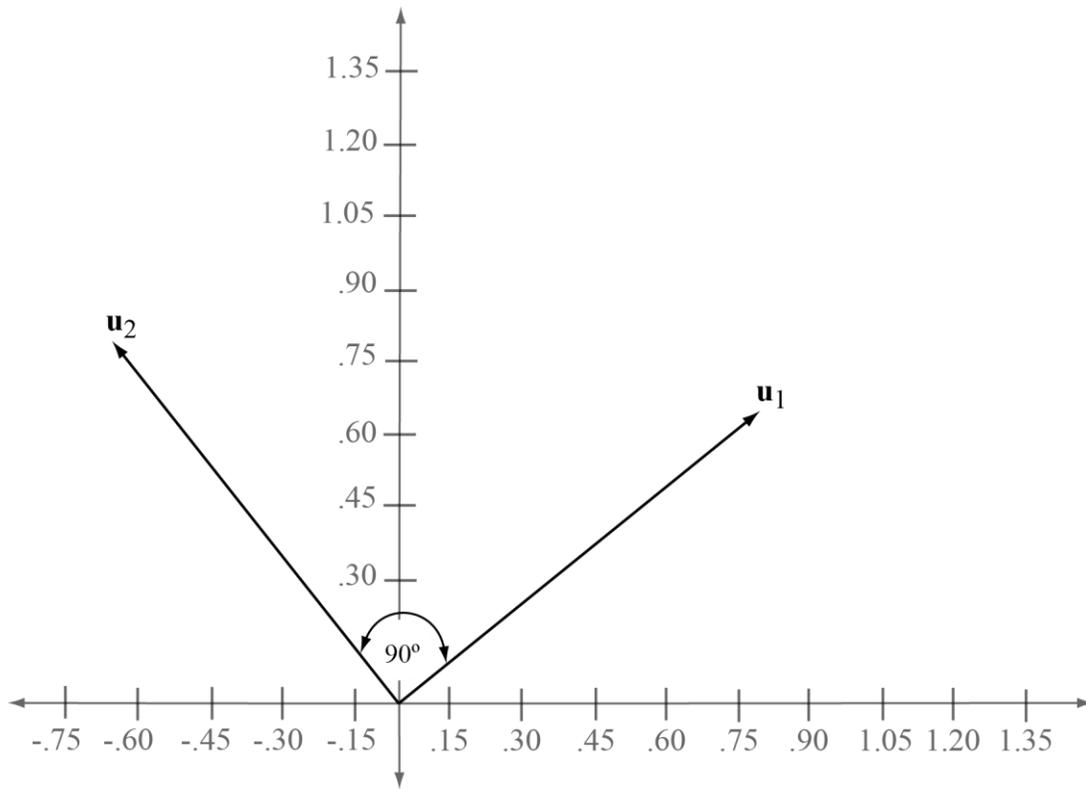
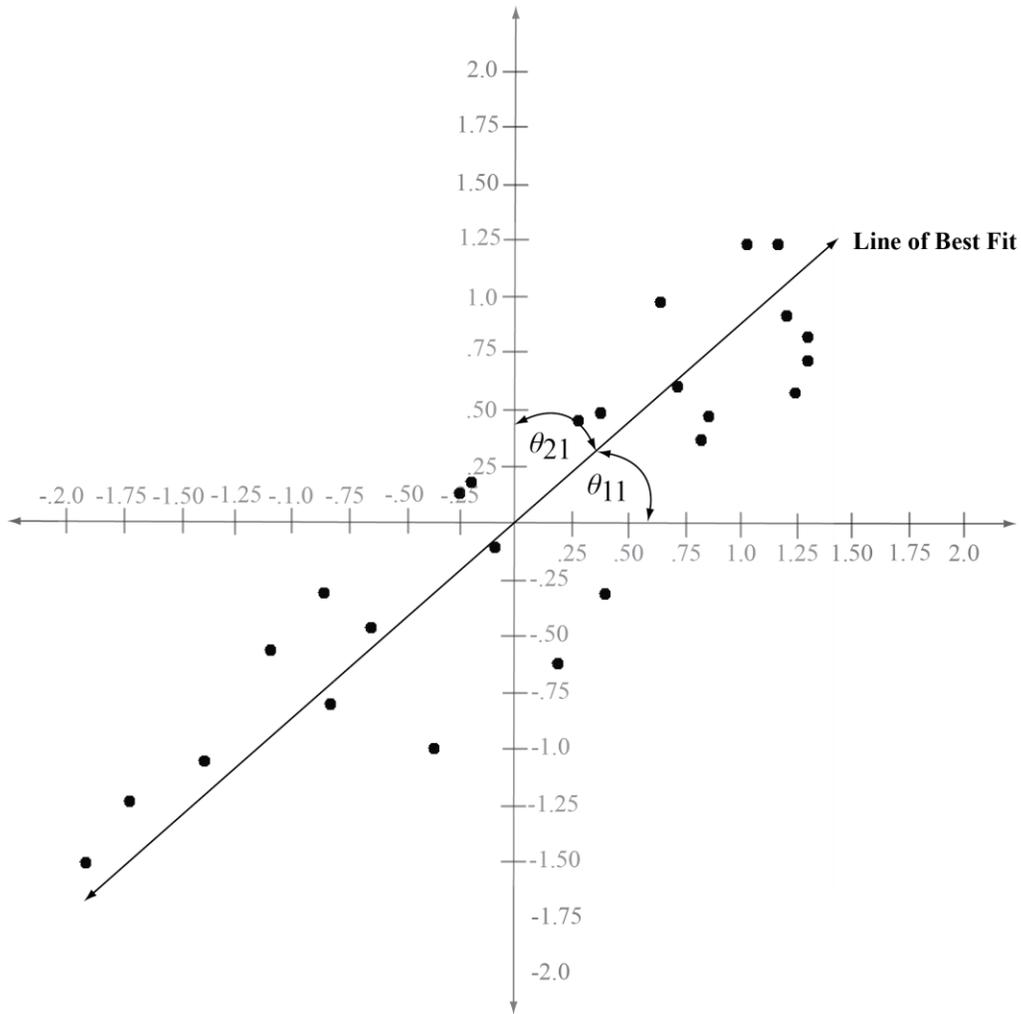


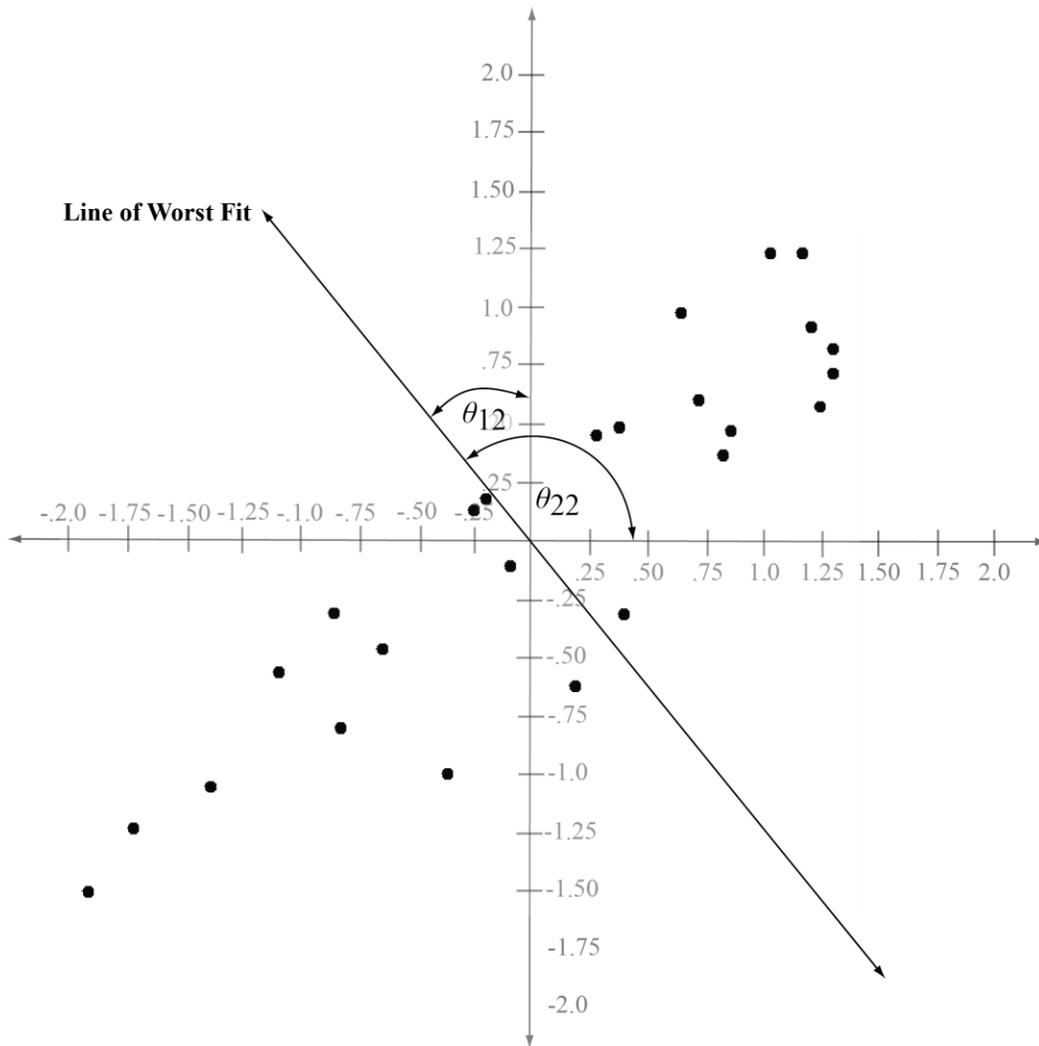
Figure 40. Geometric illustration of vector  $\mathbf{u}_1$  and vector  $\mathbf{u}_2$ .



*Figure 41.* Geometric illustration of the angle formed between the perpendicular vector  $\mathbf{u}_1$  and vector  $\mathbf{u}_2$ .



*Figure 42.* Geometric representation of the angles of the rotated axes relative to the original scatter plot's x y coordinate positions. The fitted "Line of Best Fit" represents the first principal component of the 2x2 covariance matrix  $\mathbf{S}$  from the data in Table 10.



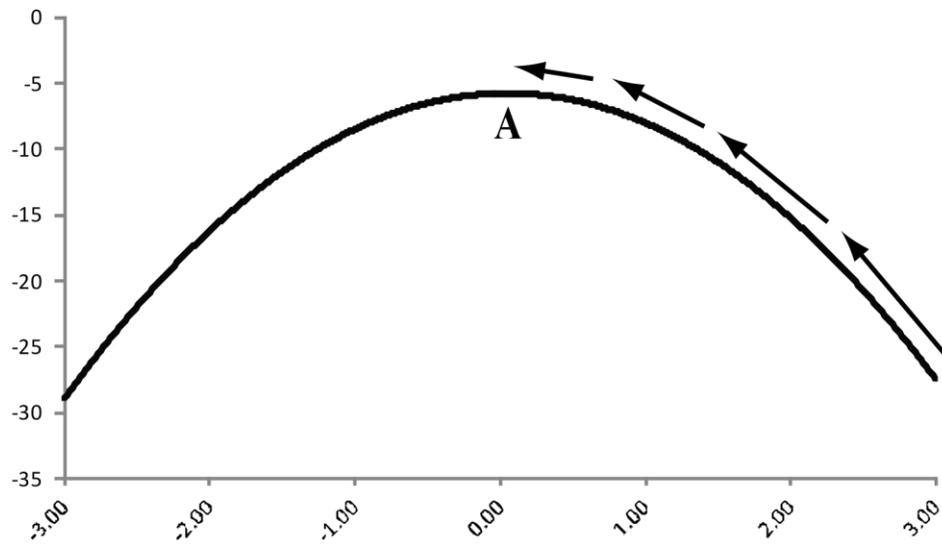
*Figure 43.* Geometric representation of the angles of the rotated axes relative to the original scatter plot's x y coordinate positions. The fitted "Line of Worst Fit" represents the second principal component of the 2x2 covariance matrix  $\mathbf{S}$  from the data in Table 10.

## Principal Components

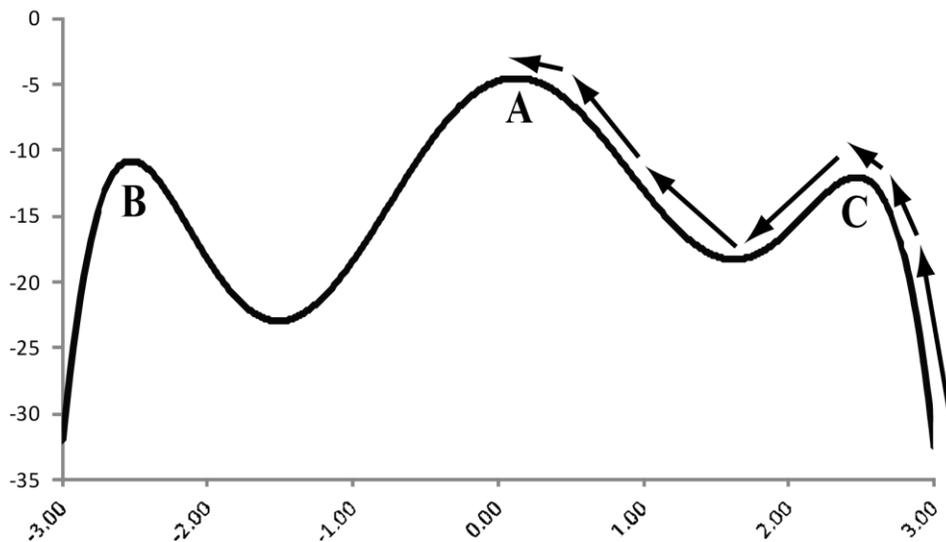
$$\mathbf{k} = \underbrace{\begin{bmatrix} k_{11} & \cdots & k_{1k} \\ k_{21} & \cdots & k_{2k} \\ k_{31} & \cdots & k_{3k} \\ \cdots & \cdots & \cdots \\ k_{i1} & \cdots & k_{ik} \end{bmatrix}}_{\text{components}} \left. \vphantom{\begin{bmatrix} k_{11} & \cdots & k_{1k} \\ k_{21} & \cdots & k_{2k} \\ k_{31} & \cdots & k_{3k} \\ \cdots & \cdots & \cdots \\ k_{i1} & \cdots & k_{ik} \end{bmatrix}} \right\} \text{scores}$$

Figure 44. Illustration of  $k$  principal components as new variables that consist of  $i$  scores.

### (A) Likelihood Function (unimodal)



### (B) Likelihood Function (trimodal)



*Figure 45.* A graphical depiction of a likelihood function to depict convergence failure. Panel A illustrates a typical unimodal likelihood function with one likely solution denoted as A. Panel B shows a trimodal likelihood function with three possible likelihood solutions denoted A, B, and C which correspond to different MLE of the mean.

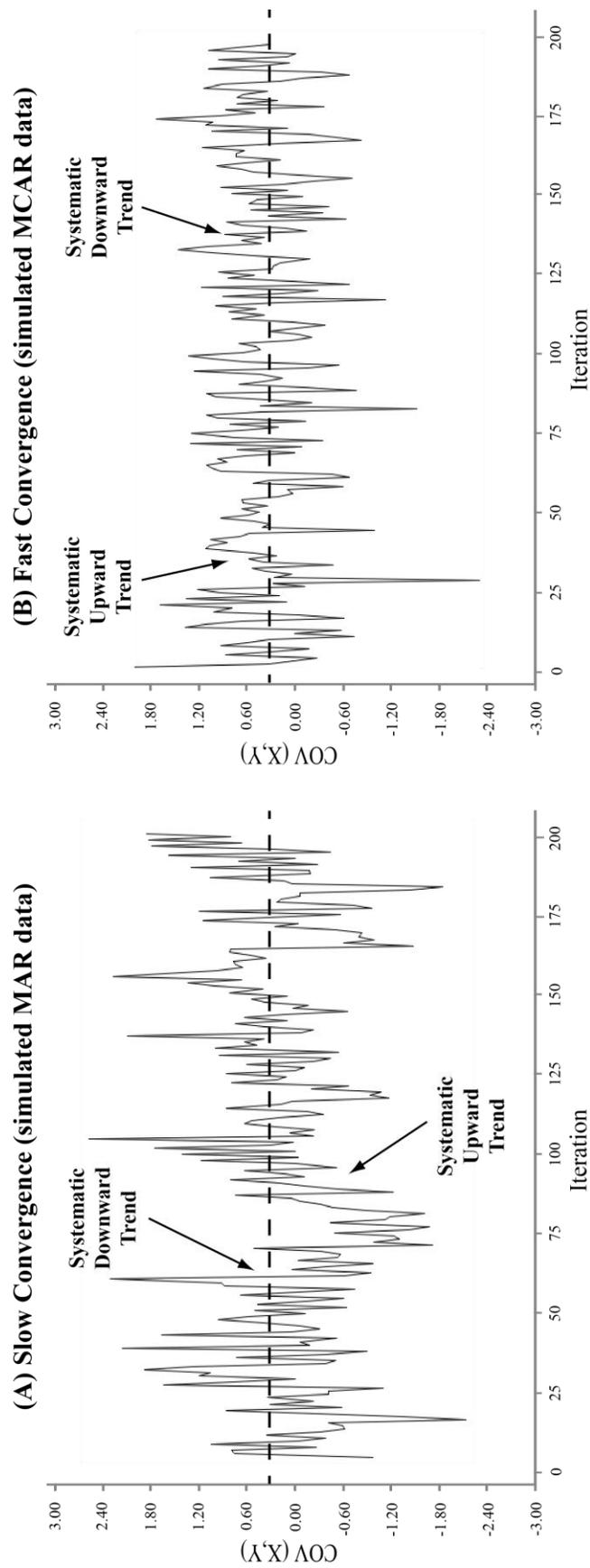


Figure 46. A graphical depiction of the data augmentation stability. Let the dashed line represent a reference line for the population value Panel A (simulated MAR data) represents relatively more instability than Panel B (simulated MCAR data). More specifically, note that the estimated data in Panel A contains larger systematic downward and upward trends relative to the estimated data in Panel B.

	$X$	$Y$	$Aux_1$	$\dots$	$Aux_p$
1					
2					
3					
4					
5					

Figure 47. Exemplar general missing data pattern with missing on all variables except X.

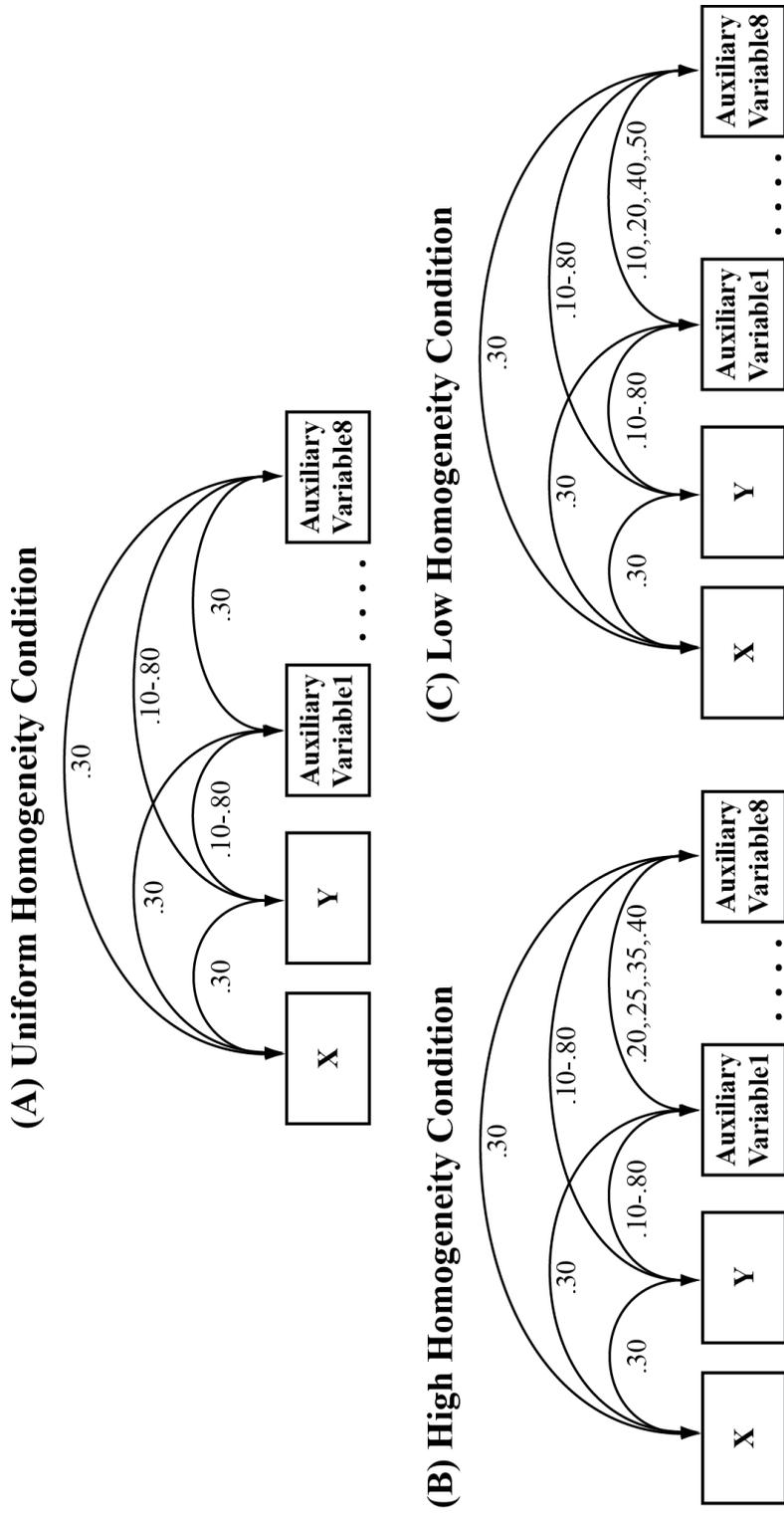
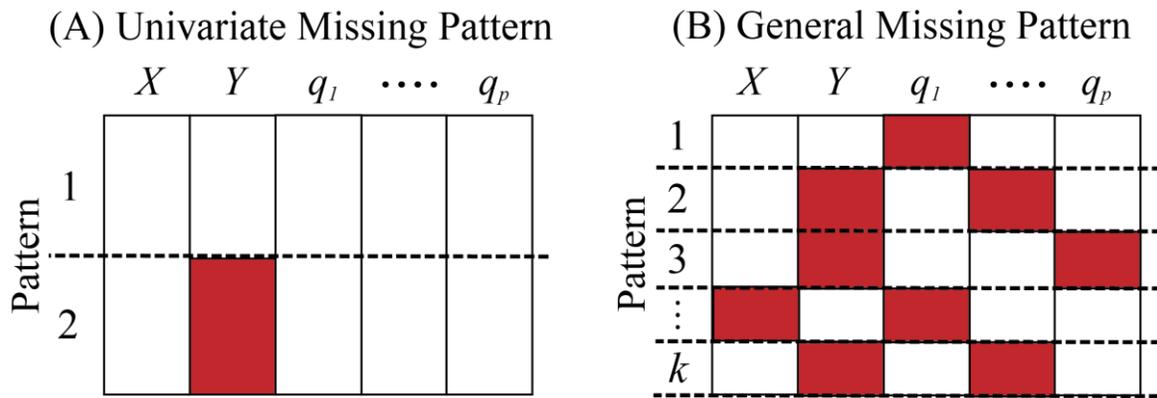


Figure 48. Illustration of the population model for the three homogeneity conditions. Panel A represents a population model with a uniform correlation of  $\rho = .30$  among the auxiliary variables. Panel B shows a high homogeneity condition with correlation magnitudes divided into fourths with  $\rho = .20, .25, .35,$  and  $.40$ . Panel C demonstrates the low homogeneity condition with correlation magnitudes divided into fourths with  $\rho = .10, .20, .40,$  and  $.50$ .



*Figure 49.* Illustration of the missing data patterns investigated across  $X$ ,  $Y$  and  $p$  auxiliary variables (note that some conditions include squared and cubed terms) where missingness is shown as shaded areas. Panel A displays a univariate missing data pattern with missingness only on  $Y$  (thus only two patterns of observed or missing data). Panel B shows a general missing data pattern with missingness scattered throughout the data in a seemingly random fashion (with  $k$  missing data patterns).

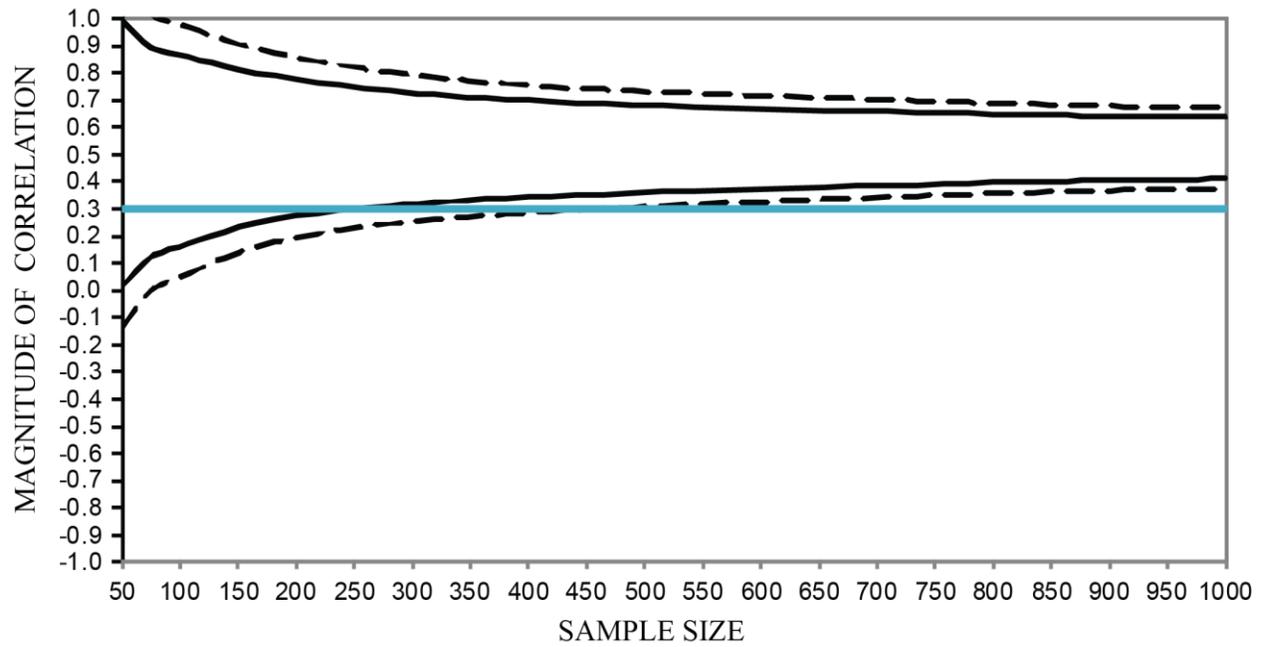


Figure 50. Illustration of simulation results showing bias of the correlation between  $X$  and  $Y$  associated with not including a non-linear cause of missingness with a 95% (dashed black line) and a 99% (solid black line) confidence interval associated with a 60% non-linear MAR mechanism where no auxiliary variables are included. Note that the solid blue line is a reference to the population correlation of  $\rho = .30$  between  $X$  and  $Y$ .

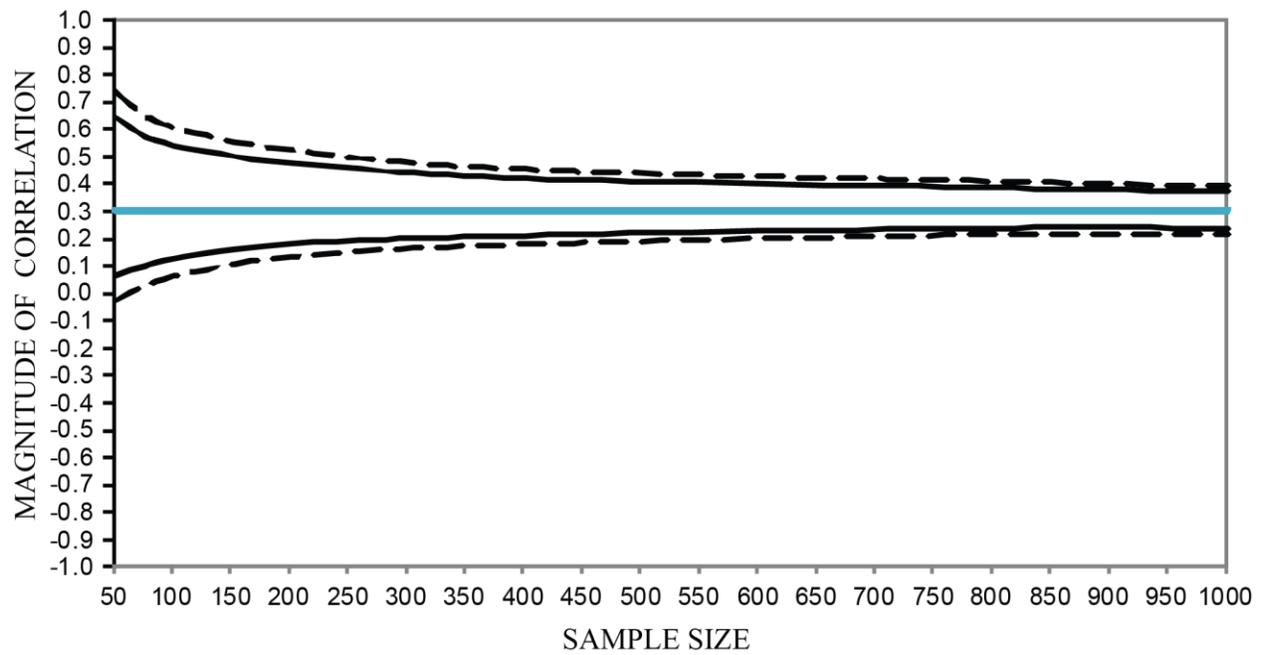


Figure 51. Illustration of simulation results showing the correlation between  $X$  and  $Y$  where the linear components (i.e.,  $AUX_1$  and  $AUX_2$ ) of the interaction term (i.e.,  $AUX_1 * AUX_2$ ) that caused missingness are included with a 95% (dashed black line) and a 99% (solid black line) confidence interval associated with a 60% non-linear MAR mechanism where all linear auxiliary variables are included. Note that the solid blue line is a reference to the population correlation of  $\rho = .30$  between  $X$  and  $Y$ .

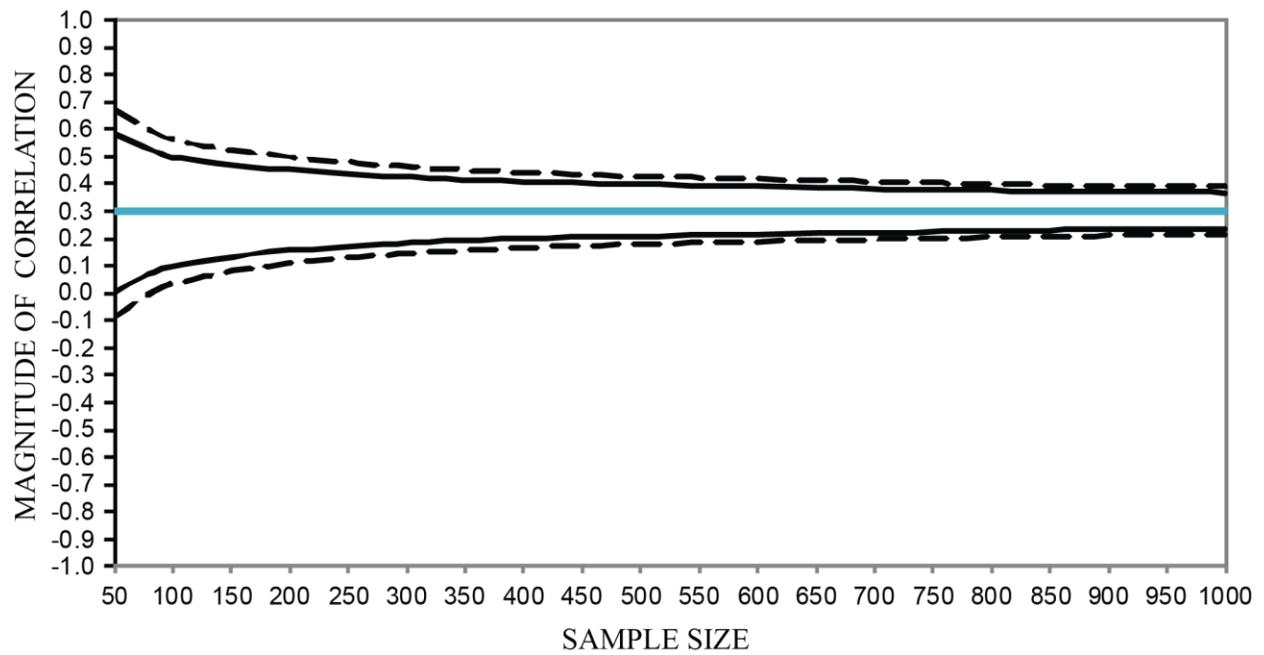
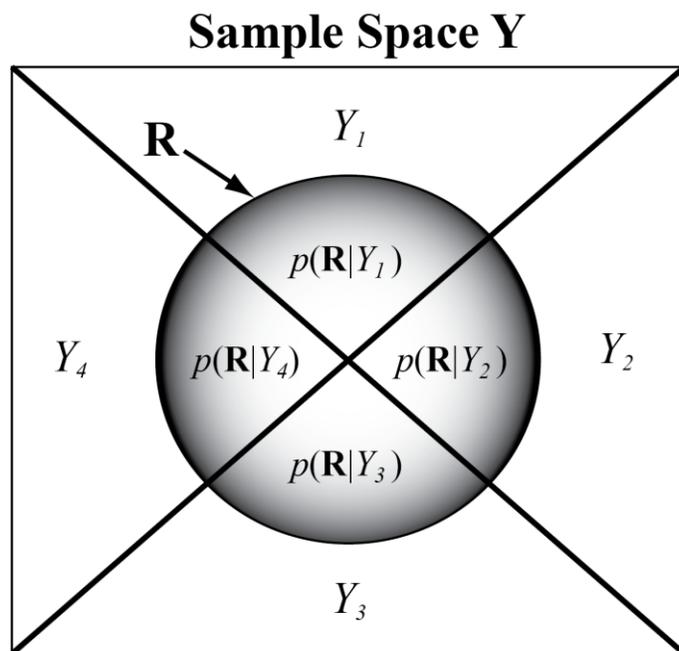


Figure 52. Illustration of simulation results showing the correlation between  $X$  and  $Y$  where the  $PCA_{AUX}$  method is used with a 95% (dashed black line) and a 99% (solid black line) confidence interval associated with a 60% non-linear MAR mechanism where a PCA auxiliary variable which contains nonlinear information is included. Note that the solid blue line is a reference to the population correlation of  $\rho = .30$  between  $X$  and  $Y$ .

$$\begin{array}{cc}
\text{(A) raw data with missingness} & \text{(B) rows with complete data only} \\
\mathbf{Y} = \begin{bmatrix} 21.15 \\ Y_{miss} \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} & \mathbf{X} = \begin{bmatrix} 21.14 \\ 20.22 \\ 20.81 \\ 20.69 \\ 19.59 \\ 18.33 \\ 15.40 \\ 20.34 \end{bmatrix} & \rightarrow & \mathbf{Y}_{obs} = \begin{bmatrix} 21.15 \\ 20.40 \\ 19.96 \\ 17.21 \\ 19.29 \\ 16.91 \\ 18.06 \end{bmatrix} & \mathbf{X}_{obs} = \begin{bmatrix} 21.14 \\ 20.81 \\ 20.69 \\ 19.59 \\ 18.33 \\ 15.40 \\ 20.34 \end{bmatrix}
\end{array}$$

*Figure 53.* The  $Y$  and  $X$  vectors for the Yates method example. Panel A demonstrates the raw data with a missing values denoted  $Y_{miss}$ . Panel B contains only associated values (i.e., rows) that are observed. That is,  $\mathbf{Y}_{obs}$  does not contain  $Y_{miss}$  and  $\mathbf{X}_{obs}$  does not contain the value 20.22 as it is in the same row (from the same case) as  $Y_{miss}$ .



**(A) Independent joint probabilities**

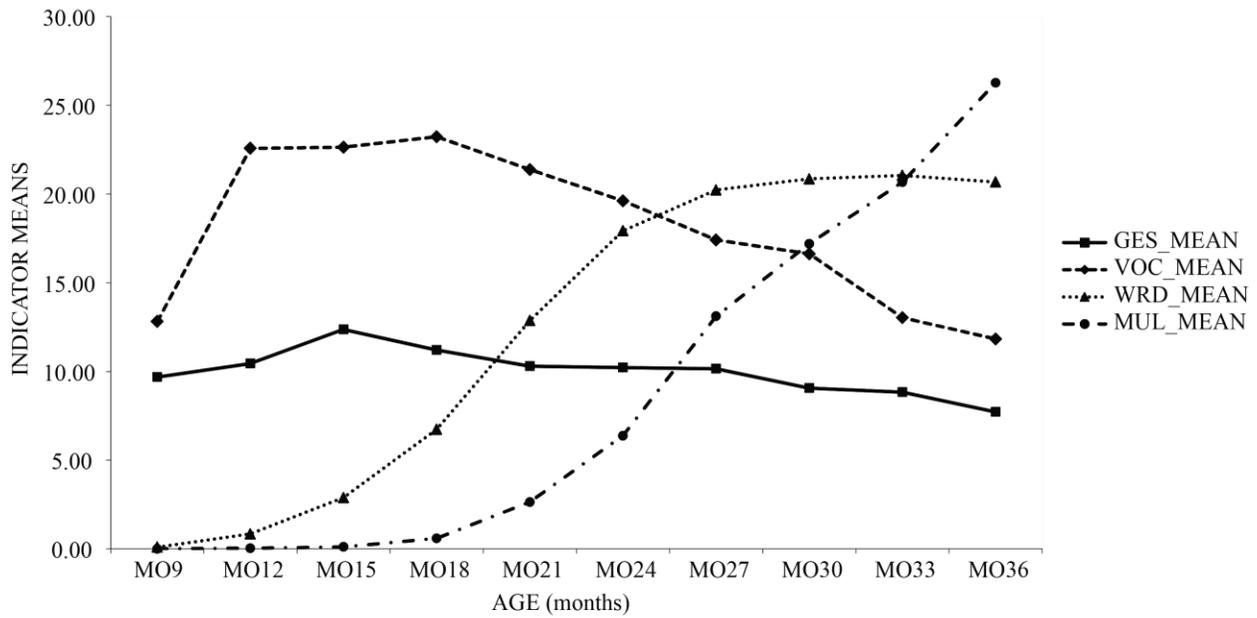
Variables	Missing Data	
	$p(R=1) = .15$	$p(R=0) = .85$
$p(Y_1) = .25$	$p(Y_1)p(R=1)$ $(.25)(.15)=.034$	$p(Y_1)p(R=0)$ $(.25)(.85)=.213$
$p(Y_2) = .25$	$p(Y_2)p(R=1)$ $(.25)(.15)=.034$	$p(Y_2)p(R=0)$ $(.25)(.85)=.213$
$p(Y_3) = .25$	$p(Y_3)p(R=1)$ $(.25)(.15)=.034$	$p(Y_3)p(R=0)$ $(.25)(.85)=.213$
$p(Y_4) = .25$	$p(Y_4)p(R=1)$ $(.25)(.15)=.034$	$p(Y_4)p(R=0)$ $(.25)(.85)=.213$

**(B) Non-independent joint probabilities**

Variables	Missing Data	
	$p(R=1) = .15$	$p(R=0) = .85$
$p(Y_1) = .25$	$p(Y_1)p(R=1)$ $= .06$	$p(Y_1)p(R=0)$ $= .19$
$p(Y_2) = .25$	$p(Y_2)p(R=1)$ $= .02$	$p(Y_2)p(R=0)$ $= .23$
$p(Y_3) = .25$	$p(Y_3)p(R=1)$ $= .03$	$p(Y_3)p(R=0)$ $= .22$
$p(Y_4) = .25$	$p(Y_4)p(R=1)$ $= .04$	$p(Y_4)p(R=0)$ $= .21$

Figure 54. A graphical depiction of joint events in a sample space (left). Panel A illustrates a table of probabilities associated with independent joint events. Panel B demonstrates a table of probabilities associated with non-independent joint events. This image shows the relationship between the missing data probability distribution **R**, and the associated data probability distribution **Y**.

(A) Key Skills Indicator Means (3 - Month Intervals) Multiple Imputation via the AUX Method



(B) Key Skills Indicator Means (3 - Month Intervals) Multiple Imputation via the PCAaux method

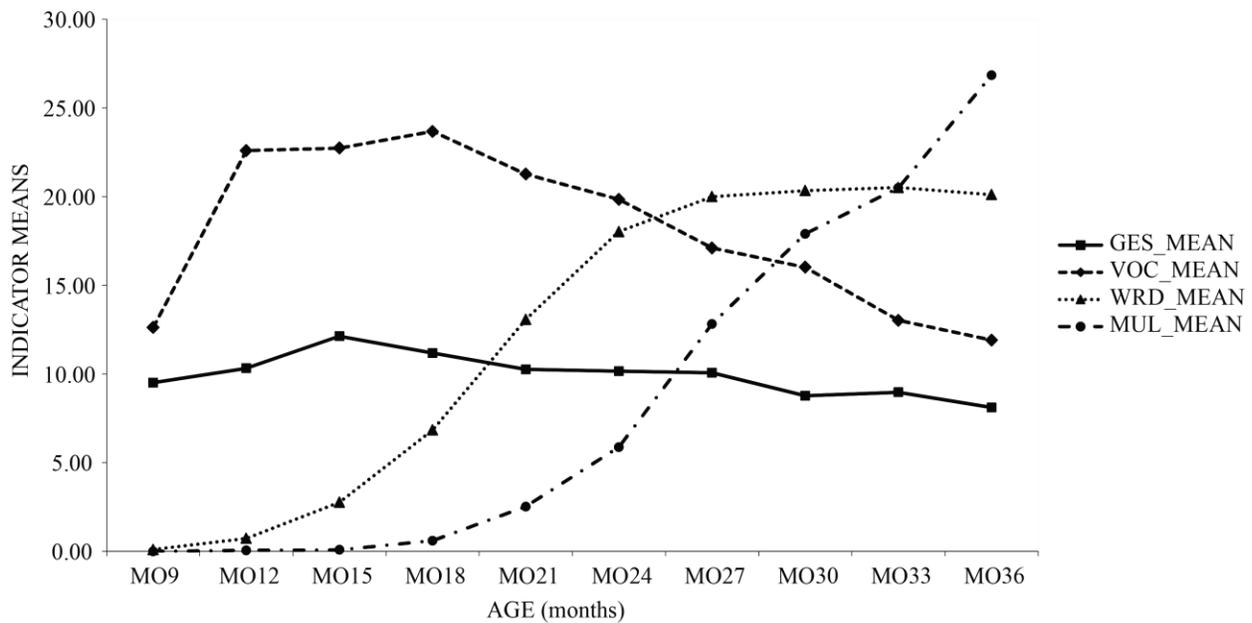


Figure 55. Mean plots of the ECI key skills from the empirical data example. Panel A demonstrates the key skills means across 3 – month intervals based on multiple imputation via the AUX method. Panel B contains the key skills means across 3 – month intervals based on multiple imputation via the PCAux method. Note that GES\_MEAN = the mean of gestures,

VOC\_MEAN = the mean of vocalizations, WRD\_MEAN = the mean of words, MUL\_MEAN = the mean of multiple words.