Examining How Missing Data Affect Approximate Fit Indices in Structural Equation Modelling Under Different Estimation Methods

by

Xijuan Zhang

B.A., University of British Columbia, 2012

M.A., University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

The University of British Columbia
(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Examining How Missing Data Affect Approximate Fit Indices in Structural Equation Modelling Under Different Estimation Methods

submitted by **Xijuan Zhang** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Psychology**.

Examining Committee:

Victoria Savalei, Professor, Psychology, UBC
Supervisor

Jeremy Biesanz, Associate Professor, Psychology, UBC
Supervisory Committee Member

Harry Joe, Professor, Statistics, UBC
Supervisory Committee Member

Lang Wu, Professor, Statistics, UBC
University Examiner

Brian O'Connor, Professor, Psychology, UBC
University Examiner

Abstract

The full-information maximum likelihood (FIML) is a popular estimation method for missing data in structural equation modeling (SEM). However, it is not commonly known that approximate fit indices (AFIs) can be distorted, relative to their complete data counterparts, when FIML is used to handle missing data. In the first part of the dissertation work, we show that two most popular AFIs, the root mean square error of approximation (RMSEA) and the comparative fit index (CFI), often approach different population values under FIML estimation when missing data are present. By deriving the FIML fit function for incomplete data and showing that it is different from the usual maximum likelihood (ML) fit function for complete data, we provide a mathematical explanation for this phenomenon. We also present several analytic examples as well as the results of two large sample simulation studies to illustrate how AFIs change with missing data under FIML. In the second part of the dissertation work, we propose and examine an alternative approach for computing AFIs following the FIML estimation, which we refer to as the FIML-Corrected or FIML-C approach. We also examine another existing estimation method, the two-stage (TS) approach, for computing AFIs in the presence of missing data. For both FIML-C and TS approaches, we also propose a series of small sample corrections to improve the estimates of AFIs. In two simulation studies, we find that the FIML-C and TS approaches, when implemented with small sample corrections, can estimate the complete data population AFIs with little bias across a variety of condi-

tions, although the FIML-C approach can fail in a small number of conditions with a high percentage of missing data and a high degree of model misspecification. In contrast, the FIML AFIs as currently computed often performed poorly. We recommend FIML-C and TS approaches for computing AFIs in SEM.

Lay Summary

In a survey study, participants often leave some questions blank due to carelessness or unwillingness to answer certain questions. This creates missing data, which can distort the results of statistical analyses. Modern missing data techniques, such as the full-information maximum likelihood (FIML) method, are designed to address this problem of missing data. However, the FIML method corrects the distorted results in some circumstances but not in all.

In this dissertation, we focus on structural equation modelling (SEM), which is an advanced statistical analysis commonly used in social sciences. We explain that in SEM, the FIML method may distort the results regarding the degree of model fit as measured by the approximate fit indices (AFIs). We propose two alternative methods called the FIML-Corrected (FIML-C) and the two-stage (TS) methods. Through computer simulations, we show these two new methods can correctly compute the AFIs; therefore, we recommend these two new methods.

Preface

I am the primary author of this PhD dissertation. I was the primary individual responsible for conducting the simulation studies involved in this research. The theoretical work in this research is a joint work with Dr. Victorial Savalei. A version of Chapters 2 to 4 has been published in the *Structural Equation Modelling* journal with the following bibliographic details:

> Zhang, X., & Savalei, V. (2019). Examining the effect of missing data on RMSEA and CFI under the normal theory full information maximum likelihood. Structural Equation Modeling, 27, 219-239.

A version of Chapters 5 and 6 have been submitted for publication.

Contents

Abstract . . . . . **iii**

Lay Summary . . . . . **v**

Preface . . . . . **vi**

Contents . . . . . **vii**

List of Tables . . . . . **x**

List of Figures . . . . . **xi**

List of Supplementary Materials . . . . . **xv**

List of Abbreviations . . . . . **xvi**

Acknowledgments . . . . . **xvii**

Dedication . . . . . **xix**

1 Introduction **1**

1.1 Missing Data Mechanisms 2

1.2 Missing Data Patterns 6

1.3 Missing Data Techniques 8

1.4 SEM and SEM AFIs 10
1.5 Past Research on the Effect of SEM AFIs under FIML Estimation 13
1.6 Dissertation Organization 15

2 SEM AFIs under FIML Estimation: Technical Details 17
2.1 Fit Function Minimum for Complete Data 18
2.2 Fit Function Minimum for Incomplete Data 20
2.2.1 Sample Fit Function Minimum for Incomplete Data 20
2.2.2 Population Fit Function Minimum for Incomplete Data 25
2.3 RMSEA and CFI for Complete and Incomplete Data 29

3 SEM AFIs under FIML Estimation: Analytical Examples 32
3.1 Change in RMSEA due to Differences in the Equations of the Fit Function Minimum 33
3.1.1 Case 1: Complete data 34
3.1.2 Case 2: MCAR data; misspecification does not involve variables with missing values 35
3.1.3 Case 3: MCAR data; misspecification involves variables with missing data 36
3.1.4 Case 4: MAR data; misspecification involves variables with missing values 37
3.2 Change in RMSEA due to Differences in Parameter Values 39
3.2.1 Case 1: Pseudo-parameter values stay the same with missing data 40
3.2.2 Case 2: Pseudo-parameter values change with missing data . . . 42

4 SEM AFIs under FIML Estimation: Simulation Studies 45
4.1 Design 46
4.2 Results 50

4.2.1 Study 1 . . . 51
4.2.2 Study 2 . . . 60
4.3 Discussion . . . 64

5 Alternative Approaches for Computing AFIs . . . 67
5.1 Alternative AFIs following FIML Estimation . . . 68
5.1.1 Population Limits for FIML-C AFIs . . . 72
5.1.2 Analytical Example for FIML-C Estimation . . . 74
5.2 Alternative AFIs following TS Estimation . . . 75
5.2.1 Population Values for TS AFIs . . . 77
5.2.2 Analytical Example for TS Estimation . . . 78
5.2.3 Derivation of Small Sample Correction in FIML-C . . . 81
5.2.4 Derivation of Small Sample Correction in TS . . . 81

6 SEM AFIs under FIML-C and TS Estimations: Simulation Studies . . . 83
6.1 Design . . . 84
6.2 Results . . . 85
6.2.1 Population Behavior . . . 85
6.2.2 Finite Sample Behavior . . . 91
6.3 Discussion . . . 101

7 Conclusion and Overall Discussion . . . 104
7.1 Limitations and Future Directions . . . 108

Bibliography . . . 114

List of Tables

Table 2.1 Notation for mean vectors and covariance matrices with incomplete data under an incorrect hypothesized model. 21

Table 4.1 Conditions in the Simulation Studies 47

Table 4.2 Complete data RMSEA and CFI for all conditions in Studies 1 and 2 . 50

Table 4.3 Variables in the Regression Analyses 59

Table 4.4 Results of the Regression Analyses 60

Table 5.1 Equations for k and k_B for FIML-C versions 73

Table 5.2 Equations for c and c_B for TS versions 80

Table 6.1 "Pseudo-Parameters" for Complete and Incomplete Data under FIML 89

Table 6.2 Additional Variables in the Regression Analyses 90

Table 6.3 Results of the Regression Analyses for Bias in the Population 90

Table 6.4 Results of the Regression Analyses for Bias in the Finite Samples . . . 100

List of Figures

Figure 1.1 An example of an SEM model. In SEM diagram, circles represent latent variables, rectangles represent observed variables, and the arrows represent relationship between variables. The relationships between the observed variables and the latent factor for the measurement model part of SEM model (see the blue box for an example); the relationships between different latent variables form the structural part of SEM model (see the red box for an example). 11

Figure 1.2 The population model for the simulation example in section 1.5. 13

Figure 4.1 Differences between DF and SF conditions. 48

Figure 4.2 RMSEA and CFI for Study 1 conditions varying in the locations of misfit and number of variables with missing data. For the conditions in this figure, the missing mechanism is MCAR, the population factor correlation is 0, and the number of correlated residuals is 2. 52

Figure 4.3 Fit function minima of the hypothesized and baseline models for selected conditions in Study 1. For the conditions in this figure, the missing mechanism is MCAR, the population factor correlation is zero, and the number of correlated residuals is two. 54

Figure 4.4 RMSEA and CFI for selected conditions in Study 1. For the conditions in this figure, the missing mechanism is MCAR. There is a single correlated residual in the population model, and the number of variables with missing data is four. 56

Figure 4.5 RMSEA and CFI for selected conditions in Study I. For conditions in this figure, the population factor correlation is 0.4; the number of correlated residuals is two, and the number of variables with missing data is four. 57

Figure 4.6 RMSEA and CFI for selected conditions in Study 2. For the conditions shown in this figure, the number of variables with missing data is six. 62

Figure 6.1 Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 1 comparing FIML, FIML-C and TS approaches. Complete data population RMSEA and CFI are also included for comparison. In these selected conditions, the number of variables with missing data is four, the number of correlated residuals is two, and the population factor correlation is zero. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals. 86

Figure 6.2 Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 2 comparing FIML, FIML-C and TS approaches. Complete data population RMSEA and CFI are also included for comparison. In these selected conditions, there are six variables that have missing data. The population model is a two-factor model with varying sizes for the factor correlation shown on the x-axis. The hypothesized model is a one-factor model. 87

Figure 6.3 Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing FIML, FIML-C and TS approaches. In these conditions, the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, the percentage of missing is 50%, and the location of misfit is the same as the location of missing data. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals. 92

Figure 6.4 Root mean square error (RMSE) in the sample RMSEA and CFI estimates for selected conditions in Study 2 comparing FIML, FIML-C and TS approaches. In these conditions, the number of variables with missing data is six, the percentage of missing is 50%, and the number of patterns is large. The population model is a two-factor model with varying sizes for the factor correlation shown on the *x*-axis. The hypothesized model is a one-factor model. 93

Figure 6.5 Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing among the best performing FIML-C and TS methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the *x*-axis. The hypothesized model is a two-factor model without any correlated residuals. 96

Figure 6.6 Bias in the sample RMSEA and CFI estimates for selected conditions in Study 2 comparing among the best performing FIML-C and TS methods. In these conditions, the number of variables with missing data is six and the missing mechanism is strong MAR. The population model is a two-factor model with varying sizes for the factor correlation shown on the *x*-axis. The hypothesized model is a one-factor model. . . 97

Figure 6.7 Root mean square error (RMSE) in sample RMSEA and CFI for selected conditions in Study 1 comparing among the best performing FIML-C and TS methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the *x*-axis. The hypothesized model is a two-factor model without any correlated residuals. 98

Figure 6.8 Root mean square error (RMSE) in sample RMSEA and CFI for selected conditions in Study 2 comparing among the best performing FIML-C and TS methods. In these conditions, the number of variables with missing data is six and the missing mechanism is strong MAR. The population model is a two-factor model with varying sizes for the factor correlation shown on the x-axis. The hypothesized model is a one-factor model. 99

Figure 7.1 Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 1 comparing FIML, FIML-C, TS, MI approaches. In these selected conditions, the number of variables with missing data is four, the number of two correlated residuals is two, and the population factor correlation is zero. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals. 111

Figure 7.2 Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing among FIML, FIML-C, TS and MI methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals. 112

List of Supplementary Materials

- Tables for the results of simulation studies.
- Sample code for generating missing data for the simulation studies.
- Sample code for computing FIML-C RMSEA and CFI.
- Sample code for computing TS RMSEA and CFI.

For online access of the Supplementary Materials, please visit the following webpage:

https://osf.io/rtp38/?view_only=15d7262e78ca4f018f4deb8b47307e5a

List of Abbreviations

- AFI: Approximate Fit Indices
- CFI: Comparative Fit Index
- EM: Expectation-Maximization
- FIML: Full Information Maximum Likelihood
- FIML-C: FIML-Corrected
- LR: Likelihood Ratio
- MAR: Missing At Random
- MCAR: Missing Completely At Random
- MCMC: Markov Chain Monto Carlo
- MI: Multiple Imputation
- MNAR: Missing Not At Random
- RMSEA: Root Mean Square Error of Approximation
- RMSE: Room Mean Square Error
- SEM: Structural Equation Modelling
- TS: Two-Stage

Acknowledgments

First and foremost, I thank my supervisor, Dr.Victoria Savalei. She has helped me in every aspect of my academic career. This dissertation work would not have been remotely possible without her support and dedication. Her advice, guidance and commitment have been invaluable for me. I am also very grateful for the financial support she has granted me throughout my graduate studies at UBC, which allows me to pursue my dream in academia while living a comfortable life. Beyond vocational or academic matters, I thank Dr.Savalei for her understanding and support during some of the difficult times in my person life. All in all, I consider myself extremely lucky to have Dr. Savalei as my graduate supervisor.

Secondly, I would also like to thank Dr. Jeremy Biesanz, Dr. Harry Joe and Dr. Ke-Hai Yuan. Specifically, I thank Dr. Jeremy Biesanz for introducing me to quantitative psychology in PSYC 359, for helping me out with various issues I had met during my graduate school years, and for the time and effort for being on my master and PhD committee. I thank Dr. Harry Joe for teaching me STAT 306, and for being on my PhD committee and writing me a research note after my dissertation proposal defense. Dr. Joe's research note and course package for STAT 306 had helped me understand many of key concepts in statistics. I thank Dr. Ke-Hai Yuan for giving the opportunity to visit him at the University of Notre Dame; Dr. Ke-Hai Yuan's advice on my dissertation work had helped me design the simulation studies in my research.

I also express my gratitude towards Dr. Oscar Olvera, who is both my friend and my

mentor. He has helped me both personally and professionally. Based on his suggestions, I took a course on mathematical proofs and another course on real analysis, both of which had greatly improved my mathematical proficiency. His humor and support had made some of the difficult times in my life more bearable.

Finally, I would like to thank my course professors, Dr. Lang Wu (STAT 300), Dr. William Welch (STAT 305), Dr. Matias Salibian-Barrera (STAT 406), and Dr. Gordon Slade (MATH 320), Dr. Brett Kolesnik (MATH 220) and Dr. Anna Levit (MATH 302), for teaching me statistics and mathematics. The courses I took with these professors had given me a lot of insight and inspiration for my own research.

This research was supported by the Social Sciences and Humanities Research Council's Doctoral Fellowship to Xijuan Zhang, the Univeristy of British Columbia's Four Year Doctoral Fellowship to Xijuan Zhang, and the Natural Sciences and Engineering Research Council's research grant to Dr. Victoria Savalei.

Dedication

I dedicate my dissertation work to the most important teachers I have met in my life: Mr. Bill Morphett (high school science teacher), Dr. Victoria Savalei (graduate school supervisor), Dr. Corey Hamm (piano teacher), and Dr. Ross Salvosa (piano teacher).

Mr. Bill Morphett's encouragement has given me the courage to pursue my dream. His kindness and compassion as a teacher has made me also want to become a teacher.

Dr. Victoria Savalei is the most influential female role model in my life. Her intelligence, independence, kindness and dedication to research have inspired me to become a female figure like her.

Dr. Corey Hamm and Dr. Ross Salvosa have not only taught how to play the piano but more importantly, how to understand and appreciate music. Music holds a very special place in my heart and so do Dr. Hamm and Dr. Salvosa.

Chapter 1

Introduction

> Missing data are a real bane to researchers across all social science disciplines. For most of our scientific history, we have approached missing data much like a doctor from the ancient world who might use bloodletting to cure disease or amputation to stem infection (e.g, removing the infected parts of one's data by using list-wise or pair-wise deletion).
>
> *Craig K.Enders*

Missing data, also known as incomplete data, are prevalent in psychological and educational research, particularly when repeated measures or longitudinal studies are involved. Osborne [29] reported that around 40% of papers in APA journals in the year 2009 described dealing with missing data. Jelicic et al. [18] examined the prevalence of missing data in longitudinal studies over six years in three developmental psychology journals, and found about half of the studies had missing data.

Historically, statistical analysis methods are developed assuming that data are complete, and proper methodology for dealing with incomplete data is hard to implement due to intensive computations. However, with the advance of computing technology, beginning in the late 1980s, more and more researchers began to study the problem of missing

data. According to Google Scholar, the number of articles with titles including the words *missing data* or *incomplete data* were 1024 in years 1990 to 1999, grew to 3505 in the years 2000–2009, and is 5400 in the past 8 years.

In addition, with the improvement in computing technology, more advanced modeling methods are made available for researchers to use. Structural equation modeling (SEM) is one of these advanced methods; it allows researchers to test complex theories involving multiple observed and latent variables. With increasing number of SEM software packages, SEM has become very popular in psychology and other social sciences.

In this dissertation, we aim to expand the research on both missing data and SEM by examining how SEM approximate fit indices (AFIs) are affected by missing data under different missing data techniques. More specifically, the research demonstrates that popular SEM AFIs can be distorted when being estimated using one of the most popular missing data techniques, the full information maximum likelihood (FIML), and such distortion can be corrected through alternative missing data techniques. In this introductory chapter, we first introduce the key topics related to the current research, including missing data mechanisms, missing data patterns, missing data techniques, SEM, and SEM AFIs. Then we review the past research on how missing data affect SEM AFIs. We end the chapter with an outline for the rest of the dissertation paper.

1.1 Missing Data Mechanisms

The most common classification of missing data is by missing data mechanism, which is first proposed by Rubin [31]. Missing data mechanism can be thought as a kind of missing data generation rule that describes the statistical relationship between variables and the probability of missing data at the population level [28, 31]. There are generally three types of missing data [32]: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) missing not at random (MNAR). In this section, we review these

three types of missing data mechanisms in both informal and formal terms.

Let us consider a dataset with n subjects and p variables denoted as $X_1, \ldots, X_p$. When we do not have missing data, our dataset should look like a matrix with n rows and p columns. When we have missing data, we can consider the missing data as unobserved values that create holes in the data matrix. Suppose only X_1 has missing values. If X_1 is MCAR, then the probability of a subject having a missing value of X_1 does not depend on its unobserved value in X_1 nor its observed values of other variables. This means that knowing the subject's values on any of the variables does not give you any information about its probability of being missing. An example of MCAR data is when the paper-form questionnaire data are missing because a house cat spilled coffee on the table. In this case, there are no observed nor missing data that can predict the probability of being missing. If X_1 is MAR, then the probability of a subject being missing depends on its observed values of other variables but does not depend on its value of X_1. In other words, MAR means "conditionally missing at random": conditional on the observed values of other variables, the probability of being missing does not depend on the value of X_1. An example of MAR data is when shy participants tend to have more missing values on the questionnaire items about their sexual orientation. In this case, we can use the items that measure the shyness of participants to predict the probability of missing data about sexual orientation. If X_1 is neither MCAR nor MAR, then X_1 is MNAR, where the probability of a subject having a missing value on X_1 depends on its value of X_1. A classical example of MNAR data is when participants with high income avoid answering questions about income. In this case, the probability of missing the income data is related to participants' own income.

To define the types of missing data mechanisms formally, let $X = (X_1, \ldots, X_p)^T$ be a random vector representing the p variables in the dataset and $x = (x_1, \ldots, x_p)^T$ represent the realizations of X. Same as above, suppose X_1 is the only random variable with missing data. Let M be a random indicator variable with $M = 1$ representing a missing value in X_1;

we call M the *missing data indicator*. MCAR occurs when the distribution of M does not depend on x:

$$P(M = 1|x) = P(M = 1)$$

$$P(M = 0|x) = 1 - P(M = 1|x) = 1 - P(M = 1) = P(M = 0).$$

To define MAR and MNAR, we have to break down x into the observed (x_{obs}) and the unobserved or missing (x_{mis}) parts of x; that is $x = (x_{\text{mis}}, x_{\text{obs}})^T$. In this case, since x_1 is the only variable with missing data, $x_{\text{mis}} = x_1$ and $x_{\text{obs}} = (x_2, \ldots, x_p)^T$. MAR occurs when the distribution of M depends on x_{obs} but not x_{mis}:

$$P(M = 1|(x_{\text{mis}}, x_{\text{obs}})^T) = P(M = 1|x_{\text{obs}})$$

$$P(M = 0|(x_{\text{mis}}, x_{\text{obs}})^T) = P(M = 0|x_{\text{obs}}).$$

Notice that MAR data becomes MCAR data when M's dependence on x_{obs} is zero. Lastly, MNAR occurs when the distribution of M depends on x_{mis}; that is when $P(M = 1|(x_{\text{mis}}, x_{\text{obs}})^T)$ and $P(M = 0|(x_{\text{mis}}, x_{\text{obs}})^T)$ cannot be simplified further.

An important concept related to the types of missing data mechanisms is *ignorability*. Ignorable data are the types of missing data that can be handled with the likelihood-based analysis such as the full information maximum likelihood (FIML) estimation method. In other words, with ignorable missing data, we can obtain consistent parameter estimates without explicitly modelling the underlying missing data mechanism. Ignorable missing data needs to satisfy the following two conditions: 1) the data are either MCAR or MAR; 2) parameters associated with the specific missing data rule are distinct from the parameters associated with the distribution of the variables in the dataset [31]. In the above example, the second condition means that the parameters associated with the distribution of M are distinct from the parameters associated with the distribution of X. To explain

why these conditions are needed, let θ and ϕ are the parameters associated with X and M, respectively, and let $f(x,m;\theta,\phi)$ denote the joint density of X and M. Because θ and ϕ are distinct, when the data are incomplete, the observed data likelihood can be obtained via the marginal of x_{obs} as follows:

$$f(x_{\text{obs}},m,\theta,\phi) = \int f(x_{\text{obs}},x_{\text{mis}};\theta)f(m|x_{\text{obs}},x_{\text{miss}};\phi)dx_{\text{mis}} \tag{1.1}$$

When the data are MCAR, $f(m|x_{\text{obs}},x_{\text{miss}};\phi) = f(m;\phi)$; when the data are MAR, $f(m|x_{\text{obs}},x_{\text{mis}};\phi) = f(m|x_{\text{obs}};\phi)$. Since neither $f(m;\phi)$ nor $f(m|x_{\text{obs}};\phi)$ involves x_{mis}, we can take $f(m;\phi)$ or $f(m|x_{\text{obs}};\phi)$ out of the integral. In other words, for MCAR or MAR data, it is sufficient to maximize $\int f(x_{\text{obs}},x_{\text{mis}};\theta)dx_{\text{mis}}$ with respect to θ if we only want to estimate θ. There are MAR data that violate the second assumption for ignorable missing data (i.e.,θ and ϕ are not distinct); in such cases, statistical methods assuming ignorability are not optimal but are generally still good for obtaining consistent estimates. Therefore, in practice, ignorable missing data imply MCAR or MAR data, and non-ignorable missing data imply MNAR data.

In addition, it is worth noting that the fact that ignorable missing data can be handled by the FIML method (i.e., producing consistent estimates) rests on the assumption that the model is correctly specified. In the case of SEM, if the hypothesized model is the same as the population model, then the FIML method is able to produce consistent model parameter estimates and model fit. On the other hand, if the hypothesized model is misspecified, the FIML method estimates the model parameters so that the "distance" between hypothesized probability distribution and the true probability distribution is as close as possible.[1] In this dissertation, we call the parameters obtained under a misspecified model the "pseudo-parameters". As we will show later, even with ignorable missing

[1]The "distance" between two probability distributions is known as the Kullback-Leibler divergence [41].

data, the FIML method does not produce consistent estimates for the "pseudo-parameters" of complete data.

Due to the important property of ignorability for MCAR and MAR data, missing data mechanism is by far the most important feature of missing data being studied in simulation studies involving missing data. Most SEM research on missing data focuses on studying missing data techniques that can handle ignorable missing data; in other words, the researchers will mainly focus on MCAR and MAR missing data in their simulation studies [e.g., 36, 37, 45]. Our research is no exception. Our research mainly focuses on missing data approaches that can be used to handle ignorable missing data; therefore, as you will see in Chapters 4 and 6, examining different types of MCAR and MAR missing data is one of the main focuses in our missing data simulation studies.

1.2 Missing Data Patterns

Missing data pattern is another way to categorize missing data. Missing data pattern refers to the arrangement of observed and missing values in a dataset [15]. It is often confused with missing data mechanism [e.g. 16]. The distinction is that a specific missing data mechanism is a missing data generation rule that describes the relationship between variables and the probability of missing, whereas a specific missing data pattern is a data configuration that describes the location of the missing values in the data.

Although missing data pattern and missing data mechanism are distinct concepts, they do affect each other. Given a specific missing data generation rule with a certain type of missing data mechanism, the number and the type of missing data pattern will be determined. For example, suppose a dataset has $X_1, \ldots, X_p$ variables, if the missing data rule is *each subject has* 20% *probability of being missing from the variable* X_1, then the missing data pattern is univariate, implying two missing patterns.

When it comes to studying missing data techniques such as FIML, missing data pat-

terns are often considered less important than missing data mechanisms, probably because missing data patterns are not directly related to the ignorability property of missing data. Nonetheless, missing data patterns have several important implications for missing data techniques. First, when missing data patterns have variables that are never observed together, some parameters such as those measuring the correlations between these variables may not be estimable from the observed data (see Example 1.7 in [23]). Second, the number of missing data patterns may affect the performance of missing data techniques [35, 36]. For example, Savalei and Bentler [35] found that the number of missing data patterns may affect the efficiency of an estimation method, and this effect can be as strong as that of the missing data mechanism. Based on these previous findings, we have varied both types of missing data mechanisms and the number of missing data patterns when designing our simulation studies. Indeed, as we will explain in detail, missing data pattern can interact with missing data mechanism in their effects on the estimation of SEM AFIs.

Finally, another importance of missing data patterns is that the loglikelihood function for missing data can be written as a sum that iterates through each missing data pattern in the dataset, weighted by the proportion of each pattern. For example, loglikelihood function of Equation 1.1 can be written as

$$\log L(\theta; x_{\text{obs}}) = \sum_{j=1}^{J} \hat{q}_j \log L(\theta; x_{\text{obs},j}), \tag{1.2}$$

where J is the number of missing data patterns in the population, and q_j is proportion of missing data in a pattern j, where $j = 1, \ldots, J$. As we will explain later, we can also write the SEM FIML fit function as a sum that iterates through the missing data patterns; doing so allows us to see how missing data affect the estimation of AFIs under the FIML estimation.

1.3 Missing Data Techniques

In older times, when computing power is limited, the most common techniques for handling missing data include listwise deletion, pairwise deletion and mean substitution. The main goal of these techniques is to get rid of the missing data so that some data analyses could be done. This is in contrast with the modern missing data techniques, which main goal is to effectively deal with the missing data so that the data analysis can be used to obtain unbiased, consistent and efficient estimates of the population parameters. Most modern data techniques are mainly designed to handle ignorable missing data (i.e., MCAR and MAR data). MNAR data are almost impossible to be dealt with unless the researchers can effectively model the underlying missing data generation rule [1].

In SEM, the most common modern approach to handling ignorable missing data is the normal theory FIML estimation [1, 2, 43]. The FIML method is available in almost all SEM software, and it is usually the go-to estimation method for missing data. As explained earlier, the FIML approach involves maximizing the observed data likelihood in order to obtain the parameter estimates. Because the likelihood function for ignorable data does not involve the parameters associated with the missing data mechanism, FIML is able to produce consistent parameter estimates and standard errors under a correctly specified model.

The other modern missing data techniques used in SEM are the multiple imputation (MI) and two-stage (TS) approaches. These two approaches are well-researched but less commonly used in SEM literature. The MI approach consists of three steps: 1) imputation, 2) analysis, and 3) pooling [31]. In the imputation step, multiple sets of the data are created, each of which contains different estimates of the missing values (i.e., each dataset has the missing data filled in). This step involves an iterative process based on the Markov Chain Monto Carlo (MCMC) algorithm, which has been implemented in common SEM software such as M*plus* and lavaan package in *R* [1]. In the analysis step, the hypothesized

model is fit to each filled-in dataset as if there were no missing data, and then the statistic of interest (e.g., parameter estimates) is computed for each dataset. In the pooling step, the results across the imputed datasets are combined into a single result. In a way, the MI approach is similar to the older regression-based imputation technique, where the imputed values are based on the regression model built with cases with no missing data. However, the older regression-based imputation technique underestimates the standard errors because the imputed values always fall right on the regression line/plane; the MI approach solves this problem by incorporating simulated random draws from the population in the imputation step (see [15] for details).

The TS approach involves a two-stage procedure for obtaining parameter estimates [43, 46]. The first stage involves fitting a saturated model, which is an unrestricted model with zero degrees of freedom, to the incomplete data in order to estimate the saturated model's mean vector and covariance matrix. The saturated model's mean vector and covariance matrix essentially estimates what the mean vector and covariance matrix would have been if the data had been complete. This stage is analogous to the imputation stage of the MI method; however, instead of imputing missing data to create a "complete" dataset, the first stage of the TS method directly estimates the mean vector and covariance matrix of the "complete" dataset. Then in the second stage, the saturated model's estimated mean vector and covariance matrix are used to minimize the complete data fit function in order to obtain consistent estimates of the model parameters. Unbiased standard errors for the parameter estimates can be obtained by using a sandwich-type covariance matrix developed based on the likelihood theory (see [36, 48] for details).

It is not hard to see that the FIML approach is the "simplest" modern missing data approach in terms of computational complexity, which underscores its popularity. One unique advantage of the MI and TS approaches over the FIML approach is that they allow the incorporation of auxiliary variables, which are variables that the researchers are not

interested to study but their inclusion may improve the estimates of model parameter or standard error [36, 48]. Past research showed that with these auxiliary variables, the TS approach can produce more stable estimates in smaller samples [36, 48].

Most of the previous SEM research comparing the FIML, MI and TS approaches focused mostly on model parameter estimates, standard errors and confidence intervals of the estimates [e.g. 12, 35–37, 48]. Only a small number of research studies , which we will review in a later section, have compared these methods in terms of estimating AFIs. Our research aims to address this gap of research. In this dissertation, we focus mainly on the FIML and TS approaches; the MI approach will be discussed in the final chapter.

1.4 SEM and SEM AFIs

What is *structural equation modelling*? This may not be an easy question to answer even for researchers who are familiar with SEM. Indeed, the word, "structural equation modeling" or "SEM", describes a diverse set of mathematical models, computer algorithms and statistical methods that involve fitting a network of constructs to data. Historically, SEM comes from three different streams of research: 1) path analysis, 2) measurement models, and 3) general estimation algorithms for statistical models [5] .

Despite the varied origins of SEM, one important theme in SEM is the modelling of theoretical constructs that cannot be directly observed in a dataset. With SEM, researchers can represent these underlying theoretical constructs by latent variables, and they can estimate these latent factors via several observed variables that serve as "indicators" of the latent variables. The indicators for a latent variable can be selected based on prior knowledge or based on exploratory factor analyses that can measure the degree to which the indicators "tap into" the latent factor. The main advantage of SEM is its flexibility in incorporating both the relationships between several observed variables and one latent variable (via the measurement model part of SEM) as well as the relationships between

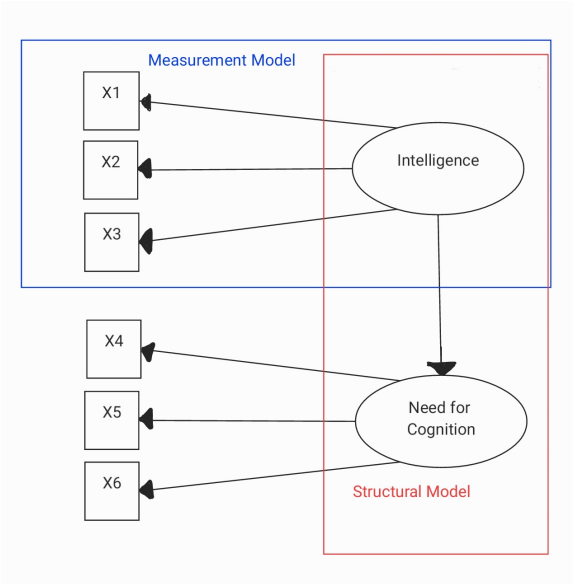


Figure 1.1: An example of an SEM model. In SEM diagram, circles represent latent variables, rectangles represent observed variables, and the arrows represent relationship between variables. The relationships between the observed variables and the latent factor for the measurement model part of SEM model (see the blue box for an example); the relationships between different latent variables form the structural part of SEM model (see the red box for an example).

several different latent variables (via the structural model part of SEM) (see Figure 1.1). When conducting SEM analysis, researchers can specify their hypothesized model that may include the structural part or the measurement part or both. Through fitting the hypothesized model to the data, researchers can obtain the estimates of the model parameters as well as the model-implied covariance matrix (a.k.a. model-based covariance matrix), which is the covariance matrix computed based on the estimates of the model parameters.

Another important theme in SEM is the measure of the overall model fit; that is, the measure of the extent to which relationships between variables as specified in the hypothesized model are representative of the true relationships found in the population. Traditionally, researchers uses the chi-square test of fit to make a binary decision on whether the model is sufficiently fit to the data. However, "all models are wrong but some are useful"; in a sense, all hypothesized models can be rejected given a large enough sample, thus defeating the purpose of the chi-square test. Therefore, in recent decades, researchers

have proposed approximate fit indices (AFIs), which measure the *degree* to which the hypothesized model is fit to the data [3, 39]. In other words, an AFI is a *continuous* metric along which to evaluate the hypothesized model's appropriateness for the data.

In our research, we will focus on the two most popular AFIs in SEM: the root mean square error of approximation (RMSEA) and comparative fit index (CFI). RMSEA measures the amount of misfit in the hypothesized model per degrees of freedom. CFI measures the amount of improvement in fit for the hypothesized model relative to the fit of the baseline model (a.k.a. the independence model), which is a null model where all variables are uncorrelated. RMSEA value is equal to or greater than zero, with lower value indicating better fit (i.e., zero indicating perfect fit) whereas CFI value ranges from zero to one, with higher value indicating better fit (i.e., one indicting perfect fit). Detailed equations for these AFIs will be provided in the later chapters. Ironically, although AFIs are supposed to measure fit on a continuum, cut-off points are still commonly used to help researchers categorize the amount of misfit. For RMSEA, a value less than 0.08 indicates good fit [8]; for CFI, a value greater than 0.9 indicates good fit [17].

Finally, we explain RMSEA and CFI's relationship with other types of AFIs. As we will show in the later chapters, both RMSEA and CFI are defined in terms of the fit function minimum values. There are other AFIs that are also defined in terms of the fit function minimum (e.g., Normed Fit Index (NFI), Tucker-Lewis Index (TLI) [3]); for these AFIs, the patterns of results in this dissertation work should also apply to them. However, for AFIs that are not defined in terms of the fit function minimum (e.g., the standardized root mean square residual (SRMR) [4] and goodness of fit index (GFI) [25]), our results may not apply. Many of these other AFIs fall out of popularity due to a variety of reasons. For example, SRMR can be very biased in smaller samples, and NFI does not account for the complexity of the hypothesized model well [3, 17]. Due to the unpopularity of these AFIs, we did not include them in our study.

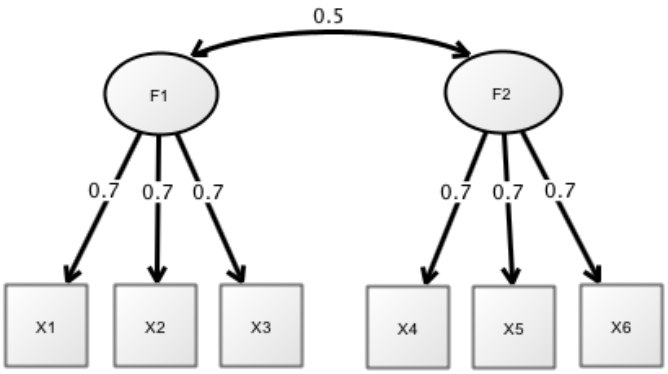


Figure 1.2: The population model for the simulation example in section 1.5.

1.5 Past Research on the Effect of SEM AFIs under FIML Estimation

The first main goal of this dissertation is to point out the potentially problematic performance of AFIs when computed following FIML estimation. It does not appear to be well-known that when AFIs are computed following FIML estimation, the resulting population values are distorted relative to their complete data counterparts. This means that the approximate fit of the same model to data drawn from the same population may be different depending on whether the data are complete or incomplete. To illustrate, we have generated a sample of $n = 1000000$ from a population that follows a correlated (a.k.a. oblique) two-factor model with standardized loadings of 0.7 and a factor correlation of 0.5 (see Figure 1.2). We fit a one-factor model to this sample. With complete data, the RMSEA and CFI are 0.203 and 0.747, respectively. However, when we randomly delete 50% data for each of the three variables loading on factor 1, the RMSEA and CFI are now 0.148 and 0.816, respectively.

We are aware of only three studies that have examined the behavior of AFIs with incomplete data under the FIML estimation; none of them noted this phenomenon. Davey et al. [10] conducted a simulation study, examining the effects of incomplete data on AFIs

in sample data. They found that with a misspecified model, sample AFIs following FIML estimation indicated better fit with higher percentage of missing data, but they do not provide an explanation for this finding. Enders and Mansolf [13] have conducted a simulation study comparing AFIs computed under the FIML and MI approaches. He found that both approaches have produced similar AFIs. More specifically, under both approaches, sample CFI stayed relatively the same with more missing data but RMSEA decreased slightly with missing data. One drawback of Enders and Mansolf [13]'s simulation study is that they used a misspecified model that was only slightly misspecified (RMSEA=0.041 and CFI=0.981 for complete data). It is impossible for the AFIs to show much improvement in fit with the addition of missing data when the model misspecification is already minor with complete data. Finally, Li and Lomax [22] conducted a simulation study, where they examined the effects of incomplete and nonnormal data on RMSEA. They found that sample RMSEA following FIML estimation had relatively little bias; however, the authors did not report any population RMSEA values, so it is unclear how the sample bias was computed.

The second main goal of this dissertation is to examine alternative methods for estimating AFIs so that the AFIs are not distorted by missing data. We are only aware of one recently published research paper that also examined alternative approaches for estimating AFIs. Lai [21] has studied the TS approach for computing RMSEA under missing data. In addition, he proposed a small sample correction to improve the TS estimation of RMSEA in finite samples with missing data. He found that across a wide variety of conditions, the TS approach with the small sample correction consistently produced RMSEA estimates that are closer to the complete data population RMSEA values relative to the FIML approach. However, Lai [21] did not explain how the small sample correction should be computed. Our paper addresses this gap in the research, and propose two computational versions for the small sample corrections under the TS estimation.

1.6 Dissertation Organization

This dissertation has two main goals. The first goal is to examine why and how AFIs are distorted by missing data under the FIML estimation. The second goal is to propose and investigate alternative computations of AFIs that can produce consistent and unbiased estimation in the presence of missing data.

Chapters 2 to 4 address the first goal of our dissertation. Summary of each of these chapters are as follows:

- Chapter 2 provides the technical details that can help us explain why AFIs can be affected by missing data under the FIML estimation method. We first show how we can rewrite the minimum of the FIML fit function in terms of the missing data patterns. Then we obtain the minimum of this fit function at the population level by figuring out the population limits of each component in the fit function minimum. Here, we show the population limits vary across different types of missing data. Finally, we show how these fit functional minimum directly affect the estimates of AFIs.

- Chapter 3 provides a few analytical examples that demonstrate how AFIs are affected with increasing missing data under FIML. Here, we give examples where AFIs change with missing data as well as examples where AFIs stay the same with more missing data. We show examples where AFIs change solely because of the differences in equations between complete and incomplete data; we also show examples where AFIs change due to both the differences in equations and the differences in the population parameter values.

- Chapter 4 presents the results from two large sample simulation studies that examine the effect of missing data on AFIs under more realistic models. We focus on large samples in order to study the behaviour of AFIs without the presence of sampling

error. Across the two simulation studies, we have mainly manipulated the amount of missing data, missing data mechanism, missing data pattern and the location of missing data relative to the location of misfit. Each of these factors have turned out to be important in the effect of missing data on AFIs under FIML.

Chapters 5 and 6 address the second goal of the dissertation. Summary of these two chapters are as follows:

- Chapter 5 proposes two alternative approaches that can address the problems of AFIs under FIML. One approach involves implementing a correction step following the FIML estimation; we call it the FIML-corrected or FIML-C approach. The second approach involves the use of the TS method. We lay out all the technical details for the two approaches and provide two analytical examples that demonstrate how these methods should be used.

- Chapter 6 presents the results of two simulation studies that compare the AFIs under the FIML-C and TS approaches relative to the original FIML approach. The design of the two simulation studies is the same as that of the simulation studies in Chapter 4 except that the studies in this chapter include the FIML-C and TS approaches as well as simulated data with more varied sample sizes. Overall, the results from the simulation studies give support for the use of the alternative methods.

Finally, in Chapter 7, we conclude the dissertation by summarizing the main results, providing recommendations for applied researchers, discussing the limitations of the current research, and suggesting a few future research directions.

Chapter 2

SEM AFIs under FIML Estimation: Technical Details

> I argued that full information maximum likelihood (FIML) has several advantages over multiple imputation (MI) for handling missing data: 1) FIML is simpler to implement (if you have the right software); 2) unlike multiple imputation, FIML has no potential incompatibility between an imputation model and an analysis model; 3) FIML produces a deterministic result rather than a different result every time.
>
> ---
>
> *Paul Allison, 2012*

Although many past research studies have shown that missing data affect AFIs such as RMSEA and CFI under the FIML estimation [10, 13, 22], none of them have provided a mathematical explanation for such a phenomenon. In this chapter, we provide the technical details to show how RMSEA and CFI are affected by missing data under FIML. Since RMSEA and CFI are functions of the fit function minimum, we start this chapter by explaining how the fit function minimum is changed with missing data. We first show the derivations of the fit function minimum for complete data at both the sample and the

population levels. We then show the derivations for incomplete data at the sample and population levels. Finally, we explain how the change in the fit function minimum affects RMSEA and CFI.

2.1 Fit Function Minimum for Complete Data

Let $x_1, \ldots, x_n$ be a random sample from p-variate normal distribution with $N(\mu, \Sigma)$. We want to test the null hypothesis that the data come from $N(\mu(\theta), \Sigma(\theta))$, where θ is a $q \times 1$ vector of model parameters. The normal-theory maximum likelihood (ML) estimator $\hat{\theta}$ maximizes the observed data log-likelihood

$$\begin{aligned} l(\theta|x_1, \ldots, x_n) =& \sum_{i=1}^{n} l_i(\theta) \\ =& C - \frac{n}{2}\Big(\log|\Sigma(\theta)| + \text{tr}|S\Sigma^{-1}(\theta)| + (\bar{x} - \mu(\theta))'\Sigma^{-1}(\theta)(\bar{x} - \mu(\theta)) \Big), \end{aligned} \tag{2.1}$$

where $\bar{x}$ and S are sample means and covariance matrix, and C does not depend on θ. We denote the maximized log-likelihood for the structured (hypothesized) model as $\hat{l}$. The model-implied means and covariance matrix are $\hat{\mu} = \mu(\hat{\theta})$ and $\hat{\Sigma} = \Sigma(\hat{\theta})$. We can also maximize Equation 2.1 under the saturated model, which includes all the unique elements in μ and Σ as model parameters. We denote the maximized log-likelihood for the saturated model as $\tilde{l}$. The estimates of means and covariance matrix under the saturated model are $\tilde{\mu} = \mu(\tilde{\theta}) = \bar{x}$ and $\tilde{\Sigma} = \Sigma(\tilde{\theta}) = S$. With complete data, the saturated model estimates are just the familiar sample means and sample covariance matrix.

Maximizing the log-likelihood in Equation 2.1 is equivalent to minimizing the familiar

ML fit function, [1] whose minimum is given by:

$$F_{\mathrm{c}}\big(\hat{\mu},\hat{\Sigma}|\bar{x},S\big)=\log|\hat{\Sigma}S^{-1}|+\mathrm{tr}(S\hat{\Sigma}^{-1})+(\bar{x}-\hat{\mu})'\hat{\Sigma}^{-1}(\bar{x}-\hat{\mu})-p, \tag{2.2}$$

where the subscript c represents complete data. The likelihood ratio (LR) test statistic is a scaled difference between the structured and the saturated log-likelihoods, and it can also be expressed in terms of the fit function minimum, as follows:

$$T_c=-2(l(\hat{\theta})-l(\tilde{\theta}))=nF_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\bar{x},S), \tag{2.3}$$

where T_c denotes the LR test statistic for complete data. In order to derive population values of AFIs, it is necessary to determine the population limit of $F_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\bar{x},S)$. When the hypothesized model is true, this limit is zero. In this article, we are primarily interested in the case when the hypothesized model is false, as this is when the AFIs become relevant for evaluating the degree of misfit. Denote the population limits of sample parameter estimates under the structured model as follows: $\hat{\theta}\rightarrow\theta_0$, and the corresponding limits of the model-implied means and covariances are given by $\hat{\mu}\rightarrow\mu_0$ and $\hat{\Sigma}\rightarrow\Sigma_0$. Under the saturated model, the sample estimates of means and covariances, $\bar{x}$ and S, will converge to μ and Σ, respectively. When the structured model is wrong, it is generally the case $\Sigma\neq\Sigma_0$ and it is sometimes the case that $\mu\neq\mu_0$ (in the presence of a mean structure). Therefore, when the data are complete, the fit function minimum at the population level is given by

$$F_c(\mu_0,\Sigma_0|\mu,\Sigma)=\log|\Sigma_0\Sigma^{-1}|+\mathrm{tr}(\Sigma\Sigma_0^{-1})+(\mu-\mu_0)'\Sigma_0^{-1}(\mu-\mu_0)-p. \tag{2.4}$$

We refer to the values θ_0 as "pseudo-parameters" [2], because they are population parame-

[1] The ML fit function used in SEM is the equivalent to the Kullback-Leibler divergence (see footnote 1 in Chapter 1).

[2] In the statistics literature, the "pseudo-parameters" are also known as the "pseudo-true values" [40] , the "least false parameter values minimizing Kullback-Leibler divergence"[9], or the "parameter vector that

ters for an incorrect model.

2.2 Fit Function Minimum for Incomplete Data

2.2.1 Sample Fit Function Minimum for Incomplete Data

Let $x_1, \ldots, x_n$ again be a random sample from the p-variate normal distribution $N(\mu, \Sigma)$. If the sample contains missing data, for each $i = 1, \cdots, n$, the corresponding observed vector $x_{\text{obs},i}$ is of dimension $p_i \times 1$. Under an ignorable missing data mechanism (i.e., MCAR or MAR), the FIML estimator $\hat{\theta}$ can be obtained by maximizing the observed data log-likelihood

$$\begin{aligned} l(\theta | x_1, \ldots, x_n) &= \sum_{i=1}^{n} l_i(\theta) \\ &= C - \frac{1}{2} \sum_i \log |\Sigma_i(\theta)| - \frac{1}{2} \sum_i \big(x_{\text{obs},i} - \mu_i(\theta)\big)' \Sigma_i^{-1}(\theta) \big(x_{\text{obs},i} - \mu_i(\theta)\big), \end{aligned} \tag{2.5}$$

where $\mu_i(\theta)$ is the relevant $p_i \times 1$ subvector of $\mu(\theta)$, $\Sigma_i(\theta)$ is the relevant $p_i \times p_i$ submatrix of $\Sigma(\theta)$, and C does not depend on θ [e.g. 24] (see Table 2.1 for a summary of the notation used in this section). As with complete data, we can obtain the structured and saturated log-likelihoods (denoted by $\hat{l}$ and $\tilde{l}$, respectively):

$$\begin{aligned} \hat{l} &= \sum_{i=1}^{n} \hat{l}_i \\ &= C - \frac{1}{2} \sum_i \log |\hat{\Sigma}_i| - \frac{1}{2} \sum_i \big(x_{\text{obs},i} - \hat{\mu}_i\big)' \hat{\Sigma}_i^{-1} \big(x_{\text{obs},i} - \hat{\mu}_i\big), \end{aligned}$$

minimizes the Kullback-Leibler divergence" [19].

Table 2.1: Notation for mean vectors and covariance matrices with incomplete data under an incorrect hypothesized model.

Description	**Population Quantities**	**Consistent Sample Estimates**
True Means and Covariance Matrix	μ, Σ	$\tilde{\mu}, \tilde{\Sigma}$
Model-implied Means and Covariance Matrix (pseudo-parameters)	$\mu_0 = \mu(\theta_{0_m}), \Sigma_0 = \Sigma(\theta_{0_m})$	$\hat{\mu} = \mu(\hat{\theta}), \hat{\Sigma} = \Sigma(\hat{\theta})$
True Means and Covariance Matrix (sub-components for pattern j)	μ_j, Σ_j	$\tilde{\mu}_j, \tilde{\Sigma}_j$
Model-implied Means and Covariance Matrix (sub-components for pattern j)	$\mu_{0_m,j}, \Sigma_{0_m,j}$	$\hat{\mu}_j, \hat{\Sigma}_j$
Pattern-specific Means and Covariance Matrix	μ_j^*, Σ_j^*	$\bar{x}_j, S_j$

Note: The subscript 0_m indicates that the population limits for missing data are different from those for complete data, which are denoted by the subscript 0.

$$\begin{aligned}\tilde{l} =& \sum_{i=1}^{n} \tilde{l}_i \\ =& C - \frac{1}{2}\sum_i \log|\tilde{\Sigma}_i| - \frac{1}{2}\sum_i \left(x_{\text{obs},i} - \tilde{\mu}_i\right)' \tilde{\Sigma}_i^{-1} \left(x_{\text{obs},i} - \tilde{\mu}_i\right).\end{aligned}$$

With the structured model, we obtain the model-implied mean vector $\hat{\mu} = \mu(\hat{\theta})$ and covariance matrix $\hat{\Sigma} = \Sigma(\hat{\theta})$; with the saturated model, we obtain the saturated model estimates $\tilde{\mu} = \mu(\tilde{\theta})$ and $\tilde{\Sigma} = \Sigma(\tilde{\theta})$, which represent the incomplete data analogues of $\bar{x}$ and S. However, in the case of incomplete data, these saturated model estimates generally do not reduce to any closed form sample quantities. These saturated model estimates are also sometimes known as the "EM" [after the Expectation-Maximization (EM) algorithm; 11] means and covariances [e.g., 14].

The LR statistic is again the rescaled difference between the two log-likelihoods; however, with incomplete data, this statistic is typically not expressed as the sample size times the minimum of a fit function. In fact, the concept of a "fit function" does not seem to

have been defined for incomplete data. In this article, we introduce this concept and infer the form of this function by taking the difference of the two maximized log-likelihoods. That is, we write the LR test statistic for missing data as follows:

$$\begin{aligned} T_m &= -2(l(\hat{\theta}) - l(\tilde{\theta})) \\ &= \sum_i \log|\hat{\Sigma}_i| + \sum_i \left(x_{\text{obs},i} - \hat{\mu}_i\right)' \hat{\Sigma}_i^{-1} \left(x_{\text{obs},i} - \hat{\mu}_i\right) - \sum_i \log|\tilde{\Sigma}_i| - \sum_i \left(x_{\text{obs},i} - \tilde{\mu}_i\right)' \tilde{\Sigma}_i^{-1} \left(x_{\text{obs},i} - \tilde{\mu}_i\right) \\ &= \sum_i \log|\hat{\Sigma}_i \tilde{\Sigma}_i^{-1}| + \sum_i \left(x_{\text{obs},i} - \hat{\mu}_i\right)' \hat{\Sigma}_i^{-1} \left(x_{\text{obs},i} - \hat{\mu}_i\right) - \sum_i \left(x_{\text{obs},i} - \tilde{\mu}_i\right)' \tilde{\Sigma}_i^{-1} \left(x_{\text{obs},i} - \tilde{\mu}_i\right) \\ &= nF_m(\hat{\mu}, \hat{\Sigma} | \tilde{\mu}, \tilde{\Sigma}), \end{aligned} \tag{2.6}$$

where the general form of the minimized FIML fit function for incomplete data is

$$\begin{aligned} F_m(\mu(\theta), \Sigma(\theta) | \tilde{\mu}, \tilde{\Sigma}, \tilde{\phi}) = \frac{1}{n} \Big(& \sum_i \log|\Sigma_i(\theta) \tilde{\Sigma}_i^{-1}| + \sum_i (x_{\text{obs},i} - \mu_i(\theta))' \Sigma_i^{-1}(\theta) (x_{\text{obs},i} - \mu_i(\theta)) \\ & - \sum_i (x_{\text{obs},i} - \tilde{\mu}_i)' \tilde{\Sigma}_i^{-1} (x_{\text{obs},i} - \tilde{\mu}_i) \Big), \end{aligned} \tag{2.7}$$

where $\tilde{\phi}$ is the missing mechanism parameter vector. When the mechanism is MCAR, the vector $\tilde{\phi}$ contains only the population probabilities and the specification of each missing data pattern. When the mechanism is MAR, the vector $\tilde{\phi}$ contains additional parameters associated with the relationships between the probability of being missing in one variable and the observed value in the other variable. Comparing Equations 2.2 and 2.7 reveals that the equations of the fit function minima for complete and incomplete data are different. In addition, when the hypothesized model is misspecified, the model-implied mean vector and covariance matrix will differ (i.e., $\hat{\mu}$ and $\hat{\Sigma}$ will not be the same for complete and incomplete data) because the model parameters depend the missing mechanism parameters

(see the next section for a detailed explanation).[3]

To figure out the corresponding population limit of Equation 2.7 , it is necessary to re-write Equation 2.7 in terms of the missing data patterns. Let $\hat{q}_j = n_j/n$ be the observed proportion of missing data in pattern J, where $j = 1,\ldots,J$ and $\sum n_j = n$. Then Equation 2.7 can be rewritten as follows:

$$\begin{aligned} F_m(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma},\tilde{\phi}) = \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \hat{\mu}_j)'\hat{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \hat{\mu}_j) \\ - \frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \tilde{\mu}_j)'\tilde{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \tilde{\mu}_j)\Big). \end{aligned} \tag{2.8}$$

In the above equation, the summations over all n have been replaced with summation over the J missing data patterns and summations over the n_j observations within each pattern; $x_{\text{obs},i}$ has been replaced with $x_{\text{obs},i(j)}$, so that raw observations are now enumerated within each pattern j , $i(j) = 1,...,n_j$. In addition, $\hat{\Sigma}_j$, $\tilde{\Sigma}_j$, $\hat{\mu}_j$, and $\tilde{\mu}_j$ represent the appropriate sub-matrices of $\hat{\Sigma}$ and $\tilde{\Sigma}$ and sub-vectors of $\hat{\mu}$ and $\tilde{\mu}$, with only rows and columns corresponding to variables observed within pattern j.

We note that in the missing data literature, a similar version of Equation 2.8 has been provided by Muthén and Muthén [27] in the M*plus* technical appendices (see Appendix 6 Equation 133 in Muthén and Muthén [27]). However, Muthén and Muthén [27]'s equation did not write out all terms associated the saturated model; instead, they expressed the part of equation associated the saturated model with one constant term. Equation 2.8 is also different from the equations presented in some of the older missing data papers [e.g., 26], which relied on the multiple-group (MG) setup to handle missing data.[4]

We now re-write Equation 2.8 in terms of the sample covariance matrices, which will

[3]Technically, we should include a subscript $\tilde{\phi}$ for $\mu_i(\theta)$ and $\Sigma_i(\theta)$ in Equation 2.7 to denote their dependency on $\tilde{\phi}$; here, we omit this for the simplicity in notations.

[4]To use the MG fit function for handling missing data, pseudo-values corresponding to cases with missing data have to be inserted in the covariance matrices of the missing data patterns , and the degrees of freedom need to be adjusted for these pseudo-values after fitting the model. See Chapter 8 of Bollen [5] for a detailed explanation.

later help us find the population limit of $F_m(\hat{\mu}, \hat{\Sigma} | \tilde{\mu}, \tilde{\Sigma}, \tilde{\phi})$. To do this, we need to define the following three "sample covariance matrices" that can be computed within each missing data pattern:

$$
\begin{aligned}
S_j &= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \bar{x}_j)(x_{\text{obs},i(j)} - \bar{x}_j)', \\
S_{\hat{\mu},j} &= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \hat{\mu}_j)(x_{\text{obs},i(j)} - \hat{\mu}_j)', \\
S_{\tilde{\mu},j} &= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \tilde{\mu}_j)(x_{\text{obs},i(j)} - \tilde{\mu}_j)',
\end{aligned}
$$

Here, the first matrix S_j is the usual sample covariance matrix within pattern j computed using the within-pattern sample mean $\bar{x}_j$; the next two matrices $S_{\hat{\mu},j}$ and $S_{\tilde{\mu},j}$ are computed using model-estimated means, either under the structured model or under the saturated model. Using these three matrices, it follows that:

$$
\begin{aligned}
S_j &= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \bar{x}_j)(x_{\text{obs},i(j)} - \bar{x}_j)' \\
&= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \hat{\mu}_j)(x_{\text{obs},i(j)} - \hat{\mu}_j)' - (\hat{\mu}_j - \bar{x}_j)(\hat{\mu}_j - \bar{x}_j)' \\
&= S_{\hat{\mu},j} - (\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)';
\end{aligned}
$$

$$
\begin{aligned}
S_j &= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \bar{x}_j)(x_{\text{obs},i(j)} - \bar{x}_j)' \\
&= \frac{1}{n_j} \sum_{i}^{n_j} (x_{\text{obs},i(j)} - \tilde{\mu}_j)(x_{\text{obs},i(j)} - \tilde{\mu}_j)' - (\tilde{\mu}_j - \bar{x}_j)(\tilde{\mu}_j - \bar{x}_j)' \\
&= S_{\tilde{\mu},j} - (\bar{x}_j - \tilde{\mu}_j)(\bar{x}_j - \tilde{\mu}_j)'.
\end{aligned}
$$

We can also write $S_{\hat{\mu},j}$ and $S_{\tilde{\mu},j}$ in terms of S_j:

$$S_{\hat{\mu},j} = S_j + (\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)'; S_{\tilde{\mu},j} = S_j + (\bar{x}_j - \tilde{\mu}_j)(\bar{x}_j - \tilde{\mu}_j)'.$$

Using these expressions and the rules of trace, starting with Equation 2.8, we can write:

$$\begin{aligned}
F_m(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma},\tilde{\phi}) &= \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \hat{\mu}_j)'\hat{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \hat{\mu}_j) \\
&\quad - \frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \tilde{\mu}_j)'\tilde{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \tilde{\mu}_j)\Big) \\
&= \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \text{tr}(\frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \hat{\mu}_j)'\hat{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \hat{\mu}_j)) \\
&\quad - \text{tr}(\frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \tilde{\mu}_j)'\tilde{\Sigma}_j^{-1}(x_{\text{obs},i(j)} - \tilde{\mu}_j))\Big) \\
&= \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \text{tr}((\frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \hat{\mu}_j)(x_{\text{obs},i(j)} - \hat{\mu}_j)')\hat{\Sigma}_j^{-1}) \\
&\quad - \text{tr}((\frac{1}{n_j}\sum_{i}^{n_j}(x_{\text{obs},i(j)} - \tilde{\mu}_j)(x_{\text{obs},i(j)} - \tilde{\mu}_j)')\tilde{\Sigma}_j^{-1})\Big) \\
&= \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \text{tr}(S_{\hat{\mu},j}\hat{\Sigma}_j^{-1}) - \text{tr}(S_{\tilde{\mu},j}\tilde{\Sigma}_j^{-1})\Big) \\
&= \sum_{j=1}^{J} \hat{q}_j \Big(\log|\hat{\Sigma}_j\tilde{\Sigma}_j^{-1}| + \text{tr}((S_j + (\bar{x}_j - \hat{\mu}_j)(\bar{x}_j - \hat{\mu}_j)')\hat{\Sigma}_j^{-1}) \\
&\quad - \text{tr}((S_j + (\bar{x}_j - \tilde{\mu}_j)(\bar{x}_j - \tilde{\mu}_j)')\tilde{\Sigma}_j^{-1})\Big).
\end{aligned} \tag{2.9}$$

2.2.2 Population Fit Function Minimum for Incomplete Data

Before obtaining the population limit of the fit function minimum for incomplete data, we elaborate on the concept of the incomplete data population. If the process that created the current sample with incomplete data is allowed to go on indefinitely so that a larger and larger sample is generated, we would eventually sample the entire population. In this

way, the current observed sample with incomplete data can be viewed as a random sample from this incomplete data population. The observed percentage of missing values in the sample is a consistent estimate of the population percentage of missing values. Further, the observed incomplete data patterns and their relative frequency are assumed to accurately reflect the underlying incomplete data population. Of course, in smaller samples not all missing data patterns that are possible in the population may be realized.

We can now proceed to obtain the population limit of the fit function for incomplete data. To obtain the population limit of Equation 2.9, we assume that the index J enumerates all of the missing data patterns that exist in the population. This means either that the sample size is large enough that all missing data patterns that exist in the population have been realized in the sample, or alternatively, that in Equation 2.9, some $\hat{q}_j$ values are zero in the sample but will approach non-zero population values; in other words, the percentage of any missing data pattern in the sample is a consistent estimate of the population probability of that pattern.

In addition, we need to determine the limits of all sample quantities in Equation 2.9 to obtain the population limit of $F_m(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma},\tilde{\phi})$. Under an ignorable missing data mechanism (i.e., MCAR or MAR), the saturated model estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ are consistent for μ and Σ. Therefore, for any missing data pattern, it is the case that $\tilde{\mu}_j \rightarrow \mu_j$ and $\tilde{\Sigma}_j \rightarrow \Sigma_j$. We also define the population "pseudo-parameters" as the limits of the corresponding sample quantities, $\hat{\theta} \rightarrow \theta_{0_m}$, $\hat{\mu} \rightarrow \mu_{0_m}$, and $\hat{\Sigma} \rightarrow \Sigma_{0_m}$, where the subscript 0_m indicates that the population limits for missing data may be different from those for complete data, which are denoted by the subscript 0. Indeed, when the hypothesized model is wrong, it is generally the case that $\Sigma \neq \Sigma_{0_m}$ and $\mu \neq \mu_{0_m}$. With incomplete data, the FIML estimates of means can be different under the structured model even when the mean structure is saturated. In addition, when the model is misspecified, the "pseudo-parameters" for complete and incomplete data will usually differ from each other, resulting in different model-implied

mean and covariance matrix estimates even in the population (i.e., $\Sigma_0 \neq \Sigma_{0_m}$ and $\mu_0 \neq \mu_{0_m}$), unless the hypothesized model has no free parameters (see Chapter 3.2 for an example).

We first state the population limit of $F_m(\hat{\mu}, \hat{\Sigma} | \tilde{\mu}, \tilde{\Sigma}, \tilde{\phi})$ in the special case when the assumption of homogeneity of means and covariances holds. This assumption is always met when the data are MCAR [e.g., 20], and it is usually not met when the data are MAR or MNAR, although it is possible to construct a counter-example [44]. Under the homogeneity of means and covariances assumption, the estimates of pattern-specific means and covariances converge to the corresponding subsets of the overall population mean and covariance matrix; that is, $\bar{x}_j \to \mu_j$, $S_j \to \Sigma_j$ for all j, where μ_j is the $p_j \times 1$ sub-vector of μ and Σ_j is the $p_j \times p_j$ sub-matrix of Σ corresponding to the variables observed in the jth missing data pattern. Therefore, the population value of Equation 2.9 in the case of MCAR data is given by:

$$
\begin{aligned}
F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) &= \sum_{j=1}^{J} q_j \Big(\log |\Sigma_{0_m,j} \Sigma_j^{-1}| + \text{tr}((\Sigma_j + (\mu_j - \mu_{0_m,j})(\mu_j - \mu_{0_m,j})') \Sigma_{0_m,j}^{-1}) \\
&\quad - \text{tr}((\Sigma_j + (\mu_j - \mu_j)(\mu_j - \mu_j)') \Sigma_j^{-1}) \Big) \\
&= \sum_{j=1}^{J} q_j \Big(\log |\Sigma_{0_m,j} \Sigma_j^{-1}| + \text{tr}(\Sigma_j \Sigma_{0_m,j}^{-1}) \\
&\quad + (\mu_j - \mu_{0_m,j})' \Sigma_{0_m,j}^{-1} (\mu_j - \mu_{0_\phi,j}) - \text{tr}\big(\Sigma_j \Sigma_j^{-1}\big) \Big) \\
&= \sum_{j=1}^{J} q_j \Big(\log |\Sigma_{0_m,j} \Sigma_j^{-1}| + \text{tr}(\Sigma_j \Sigma_{0_m,j}^{-1}) \\
&\quad + (\mu_j - \mu_{0_m,j})' \Sigma_{0_m,j}^{-1} (\mu_j - \mu_{0_m,j}) - p_j \Big),
\end{aligned}
\tag{2.10}
$$

where J is the number of missing data patterns that are possible in the population, q_j is the population probability of the jth pattern, and p_j is the number of variables in the jth missing data pattern (see Table 2.1 for notation). Note that, in the functional form, this

equation is a weighted average, by pattern probabilities, of the complete data fit function given in Equation 2.4. However, the population limits of the model-implied estimates of the means and covariances will generally differ for complete and incomplete data.

In the more general case when data are MAR, the homogeneity of means and covariances assumption is typically violated. In this case, the limits of the within-pattern estimates of means and covariances are not necessarily equal to the corresponding subcomponents of the overall population means and covariance matrix. For example, consider the simplest case of two variables, X and Y, both $N(0,1)$, where Y is missing with probability one whenever $X > 0$. Even though the population means of X and Y are both zero, the pattern-specific means will be different. In the missing pattern where X is observed but Y is missing, all sample realizations of X are positive, and thus the estimated mean of X using only the cases with this pattern will approach the mean of a standard normal distribution truncated at zero. In the general case of MAR data, let $\bar{x}_j \rightarrow \mu_j^*$, $S_j \rightarrow \Sigma_j^*$ be the pattern-specific limits of the means and covariance matrix for variables within jth pattern. The population value of the fit function minimum is given by

$$\begin{aligned} F_{\text{MAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) = \sum_{j=1}^{J} q_j \Big(& \log |\Sigma_{0_m,j} \Sigma_j^{-1}| \\ & + \text{tr}((\Sigma_j^* + (\mu_j^* - \mu_{0_m,j})(\mu_j^* - \mu_{0_m,j})') \Sigma_{0_m,j}^{-1}) \\ & - \text{tr}((\Sigma_j^* + (\mu_j^* - \mu_j)(\mu_j^* - \mu_j)') \Sigma_j^{-1}) \Big). \end{aligned} \tag{2.11}$$

Thus, in the general case of ignorable incomplete data, the fit function minimum depends on: 1) the true means and covariances (μ and Σ, where μ_j and Σ_j are the corresponding sub-components for pattern j, $j = 1, ...J$); 2) the model-implied means and covariances (μ_{0_m} and Σ_{0_m}, with $\mu_{0_m,j}$ and $\Sigma_{0_m,j}$ indicating the relevant subcomponents for pattern j); and 3) the pattern-specific population means and covariances (μ_j^* and Σ_j^*, for $j = 1, ..., J$; see Table 2.1). The model-implied means and covariances in 2) will be different for the

complete and different types of incomplete data. The pattern-specific population means and covariances in 3) also depend on the missing data mechanism; when the data are MCAR, they are equal to the corresponding subsets of the population means and covariances (i.e., $\mu_j^* = \mu_j$ and $\Sigma_j^* = \Sigma_j$), however, when the data are MAR, they will usually differ across patterns and cannot be viewed as subsets of a single vector and matrix (i.e., $\mu_j^* \neq \mu_j$ and $\Sigma_j^* \neq \Sigma_j$).

We briefly note what happens when the data are MNAR. In this case, the saturated FIML estimates of mean vector and covariance matrix, $\tilde{\mu}$ and $\tilde{\Sigma}$, are no longer consistent for μ and Σ, so the general Equation 2.11 will feature the population limits of $\tilde{\mu}$ and $\tilde{\Sigma}$, instead of μ and Σ. We assume an ignorable missing data mechanism (MCAR or MAR data) for the remainder of this dissertation.

2.3 RMSEA and CFI for Complete and Incomplete Data

In all current SEM software, the RMSEA and CFI are computed using the same equations regardless of whether the data contain missing values. For complete normal data, under the ML estimation, the LR test statistic in Equation 2.3 is used to define RMSEA and CFI as follows:

$$\begin{aligned}
\widehat{\text{RMSEA}}_{\text{ML}} &= \sqrt{\frac{\max(T_c - df, 0)}{df(n)}} = \sqrt{\max\left(\frac{F_{\text{c}}(\hat{\mu}, \hat{\Sigma}|\bar{x}, S) - \frac{df}{n}}{df}, 0\right)}; \\
\widehat{\text{CFI}}_{\text{ML}} &= 1 - \frac{\max(T_c - df, 0)}{\max\left(T_c - df, T_{c,B} - df_B, 0\right)} \\
&= 1 - \frac{\max\left(F_{\text{c}}(\hat{\mu}, \hat{\Sigma}|\bar{x}, S) - \frac{df}{n}, 0\right)}{\max\left(F_c(\hat{\mu}_B, \hat{\Sigma}_B|\bar{x}, S) - \frac{df_B}{n}, F_{\text{c}}(\hat{\mu}, \hat{\Sigma}|\bar{x}, S) - \frac{df}{n}, 0\right)},
\end{aligned} \tag{2.12}$$

where the subscript B stands for the baseline model, which assumes all variables are uncorrelated with each other; that is, df_B, $\hat{\mu}_B$, $\hat{\Sigma}_B$ and $T_{c,B}$ are the baseline model's degrees

of freedom, model-implied means, model-implied covariance matrix, and LR test statistic, respectively. As mentioned in Section 1.4, as the model fit increases, RMSEA gets closer to zero and CFI gets closer to one. In the rare case when both the numerator and the denominator in the CFI computation are zero, the convention is to set CFI to one.

In the presence of missing data, under the FIML estimation, RMSEA and CFI are computed in the same way as the above equations except we use the corresponding LR test statistic for missing data in Equation 2.6, as follows:

$$\begin{aligned}
\widehat{\text{RMSEA}}_{\text{FIML}} &= \sqrt{\frac{\max(T_c - df, 0)}{df(n)}} = \sqrt{\max\Big(\frac{F_{\text{m}}(\hat{\mu}, \hat{\Sigma}|\tilde{\mu}, \tilde{\Sigma}, \phi) - \frac{df}{n}}{df}, 0\Big)}; \\
\widehat{\text{CFI}}_{\text{FIML}} &= 1 - \frac{\max(T_c - df, 0)}{\max\Big(T_c - df, T_{c,B} - df_B, 0\Big)} \\
&= 1 - \frac{\max\Big(F_{\text{m}}(\hat{\mu}, \hat{\Sigma}|\tilde{\mu}, \tilde{\Sigma}, \phi) - \frac{df}{n}, 0\Big)}{\max\Big(F_{\text{m}}(\hat{\mu}_B, \hat{\Sigma}_B|\tilde{\mu}, \tilde{\Sigma}, \phi) - \frac{df_B}{n}, F_{\text{m}}(\hat{\mu}, \hat{\Sigma}|\tilde{\mu}, \tilde{\Sigma}, \phi) - \frac{df}{n}, 0\Big)}.
\end{aligned} \tag{2.13}$$

We now show the population limits of RMSEA and CFI under complete and incomplete data. For complete data, we can find the population limits of RMSEA and CFI by using the population fit function minima for complete data in Equation 2.4, as follows:

$$\begin{aligned}
\text{RMSEA}_{\text{ML}} &= \sqrt{\frac{F_{\text{c}}(\mu_0, \Sigma_0|\mu, \Sigma)}{df}}; \\
\text{CFI}_{\text{ML}} &= 1 - \frac{F_{\text{c}}(\mu_0, \Sigma_0|\mu, \Sigma)}{F_{\text{c}}(\mu_{B,0}, \Sigma_{B,0}|\mu, \Sigma)},
\end{aligned} \tag{2.14}$$

where $\mu B, 0$ and $\Sigma_{B,0}$ are the population limits of the model-implied means of covariances under the baseline model. For incomplete data, we just use the corresponding population

fit function minima for incomplete data, as follows:

$$\begin{aligned} \text{RMSEA}_{\text{FIML}} &= \sqrt{\frac{F_{\text{m}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)}{df}} \\ \text{CFI}_{\text{FIML}} &= 1 - \frac{F_{\text{m}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)}{F_{\text{m}}(\mu_{B,0_m}, \Sigma_{B,0_m} | \mu, \Sigma, \phi)}. \end{aligned} \tag{2.15}$$

In summary, the AFIs' equations show that the AFIs depend on the fit function minima, and they may estimate different population values depending on the presence (and type) of missing data. In the special case when the model is exactly correct, all fit function minima will be zero in the population, and AFIs from complete and any type of incomplete data will agree asymptotically. However, it is a safe assumption that the model is never exactly correct in the population, and the complete and incomplete data AFIs will converge to different population values. For complete data, the fit function minimum in the population is given by Equation 2.4; for MCAR data, it is given by Equation 2.10, and for MAR data, it is given by the most general equation, Equation 2.11. It is worth emphasizing that even this categorization is incomplete: there is a *separate* population value of the RMSEA and CFI for *each* specific type of MCAR or MAR data, depending on the missing data proportion, location, patterns, and (in the case of MAR) conditioning rules.

Chapter 3

SEM AFIs under FIML Estimation: Analytical Examples

> When attempting to assess how well a model fits a particular dataset, one must realize at the outset that the classic hypothesis-testing approach is inappropriate.
>
> *James H.Steiger, 1980*

In this chapter, we demonstrate, with a few analytical examples, how RMSEA under the FIML estimation changes with the presence and type of missing data. As shown in Chapter 2, the equations of the fit function minimum differ under complete and incomplete data and under different types of incomplete data; consequently, AFIs can also be different due to the differences in the equations of the fit function minimum. In the first section of this chapter, we will present examples where RMSEA stays the same and examples where RMSEA changes with missing data solely due to the differences in the equations of the fit function minimum. In addition, under FIML, AFIs may also change with missing data due to the differences in the parameter values. In the second section, we will present an example where RMSEA changes due to both the differences in the equations and the

differences in the parameter values.

3.1 Change in RMSEA due to Differences in the Equations of the Fit Function Minimum

In this example, the hypothesized model is fully constrained (i.e., it has no free parameters), so the model-implied means and covariances do not differ for complete and different types of incomplete data, greatly simplifying computations. All observed differences are therefore only due to the different forms of the fit function.

Let $X_1, \ldots, X_6$ follow a multivariate normal distribution $N(\mu, \Sigma)$ with the following population covariance matrix and vector of means:

$$\Sigma = \begin{pmatrix} 1.00 & & & & & \\ 0.89 & 1.00 & & & & \\ 0.49 & 0.49 & 1.00 & & & \\ 0.00 & 0.00 & 0.00 & 1.00 & & \\ 0.00 & 0.00 & 0.00 & 0.49 & 1.00 & \\ 0.00 & 0.00 & 0.00 & 0.49 & 0.49 & 1.00 \end{pmatrix}, \mu = (0,0,0,0,0,0)'. \tag{3.1}$$

This covariance structure is consistent with that of a two-factor model with orthogonal factors and three indicators per factor (with loadings of 0.7), plus a correlated residual (of size 0.4) between X_1 and X_2. The model fit to data is always a fully constrained model, which is the same as the population model but without the correlated residual. The model-

implied covariance matrix and mean vector are

$$\Sigma_0 = \begin{pmatrix} 1.00 & & & & & \\ 0.49 & 1.00 & & & & \\ 0.49 & 0.49 & 1.00 & & & \\ 0.00 & 0.00 & 0.00 & 1.00 & & \\ 0.00 & 0.00 & 0.00 & 0.49 & 1.00 & \\ 0.00 & 0.00 & 0.00 & 0.49 & 0.49 & 1.00 \end{pmatrix}, \mu_0 = (0,0,0,0,0,0)'. \tag{3.2}$$

Because the hypothesized model has no free parameters, no fit function minimization is required to obtain "parameter estimates": all values are fixed a priori. Thus, the model-implied estimates μ_0 and Σ_0 will be the same in all of the examples considered below (given by Equation 3.2). However, one can still evaluate the fit function at these estimates to obtain the "fit function minimum" for the purposes of computing AFIs. It is important to note that the misfit is caused by the correlated residual between X_1 and X_2; therefore, the deviation of the fit function minimum from zero will always be due to the difference in the value of the covariance between X_1 and X_2 in Σ versus Σ_0. Because the fit function has a different form for complete and incomplete data, the numeric values of the AFIs can still differ even in this simplified example.

3.1.1 Case 1: Complete data

When there are no missing values, the fit function "minimum" in Equation 2.4 is given by

$$\begin{aligned} F_c(\mu_0, \Sigma_0 | \mu, \Sigma) &= \log|\Sigma_0 \Sigma^{-1}| + \mathrm{tr}(\Sigma \Sigma_0^{-1}) + (\mu - \mu_0)' \Sigma_0^{-1} (\mu - \mu_0) - p \\ &= 1.200 + 5.612 + 0 - 6 = 0.812. \end{aligned}$$

The corresponding population RMSEA calculated using Equation 2.14 is

$$\text{RMSEA}_{\text{ML}} = \sqrt{\frac{F_{\text{c}}(\mu_0, \Sigma_0 | \mu, \Sigma)}{df}} = \sqrt{\frac{0.812}{27}} = 0.173.$$

We will use this RMSEA value obtained under complete data as a benchmark to compare with values obtained under incomplete data.[1]

3.1.2 Case 2: MCAR data; misspecification does not involve variables with missing values

Now suppose that 20% of the values on X_6 are missing completely at random. In this case, there are $J = 2$ missing data patterns, with q_1=0.8 (probability of the complete data pattern) and $q_2 = 0.2$ (probability of the incomplete data pattern), and with $p_1 = 6$ and $p_2 = 5$ (number of observed variables in each pattern). Then, Equation 2.10 yields

$$\begin{aligned} F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) =& q_1 \Big(\log |\Sigma_{0_m,1} \Sigma_1^{-1}| + \text{tr}(\Sigma_1 \Sigma_{0_m,1}^{-1}) + (\mu_1 - \mu_{0_m,1})' \Sigma_{0_m,1}^{-1} (\mu_1 - \mu_{0_m,1}) - p_1 \Big) \\ &+ q_2 \Big(\log |\Sigma_{0_m,2} \Sigma_2^{-1}| + \text{tr}(\Sigma_2 \Sigma_{0_m,2}^{-1}) + (\mu_2 - \mu_{0_m,2})' \Sigma_{0_m,2}^{-1} (\mu_2 - \mu_{0_m,2}) - p_2 \Big) \\ =& (0.8)(0.812) + (0.2)(0.812) = 0.812. \end{aligned}$$

Because the hypothesized model is fully constrained, $\Sigma_{0_m} = \Sigma_0$ is given by Equation 3.2. The first pattern is the complete data pattern, so that $\Sigma_{0_m,1} = \Sigma_0$, $\Sigma_1 = \Sigma$, $\mu_{0_m,1} = \mu_0$, and $\mu_1 = \mu$; consequently, the component of $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ for the first pattern is the same as $F_c(\mu_0, \Sigma_0 | \mu, \Sigma)$ (i.e., it equals 0.812).

For the incomplete data pattern, $\Sigma_{0_m,2}$ and Σ_2 are the 5×5 sub-matrices of Σ_0 and Σ with the last row and column deleted, and $\mu_{0_m,2}$ and μ_2 are the 5×1 sub-vectors of μ_0 and μ with the last element deleted. Because Σ and Σ_0 are block-diagonal (the factors are orthogonal), model misfit caused by the correlated residual between X_1 and X_2 does not prop-

[1] We cannot compute the traditionally defined CFI because the traditional independence model is not nested within the highly restrictive hypothesized model used in this example.

agate to variables loading on the second factor. Since the variable with missing values (X_6) only loads on the second factor, the component of $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ for the second pattern turns out to be the same as $F_c(\mu_0, \Sigma_0 | \mu, \Sigma)$, and the entire $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ is the same as $F_c(\mu_0, \Sigma_0 | \mu, \Sigma)$.

In sum, compared to the complete data case (Case 1), in Case 2 the hypothesized model fit function minimum and consequently the RMSEA stay the same. The reason the fit function minimum for the hypothesized model stays the same is that the variable with missing data (i.e., X_6) contributes no information about the amount of misfit (which involves the covariance between X_1 and X_2), due to the nature of the model.

3.1.3 Case 3: MCAR data; misspecification involves variables with missing data

Next, we consider the case where the location of misfit in the covariance structure involves covariances among variables with missing data. Suppose that 20% of the values on X_1 (rather than X_6, as was in Case 2) are missing completely at random. As in Case 2, $J = 2$, $p_1 = 6$, $q_1 = 0.8$, $p_2 = 5$, and $q_2 = 0.2$. However, the fit function minimum value is now given by

$$\begin{aligned} F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) =& q_1 \big(\log |\Sigma_{0_m,1} \Sigma_1^{-1}| + \text{tr}(\Sigma_1 \Sigma_{0_m,1}^{-1}) + (\mu_1 - \mu_{0_m,1})' \Sigma_{0_m,1}^{-1} (\mu_1 - \mu_{0_m,1}) - p_1 \big) \\ &+ q_2 \big(\log |\Sigma_{0_m,2} \Sigma_2^{-1}| + \text{tr}(\Sigma_2 \Sigma_{0_m,2}^{-1}) + (\mu_2 - \mu_{0_m,2})' \Sigma_{0,2}^{-1} (\mu_2 - \mu_{0_m,2}) - p_2 \big) \\ =& (0.8)(0.812) + (0.2)(0) = 0.650. \end{aligned}$$

The first pattern is the complete data pattern, and the corresponding component of $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ has the same value. However, the second pattern now omits X_1: $\Sigma_{0_m,2}$ and Σ_2 are the 5×5 submatrices of Σ_0 and Σ with the first row and column deleted, and $\mu_{0_m,2}$ and μ_2 are the 5×1 subvectors of μ_0 and μ with the first element deleted. In this case, $\Sigma_{0_m,2}$ and Σ_2 no longer contain the covariance between X_1 and X_2, which is

the one covariance that is misspecified. Thus, $\Sigma_{0_m,2} = \Sigma_2$, and the second component of $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m}|\mu, \Sigma, \phi)$ is zero. As a result, the fit function minimum for the hypothesized model is smaller. This fit function minimum directly affects RMSEA:

$$\text{RMSEA}_{\text{FIML}} = \sqrt{\frac{F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m}|\mu, \Sigma, \phi)}{df}} = \sqrt{\frac{0.650}{27}} = 0.155.$$

This example illustrates that when the variables with missing data are also involved in the misspecification, the fit function minimum for the hypothesized model and, consequently, the RMSEA generally decrease relative to their complete data counterparts. We note that for a hypothesized model (such as the one in our example) where only part of the model is severely misspecified, an interaction between the location of misfit (relative to the location of missing data) and the effect of missing data on RMSEA is expected. If variables corresponding to the part of the model that is severely misspecified part are missing, then the model will show better fit when assessed by the RMSEA.

3.1.4 Case 4: MAR data; misspecification involves variables with missing values

The last example involves MAR data. Suppose that X_1 is missing with probability one whenever $X_2 > 0.842$, which is the z-score corresponding to the 80th percentile of a normal distribution. This implies that X_1 is missing with 20% probability. As before, $J = 2$, $p_1 = 6$, $p_2 = 5$, $q_1 = 0.8$ and $q_2 = 0.2$.

The correct equation for the fit function minimum is now given by Equation 2.11 instead of Equation 2.10. To compute Equation 2.11, we require pattern-specific population means and covariance matrices, that is, μ_j^* and Σ_j^* for $j = 1, 2$. In this example, even for the complete data pattern, $\mu_1^* \neq \mu$ and $\Sigma_1^* \neq \Sigma$. The reason is that in the complete data pattern, X_2 is distributed as a standard normal variable truncated at 0.842; in addition, X_1 and X_3 will no longer have normal distributions, as they will tend to have more negative than

positive values observed (by virtue of being correlated with X_2). We have used the tmvtnorm package in *R* to obtain the population covariance matrix and means of the truncated multivariate normal distributions corresponding to this example, yielding:

$$\Sigma_1^* = \begin{pmatrix} 0.670 & & & & & \\ 0.519 & 0.583 & & & & \\ 0.308 & 0.286 & 0.900 & & & \\ 0.00 & 0.00 & 0.00 & 1.00 & & \\ 0.00 & 0.00 & 0.00 & 0.49 & 1.00 & \\ 0.00 & 0.00 & 0.00 & 0.49 & 0.49 & 1.00 \end{pmatrix}, \mu_1^* = (-0.311, -0.350, -0.171, 0, 0, 0)',$$

$$\Sigma_2^* = \begin{pmatrix} 0.219 & & & & & \\ 0.107 & 0.812 & & & & \\ 0.00 & 0.00 & 0.00 & 1.00 & & \\ 0.00 & 0.00 & 0.00 & 0.49 & 1.00 & \\ 0.00 & 0.00 & 0.00 & 0.49 & 0.49 & 1.00 \end{pmatrix}, \mu_2^* = (1.400, 0.686, 0, 0, 0)'.$$

Substituting all these components into Equation 2.11, we get

$$\begin{aligned} F_{\text{MAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) =& q_1 \Big(\log |\Sigma_{0_m,1} \Sigma_1^{-1}| + \text{tr}((\Sigma_1^* + (\mu_1^* - \mu_{0_m,1})(\mu_1^* - \mu_{0_m,1})') \Sigma_{0_m,1}^{-1}) \\ & - \text{tr}((\Sigma_1^* + (\mu_1^* - \mu_1)(\mu_1^* - \mu_1)') \Sigma_1^{-1}) \Big) \\ & + q_2 \Big(\log |\Sigma_{0_m,2} \Sigma_2^{-1}| + \text{tr}((\Sigma_2^* + (\mu_2^* - \mu_{0_m,2})(\mu_2^* - \mu_{0_m,2})') \Sigma_{0_m,2}^{-1}) \\ & - \text{tr}((\Sigma_2^* + (\mu_2^* - \mu_2)(\mu_2^* - \mu_2)') \Sigma_2^{-1}) \Big) \\ =& (0.8)(1.200 + 5.248 - 5.706) + (0.2)(0 + 6.178 - 6.178) \\ =& 0.594. \end{aligned}$$

As before, because the first pattern is the complete data pattern, $\Sigma_{0_m,1} = \Sigma_0$, $\Sigma_1 = \Sigma$,

$\mu_{0_{m,1}} = \mu_0$, and $\mu_1 = \mu$. However, the addition of the truncated covariances and means (i.e., Σ_1^* and μ_1^*) into this equation means that the component of $F_{\text{MAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ corresponding to the first pattern is different from that of $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ in Cases 2 and 3. For the second pattern, as in Case 3, $\Sigma_{0_{m,2}}$ and Σ_2 are the 5×5 submatrices of Σ_0 and Σ with the first row and column deleted, and $\mu_{0_{m,2}}$ and μ_2 are the 5×1 subvectors of μ_0 and μ with the first element deleted. Because the misfit associated covariance between X_1 and X_2 is eliminated, $\Sigma_{0_{m,2}} = \Sigma_2$, so that the component of $F_{\text{MAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)$ corresponding to the second pattern is zero.

The population RMSEA is given by:

$$\text{RMSEA}_{\text{FIML}} = \sqrt{\frac{F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)}{df}} = \sqrt{\frac{0.594}{27}} = 0.148.$$

Overall, this set of examples shows that different missing mechanisms can yield different fit function minima and AFIs. We have not shown that missing data percentage, the number of missing data patterns, and the strength of the missing data mechanism (in case of MAR) can also affect the fit function minimum and the AFIs. These variables will be considered in the simulation study described in the next chapter.

3.2 Change in RMSEA due to Differences in Parameter Values

In this section, we will show that when data change from complete to incomplete, models with pseudo-parameters may have the same or different model-implied covariance matrix, which in turn affects fit function minimum and AFIs.

3.2.1 Case 1: Pseudo-parameter values stay the same with missing data

We first consider the case where the model-implied covariance matrix stays the same. Let X and Y be two random variables that follow a multivariate normal distribution with the population covariance matrix and mean vector given by

$$\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \text{ and } \mu = (0,0)'. \tag{3.3}$$

Let the hypothesized model be a special case of the simple regression model: $Y = X + \zeta$, where the regression coefficient is fixed to one. This model is misspecified (since X and Y are orthogonal in the population). Let $\text{var}(X) = \eta$ and $\text{var}(\zeta) = \psi$. We assume a saturated mean structure and a zero correlation between X and ζ. The model-implied covariance matrix and mean vector are then

$$\Sigma(\theta_0) = \begin{pmatrix} \eta & \eta \\ \eta & \eta + \psi \end{pmatrix} \text{ and } \mu_0 = (0,0)', \tag{3.4}$$

where the model parameters are given by $\theta_0 = (\eta, \psi)'$.

To obtain the "pseudo-parameters" θ_0 with complete data, we minimize

$$\begin{aligned} F_c(\mu_0, \Sigma(\theta_0)|\mu, \Sigma) &= \log|\Sigma(\theta_0)\Sigma^{-1}| + \text{tr}(\Sigma\Sigma^{-1}(\theta_0)) + (\mu - \mu_0)'\Sigma^{-1}(\theta_0)(\mu - \mu_0) - p \\ &= \log(4\psi\eta) + \frac{1}{\psi} + \frac{1}{2\eta} - 2, \end{aligned} \tag{3.5}$$

where the second expression has been obtained by substituting Equations 3.3 and 3.4 into

the first expression and then simplifying. The partial derivatives of $F_c(\mu_0, \Sigma(\theta_0)|\mu, \Sigma)$ are

$$\frac{\partial F}{\partial \eta} = \frac{1}{\eta} - \frac{1}{2\eta^2} \text{ and } \frac{\partial F}{\partial \psi} = \frac{1}{\psi} - \frac{1}{\psi^2}.$$

Setting them to 0, we obtain $\theta_0 = (\eta, \psi)' = (\frac{1}{2}, 1)'$. Substituting these values into Equations 3.4 and 3.5, we obtain

$$\Sigma_0 = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \text{ and } F_c(\mu_0, \Sigma_0|\mu, \Sigma) = 0.693. \tag{3.6}$$

The population RMSEA can then be computed using Equation 2.14:

$$\text{RMSEA}_{\text{ML}} = \sqrt{\frac{F_c(\mu_0, \Sigma_0|\mu, \Sigma)}{df}} = \sqrt{\frac{0.6931}{1}} = 0.148.$$

Now suppose there are 50% missing data on Y, missing completely at random (MCAR). In this case, $J = 2$, $p_1 = 2$, $p_2 = 1$, $q_1 = q_2 = 0.5$. The fit function becomes

$$\begin{aligned} F_{\text{MCAR}}(\mu(\theta_{0_m}), \Sigma(\theta_{0_m})|\mu, \Sigma, \phi) =& q_1\Big(\log|\Sigma_1(\theta_{0_m})\Sigma_1^{-1}| + \text{tr}(\Sigma_1\Sigma_1^{-1}(\theta_{0_m})) \\ &+ (\mu_1 - \mu_{0,1})'\Sigma_1^{-1}(\theta_{0_m})(\mu_1 - \mu_{0,1}) - p_1\Big) \\ &+ q_2\Big(\log|\Sigma_2(\theta_{0_m})\Sigma_2^{-1}| + \text{tr}(\Sigma_2\Sigma_2^{-1}(\theta_{0_m})) \\ &+ (\mu_2 - \mu_{0,2})'\Sigma_2^{-1}(\theta_{0_m})(\mu_2 - \mu_{0,2}) - p_2\Big) \\ =& \frac{1}{2}\Big(\log(4\psi\eta) + \frac{1}{\psi} + \frac{1}{2\eta} - 2\Big) \\ &+ \frac{1}{2}\big(\log(2\eta) + \frac{1}{2\eta} - 1\big), \end{aligned} \tag{3.7}$$

where $\Sigma_1 = \Sigma$, $\mu_1 = \mu$, $\Sigma_1(\theta_{0_m}) = \Sigma(\theta_{0_m})$, $\mu_{0,1} = \mu_0$, $\Sigma_2 = 0.5$, $\mu_2 = 0$, $\Sigma_2(\theta_{0_m}) = \eta$, and $\mu_{0,2} = 0$. The partial derivatives for this fit function are the same as those for complete data. Therefore, the "pseudo-parameters" and the model-implied matrix for this incom-

plete data is the same as those for complete data, shown in Equation 3.6. Substituting the parameters into Equation 3.7, we find that $F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m}|\mu, \Sigma, \phi) = 0.3459$, which makes $\text{RMSEA}_{\text{FIML}} = 0.5887$. Therefore, in this case, the fit function minima and RMSEAs for complete and incomplete data are different solely due to the differences in their equations.

3.2.2 Case 2: Pseudo-parameter values change with missing data

We now extend this example to the situation where the model-implied covariance matrix changes when there are missing data. The population covariance matrix and mean vector are again given by Equationi 3.3. The hypothesized model is again $Y = X + \zeta$, but we now fix $\text{var}(X) = 0.5$. The model-implied covariance matrix and mean vector are

$$\Sigma(\theta_0) = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & \psi + 0.5 \end{pmatrix} \text{ and } \mu_0 = (0,0)', \tag{3.8}$$

where the parameter is just $\theta_0 = \psi$.

When data are complete, the fit function we need to minimize is the same as Equation 3.5 except we can substitute 0.5 for η, so that

$$F_c(\mu_0, \Sigma(\theta_0)|\mu, \Sigma) = \log(2\psi) + \frac{1}{\psi} - 1. \tag{3.9}$$

The derivative with respect to ψ is again

$$\frac{dF}{d\phi} = \frac{1}{\psi} - \frac{1}{\psi^2}, \tag{3.10}$$

yielding $\theta_0 = \psi = 1$ and $F_c(\mu_0, \Sigma_0|\mu, \Sigma) = 0.6931$, which are the same as the ones in the previous example. The population RMSEA is different from the previous complete data

RMSEA because the df now equals to two:

$$\text{RMSEA}_{\text{ML}} = \sqrt{\frac{F_{\text{c}}(\mu_0, \Sigma_0 | \mu, \Sigma)}{df}} = \sqrt{\frac{0.6931}{2}} = 0.5887.$$

Next, suppose that there are 50% MCAR missing data on X. Therefore, again, $J = 2$, $p_1 = 2$, $p_2 = 1$, $q_1 = q_2 = 0.5$. The fit function becomes

$$\begin{aligned} F_{\text{MCAR}}(\mu(\theta_{0_m}), \Sigma(\theta_{0_m}) | \mu, \Sigma, \phi) =& q_1 \Big(\log |\Sigma_1(\theta_{0_m}) \Sigma_1^{-1}| + \text{tr}(\Sigma_1 \Sigma_1^{-1}(\theta_{0_m})) \\ & + (\mu_1 - \mu_{0,1})' \Sigma_1^{-1}(\theta_{0_m})(\mu_1 - \mu_{0,1}) - p_1 \Big) \\ & + q_2 \Big(\log |\Sigma_2(\theta_{0_m}) \Sigma_2^{-1}| + \text{tr}(\Sigma_2 \Sigma_2^{-1}(\theta_{0_m})) \\ & + (\mu_2 - \mu_{0,2})' \Sigma_2^{-1}(\theta_{0_m})(\mu_2 - \mu_{0,2}) - p_2 \Big) \\ =& \frac{1}{2} \big(\log(2\psi) + \frac{1}{\psi} - 1 \big) \\ & + \frac{1}{2} \big(\log(2\psi + 1) + \frac{1}{2\psi + 1} - 1 \big), \end{aligned} \tag{3.11}$$

where $\Sigma_1 = \Sigma$, $\mu_1 = \mu$, $\Sigma_1(\theta_{0_m}) = \Sigma(\theta_{0_m})$, $\mu_{0,1} = \mu_0$, $\Sigma_2 = 0.5$, $\mu_2 = 0$, $\Sigma_2(\theta_{0_m}) = \psi + 0.5$, and $\mu_{0,2} = 0$. Notice that this fit function is very different from Equation 3.9. Therefore, the derivative is also different from Equation 3.10:

$$\frac{dF}{d\psi} = \frac{1}{2}\big(\frac{1}{\psi} - \frac{1}{\psi^2}\big) + \frac{1}{2}\big(\frac{2}{2\psi + 1} - \frac{2}{(2\psi + 1)^2}\big). \tag{3.12}$$

Setting the derivative to 0, we find that $\theta_{0_m} = \psi = 0.7378$, [2] which yields

$$\Sigma_{0_m} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1.2378 \end{pmatrix} \text{ and } F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi) = 0.5274. \tag{3.13}$$

[2] Using the rational root theorem, we can show that the function in Equation 3.12 has no rational root. We solved for ψ by graphing the function.

With this fit function minimum, we obtain the population RMSEA:

$$\text{RMSEA}_{\text{FIML}} = \sqrt{\frac{F_{\text{MCAR}}(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma, \phi)}{df}} = \sqrt{\frac{0.5274}{2}} = 0.5135.$$

Therefore, in this case, the fit function minima and the RMSEA values are different between complete and incomplete data not only due to differences in equations but also due to the differences in the "pseudo-parameters" and the model-implied covariance matrices. In summary, we have shown in this section that when the data change from complete to incomplete, the "pseudo-parameters' may change, which in turn affects the fit function minimum and RMSEA.

Chapter 4

SEM AFIs under FIML Estimation: Simulation Studies

> A simulation is the imitation of the operation of a real-world process or system over time. Whether done by hand or on a computer, simulation involves the generation of an artificial history of a system and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system.
>
> *Jerry Banks, John S. Carson II, Barry L. Nelson, David M. Nicol, 2010*

In this chapter, we describe two large sample stimulation studies designed to demonstrate, with more realistic models, how factors such as the location of missing data relative to the location of misfit, the percentage of missing data, and the type of missing data mechanism affect the RMSEA and the CFI computed using FIML estimation. We focus on large samples to mimic what happens at the population level. It is important to first establish the differences in the population so that they are not obfuscated by the presence of sampling error.

4.1 Design

Table 4.1 shows the design of the two simulation studies. In both studies, data were generated from confirmatory factor analysis (CFA) models. The population model was always a two-factor model with six indicators per factor, all loadings of 0.7, and unit variances for all observed and latent variables. The population model varied in the value of the factor correlation and the number and size of correlated residuals (if any) across studies and study conditions. For each population model, we generated $n = 1000000$ normally distributed observations using the simulData() function in the lavaan package [30] in *R* (see Supplementary Materials for sample code).

The two studies differed in the type and location of misfit in the hypothesized model. In Study 1, we varied the number of correlated residuals (1 or 2), the size of the correlated residuals (0, 0.1, 0.2, 0.3, or 0.4) and the strength of the factor correlation (0, 0.4, or 0.8) in the population model. The hypothesized model was always a two-factor model without correlated residuals. Thus, misfit in the hypothesized model was most directly related to the indicators of the factor where correlated residuals appeared in the population model, although misfit can propagate throughout the model via the factor correlation (when it is not zero). In addition, the location of correlated residuals (i.e., location of misfit) was varied relative to the location of missing data: 1) in the "Same Factor" (SF) conditions, variables with correlated residuals and variables with missing data loaded on the same factor; [1] 2) in "Different Factor" (DF) conditions, variables with correlated residuals and variables with missing data loaded on different factors (see Figure 4.1). In Study 2, the population model did not have any correlated residuals; instead, it had a factor correlation of varying size (0.2 to 1, see Table 4.1). The hypothesized model was always a one-factor model. Thus, in Study 2, model misfit increased as the factor correlation in the

[1] In most SF conditions, the variables with missing data had correlated residuals in the population model. However, in the SF conditions where four variables have missing data but only two variables have a correlated residual, two of the variables with missing data will not include a correlated residual.

Table 4.1: Conditions in the Simulation Studies

Study 1	
Number of Variables with Missing Data (2 levels)	2, 4
Percentage of Missing Data in Each Variable with Missing Data (3 levels)	0%, 20%, 50%
Location of Misfit (2 levels)	Same factor (SF) conditions: Variables involving misfit and those involving missing data load on the same factor. Different factor (DF) conditions: Variables involving misfit and those involving missing data load on different factors.
Missing Mechanism (3 levels)	MCAR, Weak MAR, Strong MAR
Model ($2 \times 5 \times 3$ =30 levels)	The population model is a two-factor model (six indicators loading on each factor) that varies in the following features: • Number of correlated residuals: 1, 2 • Size of correlated residuals: 0, 0.1, 0.2, 0.3, 0.4 • Factor correlation: 0, 0.4, 0.8 The hypothesized model is always a correlated two-factor model without any correlated residuals.
Study 2	
Number of Variables with Missing Data (3 levels)	2, 4, 6
Percentage of Missing Data in Each Variable with Missing Data (3 levels)	0%, 20%, 50%
Number of Missing Data Patterns (2 levels)	Minimum: Always 2 patterns Maximum: 4, 16 and 64 patterns when 2, 4 and 6 variables have missing data, respectively.
Missing Mechanism (3 levels)	MCAR, Weak MAR, Strong MAR
Model (9 levels)	The population model is a two-factor model (six indicators loading on each factor) that varies in the factor correlation: 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2. The hypothesized model is always a one-factor model.

Note: Both studies have factorial designs. In total, there are $2 \times 3 \times 2 \times 3 \times 30 = 1080$ conditions in Study 1 and there are $3 \times 3 \times 2 \times 3 \times 9 = 486$ conditions in Study 2.

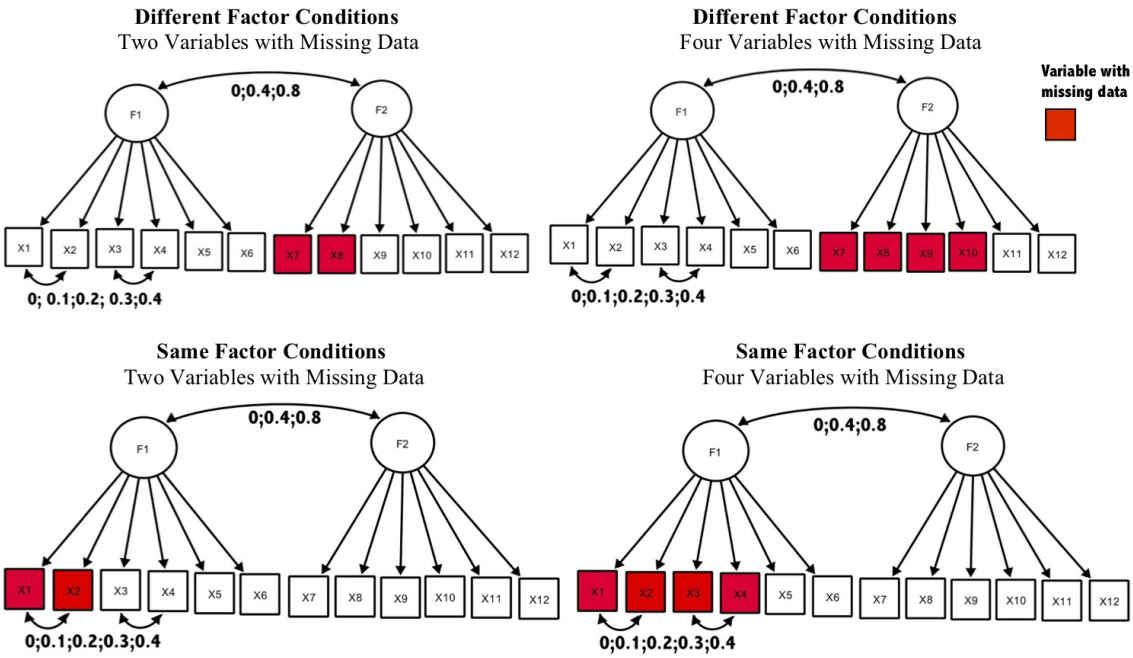


Figure 4.1: Differences between DF and SF conditions.

population model decreased. This type of misfit affects the entire covariance structure, but it particularly affects the covariances among indicators of different factors.

In both studies, we varied the percentage of missing data by deleting 0%, 20% or 50% of values in each variable designated to contain missing data. The number of such variables varied within and across studies (described shortly). In both studies, we studied three missing data mechanisms: one MCAR mechanism and two two types of MAR mechanism (weak and strong). To create MCAR data, we randomly selected rows for deletion. To generate MAR data, we first specified a cut-off point for a conditioning variable without missing data that loaded on the same factor. For 20% missing data, the cutoff point for the conditioning variable was 0.842 (i.e., the 20th quantile of the standard normal distribution); for 50% missing data, the cutoff point was 0. In the strong MAR conditions, the probability of missing data was 1 if the conditioning variable exceeded the specified cutoff, and 0 otherwise. In the weak MAR conditions, the probability of missing data was 0.75 if the conditioning variable exceeded the specified cutoff, and 0.25 otherwise.

The two studies differed in the number of variables with missing data and the number of missing data patterns. In Study 1, the number of variables with missing data was either 2 or 4; in Study 2, this number was either 2, 4 or 6. In Study 1, the variables with missing values were jointly missing, creating the minimum number of missing data patterns (i.e., two patterns). In Study 2, we also added the conditions where the variables were not jointly missing, resulting in the maximum number of possible patterns (see Table 1 for exact numbers). For MCAR conditions, data with the maximum number of possible patterns were created by creating missingness for each variable independently. For MAR conditions, such data were created by using different conditioning variable for each variable with missing data.

In summary, both studies have a factorial design, with 1080 conditions in Study 1 and 486 conditions in Study 2 (see 4.1). In both studies, we manipulated the degree of misfit, the amount of missing data, and the missing data mechanism. The main difference between the two studies is that in Study 1, we manipulated the location of misfit relative to the location of missing data; in Study 2, we fixed the location of misfit but manipulated the number of missing data patterns. The population values of the RMSEA and the CFI with complete data across different study conditions are given in Table 4.2. As this table illustrates, the complete data population RMSEA varies from 0 to 0.175 in Study 1 and from 0 to 0.192 in Study 2, while the complete data population CFI varies from 0.769 to 1 in Study 1 and from 0.538 to 1 in Study 2. The values in this table are taken to be benchmarks to which the incomplete data RMSEA and CFI will be compared. Our main question of interest is how much these values change with the introduction of incomplete data.

Table 4.2: Complete data RMSEA and CFI for all conditions in Studies 1 and 2

Study 1							
		FC=0		**FC=0.4**		**FC=0.8**	
Number of CRs	**Size of CR**	**RMSEA**	**CFI**	**RMSEA**	**CFI**	**RMSEA**	**CFI**
One CR	**0.0**	0.000	1.000	0.000	1.000	0.000	1.000
	0.1	0.022	0.994	0.022	0.994	0.023	0.994
	0.2	0.044	0.977	0.045	0.977	0.048	0.976
	0.3	0.067	0.950	0.069	0.948	0.076	0.943
	0.4	0.086	0.927	0.089	0.923	0.105	0.902
Two CRs	**0.0**	0.000	1.000	0.000	1.000	0.000	1.000
	0.1	0.033	0.987	0.033	0.987	0.034	0.988
	0.2	0.069	0.947	0.069	0.948	0.071	0.951
	0.3	0.113	0.874	0.113	0.876	0.115	0.882
	0.4	0.166	0.773	0.168	0.773	0.175	0.769

Study 2		
FC	**RMSEA**	**CFI**
1.0	0.000	1.000
0.9	0.045	0.979
0.8	0.078	0.932
0.7	0.106	0.872
0.6	0.129	0.803
0.5	0.149	0.730
0.4	0.167	0.654
0.3	0.182	0.586
0.2	0.192	0.538

Note: FC=Factor Correlation; CR=Correlated Residual; MCAR=Missing Completely At Random; MAR=Missing At Random.

4.2 Results

Below we summarized the major patterns of results using a series of figures and regression analyses. In all figures, the corresponding population quantities (i.e., RMSEA, CFI, or population fit function minimum) with complete data are shown by the red (solid) lines. The AFIs' values for complete data are also shown in Table 4.2; these values illustrate the benchmark with which we compare the incomplete data AFIs. For the regression analyses, we computed the absolute bias for the population AFIs by finding the absolute differences between the complete data population values (i.e., $RMSEA_{ML}$ or CFI_{ML} from

Equation 2.14) and the corresponding incomplete data population values (estimated from $n = 1000000$), and then we used the features of the missing data (shown in Table 4.1) to predict the absolute bias. In addition, we provided the full simulation results in the Supplementary Materials. [2]

4.2.1 Study 1

Figures 4.2-4.5 present selected results from Study 1. Figure 4.2 shows the RMSEA and CFI values for the conditions with MCAR data, population factor correlation of zero, and two correlated residuals of varying size (shown on the x-axis). Both DF ("Different Factors") and SF ("Same Factor") conditions are shown. The maximum discrepancy between complete and incomplete data AFIs occurs in the SF conditions with 50% missing data on four variables, when the correlated residuals are of size 0.4: the complete data RMSEA and CFI are 0.166 and 0.773, respectively, while the incomplete data RMSEA and CFI are 0.119 and 0.831, respectively.

Although the AFIs measure model fit on a continuum, cut-off points are commonly used to help researchers categorize the amount of misfit. For example, some methodologists have suggested RMSEA less than 0.08 indicate good fit [8], and CFI greater than 0.9 indicate good fit [17]. Figure 4.2 illustrates several conditions where missing data cause the AFIs to cross these recommended cutoff points. For example, in the SF conditions where four variables had missing data and the correlated residuals were of size 0.3, RMSEA decreased from 0.113 to 0.080 and CFI increased from 0.874 to 0.912 as the percentage of missing data increased from 0% to 50%. Thus, researchers may arrive at different conclusions about model fit depending on whether missing data are present.

Figure 4.2 also shows that the pattern of results for RMSEA is different from that

[2]The tables in the supplementary materials combine the results from these two simulation studies with those from the simulation studies in 6. For the simulation studies' results in this Chapter, please refer to the row under FIML and $n = 1000000$ in each table in the Supplementary Materials

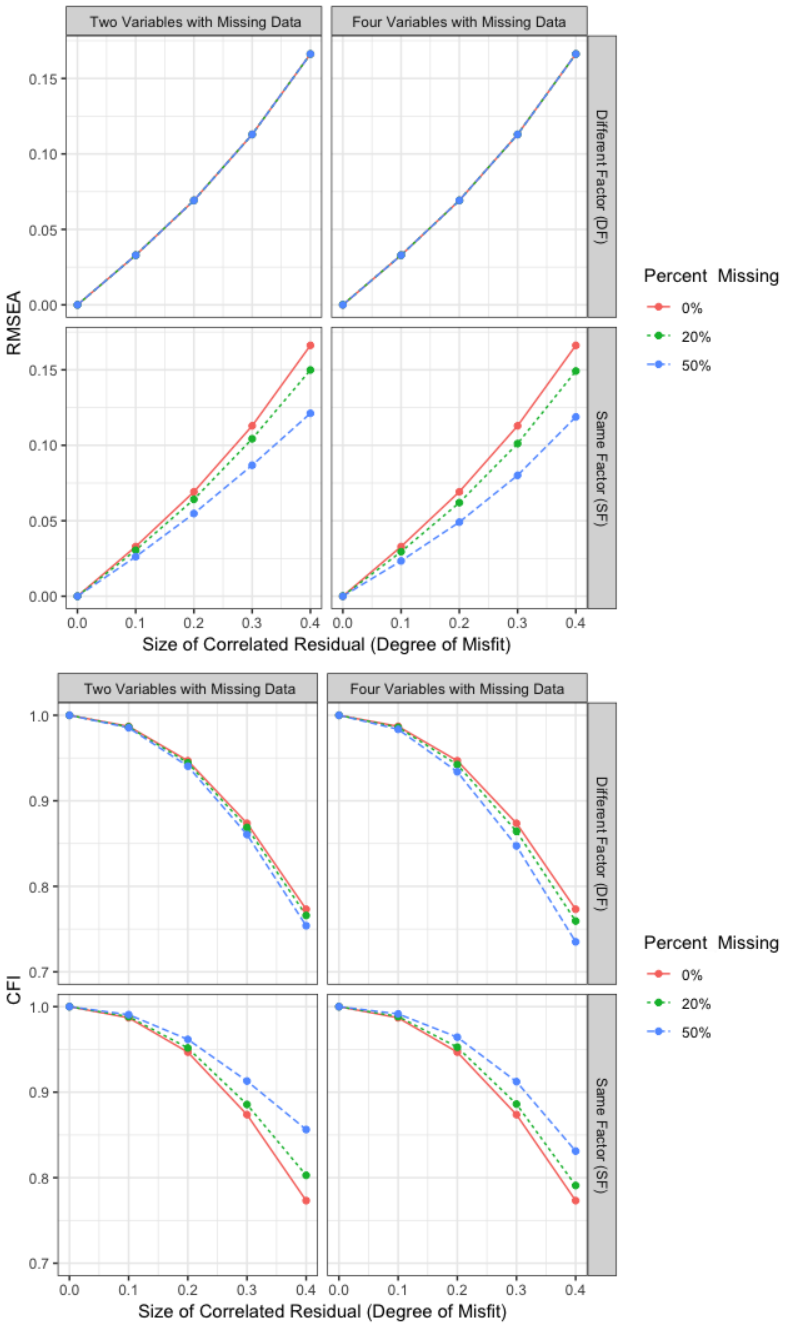


Figure 4.2: RMSEA and CFI for Study 1 conditions varying in the locations of misfit and number of variables with missing data. For the conditions in this figure, the missing mechanism is MCAR, the population factor correlation is 0, and the number of correlated residuals is 2.

for CFI. Because the factor correlation was zero, misfit associated with the covariance structure of indicators of one factor did not propagate to affect the covariance structure of the indicators of the other factor. Therefore, in the DF conditions where the indicators containing correlated residuals in the population model were different from indicators with missing data, the values of the RMSEA did not visibly change with missing data. In contrast, in SF conditions, the RMSEA values generally decreased, indicating better fit, as missing data increased (higher percentage of missing data or more variables with missing data), and the rate of decrease was higher for higher levels of misfit (i.e., larger size of correlated residuals).

The pattern was more complex for the CFI (see the second panel of Figure 4.2). In DF conditions, CFI decreased, indicating worse fit with more missing data. This pattern was opposite of that for the RMSEA. However, in the SF conditions, the CFI values increased with more missing data, indicating better fit. To explain this pattern of results, we examined the fit function minima for the hypothesized and baseline models separately. Figure 4.3 shows these values for the same conditions as in Figure 4.2. In the DF conditions, the fit function minimum for the hypothesized model stayed approximately the same with more missing data; however, it decreased for the baseline model with more missing data, especially for greater levels of misfit. The reason is that in the baseline model, which hypothesizes uncorrelated variables, the misspecification affected every part of the model, and was thus always entangled with the location of missing data. In the SF conditions, the fit function minima for both the hypothesized and the baseline models decreased with more missing data, but the rate of decrease for the hypothesized model was larger, especially for models with greater misfit. As a result, in the SF conditions, CFI increased with more missing data.

Figure 4.4 examines whether the patterns found in Figure 4.2 extend to the case when the factor correlation is not zero. In the conditions shown in Figure 4.4, the missing data

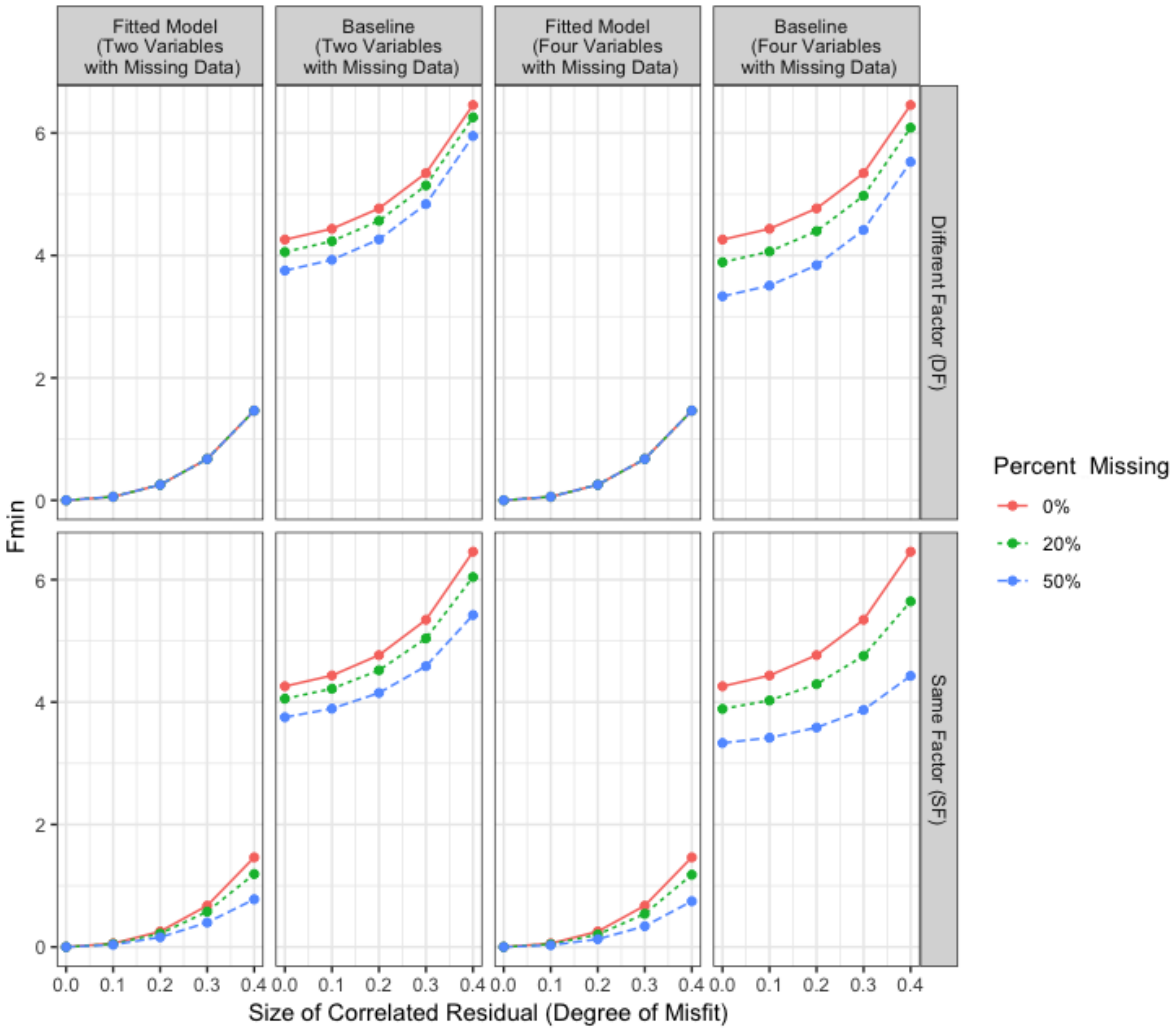


Figure 4.3: Fit function minima of the hypothesized and baseline models for selected conditions in Study 1. For the conditions in this figure, the missing mechanism is MCAR, the population factor correlation is zero, and the number of correlated residuals is two.

are still MCAR; the population model has one correlated residual and four variables with missing data; three values of the factor correlation (0, 0.4, and 0.8) are shown in separate panels. Note that the overall misfit is smaller in this figure compared to Figure 4.2, but the range of y-axis is kept the same to ensure comparability of figures. Interestingly, when the population factor correlation was non-zero, the RMSEA values still barely changed with the amount of missing data in the DF conditions: even when the correlated residual was 0.4 and the factor correlation was 0.8, the RMSEA changed from 0.105 to 0.102 as

the percentage of missing data changed from 0% to 50%. The finding that the value of the factor correlation in the population model had a small effect on the distortion in the incomplete data AFIs also generalized to other conditions of the study (see Supplementary Materials). Thus, even though misfit in the indicators of one factor can theoretically propagate across the factor correlation to affect the indicators of the other factor, this did not seem to actually occur to the degree that would affect the AFIs that much.

Finally, Figure 4.5 shows the impact of different missing data mechanisms on the AFIs in selected conditions (two correlated residuals, four variables with missing data, and factor correlation of 0.4). In these conditions, the largest change in the AFIs due to missing data occurred when the percentage of missing data was 50%, the missing mechanism was strong MAR, and the size of the correlated residuals was 0.4: the complete data RMSEA and CFI were 0.168 and 0.773, respectively, while the incomplete data RMSEA and CFI were 0.116 and 0.832, respectively. Overall, the patterns of change in the AFIs with more missing data were similar to those in Figure 4.2 and consistent across different missing mechanisms. However, the missing mechanism moderated the rate of change in the AFIs with the increasing percentage of missing data, although this effect was not always the same. For example, in the SF conditions, as the proportion of missing data increased from 0% to 20%, the RMSEA values in the weak MAR condition decreased at a *faster* rate than those in the MCAR and strong MAR conditions, but when the proportion of missing data increased from 20% to 50%, the RMSEA values for the weak MAR data decreased at a *slower* rate than those in the other two conditions. This effect of missing mechanism on the rate of change of the AFIs was similar in other study conditions not shown in Figure 4.5.

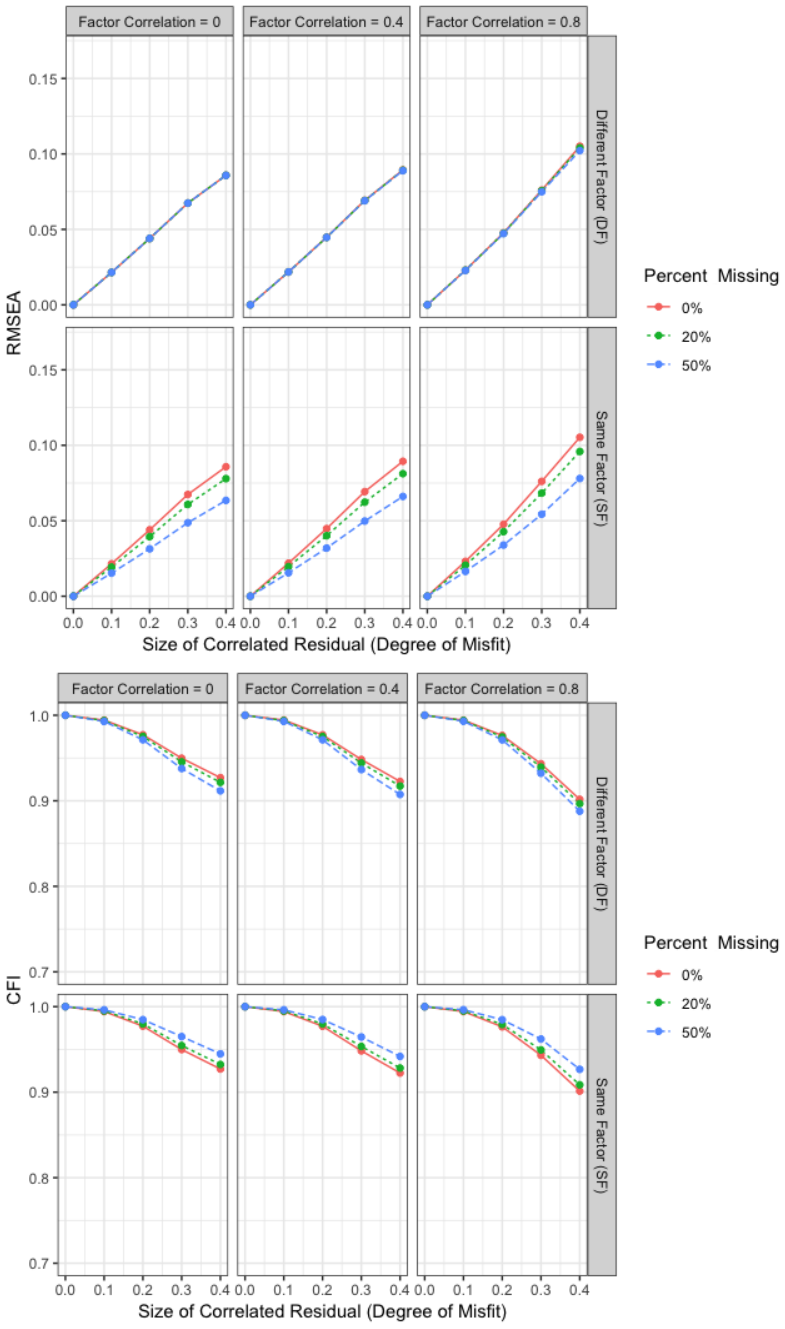


Figure 4.4: RMSEA and CFI for selected conditions in Study 1. For the conditions in this figure, the missing mechanism is MCAR. There is a single correlated residual in the population model, and the number of variables with missing data is four.

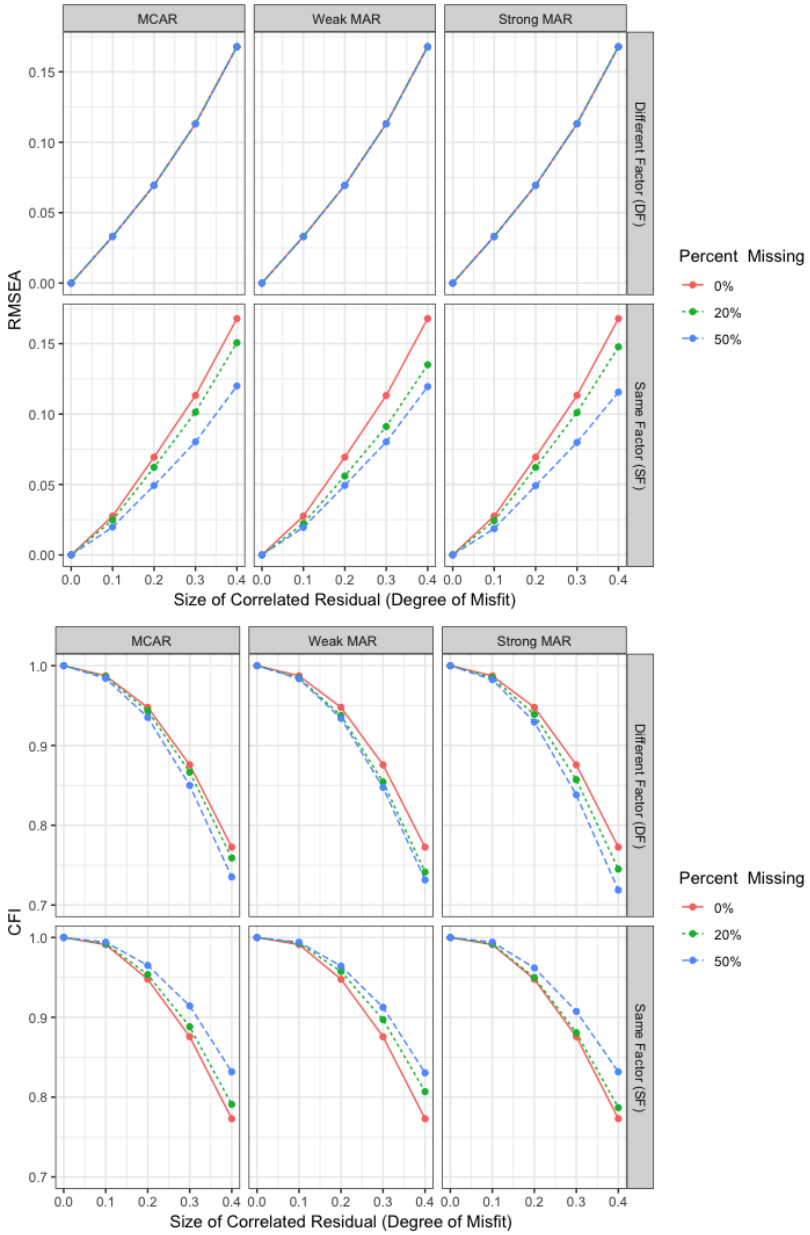


Figure 4.5: RMSEA and CFI for selected conditions in Study I. For conditions in this figure, the population factor correlation is 0.4; the number of correlated residuals is two, and the number of variables with missing data is four.

Regression Analyses

We also conducted regression analyses to examine whether missing data percentage, missing data mechanism, factor correlation in the population model, degree of misfit (measured by the size of the correlated residuals), location of misfit, and the interaction between the missing data percentage and location of misfit can predict the absolute bias of the AFIs. To simplify the analyses, we held the number of correlated residuals at two and the number of variables with missing data at four. [3] We coded the features of the missing data into factor or numeric variables, and explained these variables in Table 4.3.

Table 4.4 shows the full results of the regression analyses. Consistent with the results shown in Figures 4.4 and 4.5, missing data mechanism and factor correlation had very small effect on the absolute bias of the AFIs. Consistent with the results in Figures 4.2 and 4.3, there was an interaction between the percentage of missing data and the location of missing data. For RMSEA, when the location of missing data was DF, holding all other variables constant, the absolute bias did not change as missing data increased; however, when the location was SF, as missing data increased from 20% to 50%, the absolute bias, on average, increased by 0.012 unit. For CFI, when the location was DF, holding all other variables constant, the absolute bias increased by 0.008 unit on average; but when the location was SF, the absolute bias increased by 0.015 unit. In addition to the location and percentage of missing data, the degree misfit can have an effect on the bias, although its effect was relatively small.

Overall, Study 1 shows that the biggest distortion in the incomplete data AFIs relative to complete data AFIs occurred when the misspecification was large and when the amount of missing data was high (e.g., greater number of variables with missing data, greater pro-

[3] The number of correlated residuals and the size of the correlated residuals both measure the degree of misfit; to simplify the analyses, we held the number of correlated residuals constant and included the the size of the correlated residuals in the regression. Similarly, because the number of variables with missing data and the percentage of missing data both measure the amount of missing data, we only included the percentage of missing data in the analyses.

Table 4.3: Variables in the Regression Analyses

Study 1	
Missing	percentage of missing data in each variable with missing data. Missing is considered a categorical variable that equals 20 or 50 when the percentage missing is 20% or 50%, respectively; 20 is the reference group.
Location	location of misfit relative to location of missing data. Location is considered a categorical variable that is either SF or DF representing the SF or DF condition.
Mechanism	missing data mechanism. Mechanism is considered a numerical variable variable that equals to 0, 1 or 2 when the missing data mechanism is MCAR, weak MAR or strong MAR, respectively.
FactorCor	factor correlation in the population model. FactorCor is considered a numerical variable that equals to 0, 1 or 2 when the factor correlation is 0, 0.4 or 0.8, respectively.
Misfit	size of the correlation residuals. Misfit is considered a numerical variable that equals to 0, 1, 2, 3, or 4 when the size of the correlated is 0, 0.1, 0.2, 0.3, or 0.4, respectively.
Study 2	
Missing	percentage of missing data in each variable with missing data. Missing is considered a categorical variable that equals 20 or 50 when the percentage missing is 20% or 50%, respectively; 20 is the reference group.
Pattern	whether the missing data contain the minimum or maximum number of missing data patterns. Location is considered a categorical variable that is either min or max representing the conditions with either the minimum or maximum number of missing data patterns; min is the reference group.
Mechanism	missing data mechanism. Mechanism is considered a numerical variable that equals to 0, 1 or 2 when the missing data mechanism is MCAR, weak MAR or strong MAR, respectively.
Misfit	size of the factor correlation. Misfit is considered a numerical variable that equals to 0, 1, 2, 3, 5, 6, 7, 8 or 9 when the factor correlation is 1, 0.9, 0.8, 0.7, 0.6, 0.4, 0.3 or 0.2, respectively.

Table 4.4: Results of the Regression Analyses

Study 1
DF as the reference group for the Location variable:
$\text{BIAS}_{\text{RMSEA}} = -0.010 + 0.000\text{Missing} + 0.010\text{Location} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.005\text{Misfit} + 0.012(\text{Location})(\text{Missing})$
$\text{BIAS}_{\text{CFI}} = -0.010 + 0.008\text{Missing} + 0.001\text{Location} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.009\text{Misfit} + 0.006(\text{Location})(\text{Missing})$
SF as the reference group for the Location variable:
$\text{BIAS}_{\text{RMSEA}} = -0.001 + 0.012\text{Missing} + 0.010\text{Location} + 0.002\text{Mechanism} + 0.001\text{FactorCor} + 0.005\text{Misfit} + 0.012(\text{Location})(\text{Missing})$
$\text{BIAS}_{\text{CFI}} = -0.011 + 0.015\text{Missing} + 0.001\text{Location} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.009\text{Misfit} - 0.006(\text{Location})(\text{Missing})$
Study 2
$\text{BIAS}_{\text{RMSEA}} = -0.010 + 0.020\text{Missing} + 0.004\text{Pattern} + 0.001\text{Mechanism} + 0.005\text{Misfit}$
$\text{BIAS}_{\text{CFI}} = -0.011 + 0.040\text{Missing} + 0.008\text{Pattern} + 0.001\text{Mechanism} + 0.012\text{Misfit}$

Note: The coding of the variables is explained in Table 4.3.

portion of missing data on those variables). Another important variable was the location of misfit relative to the location of missing data: in the DF conditions, hardly any change was observed in the RMSEA. Finally, missing data mechanism can also be an important factor that determined how missing data affect the AFIs, but its effects were more subtle and more complicated.

4.2.2 Study 2

Figure 4.6 shows the RMSEA and CFI values for conditions with six variables with missing data. Because Study 2 includes conditions with greater misfit than Study 1 (see Table 4.2), the y-axis range in Figure 5 is greater than that in Figures 4.2-4.5. In this study, the hypothesized model was always the one-factor model, whereas the population model was a two-factor model without correlated residuals but with a factor correlation of varying size. In this case, the amount of misfit was directly related to the factor correlation in the population model; the x-axis in Figure 4.6 shows the population factor correlation varying

from least (correlation of 1) to most misfit (correlation of 0.2).

Figure 4.6 illustrates that the impact of missing data on AFIs can actually be quite large. In all panels of this figure, the curves corresponding to complete versus incomplete data RMSEA and CFI are much more widely separated than those in Study 1. This is due to both more severe model misspecification and to the inclusion of a condition where half (6 out of 12) of the variables contain missing data. In the most extreme case, in the weak MAR conditions with the maximum number of missing data patterns and the factor correlation of 0.2, the RMSEA changed from 0.192 to 0.114 (a 40% decrease) and CFI changed from 0.538 to 0.758 (a 41% increase) as the percentage of missing data per variable increased from 0% to 50%. In several of the shown conditions, the RMSEA crossed the recommended cutoff of 0.08 and CFI crosses the recommended cutoff of 0.9 as the percentage of missing data increased. For example, in the strong MAR conditions with the maximum number of missing data patterns and the factor correlation of 0.7, RMSEA decreased from 0.106 to 0.070 and CFI increased from 0.872 to 0.913 as missing data increased from 0% to 50%.

Several other patterns of results in Study 2 are noteworthy. First, the RMSEA decreased and CFI increased with missing data in all conditions. This pattern can be explained by the overlap between the location of misfit and the location of missing data. The misfit in Study 2 always affected all variables. This pattern was consistent with the results in Study 1, where RMSEA always decreased and CFI always increased with missing data in the SF conditions. Second, and consistent with Study 1, the missing mechanism affected the rate of change in the AFIs as the percent of missing data increased, and this effect of missing mechanism was different for different missing data percentage changes. Finally, as the percent of missing data increased, the RMSEA decreased and CFI increased at a faster rate in the conditions where the number of missing patterns was maximum. For example, for the weak MAR conditions with 0.2 factor correlation and six variables

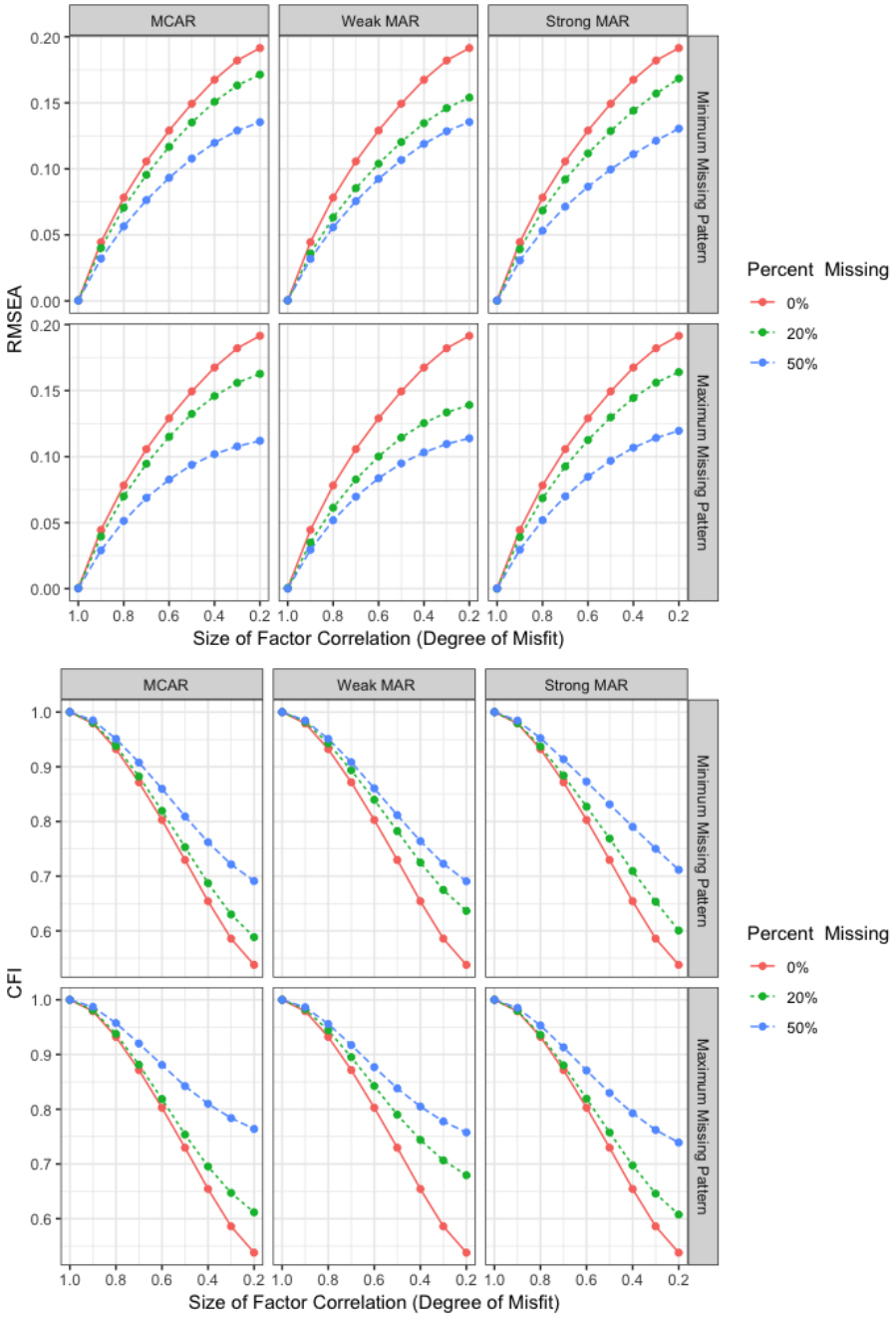


Figure 4.6: RMSEA and CFI for selected conditions in Study 2. For the conditions shown in this figure, the number of variables with missing data is six.

containing missing data (see Figure 4.6), when the number of missing patterns was maximum, the RMSEA decreased from 0.192 to 0.139 (a 27% decrease) and CFI increased from 0.538 to 0.679 (a 26% increase) as the missing data percentage changed from 0% to 20% . However, when the number of missing patterns was minimum, the RMSEA decreased from 0.192 to 0.154 (a 19% decrease) and CFI increased from 0.538 to 0.637 (a 18% increase). These patterns of results also hold in the conditions not shown in Figure 4.6; however, as the omitted conditions contain fewer variables (two or four) with missing data, the effects were smaller (see Supplementary Materials).

Regression Analyses

For the regression analyses, we examined whether missing data percentage, missing data mechanism, missing data pattern and degree of misfit (measured by the size of the factor correlation) can affect the absolute bias of the AFIs. To simplify the analyses, we held the number of variables with missing data at six.[4] Table 4.3 shows how we coded the features of the missing data in Study 2. Table 4.4 shows that the results of the analyses. The regression analyses showed the missing data mechanism and pattern had very little effect on the AFIs' bias. On the other hand, the percentage of missing data had the largest effect on the bias of the AFIs. For RMSEA, holding all variables constant, the absolute, on average, increased by 0.021 unit as missing data increased from 20% to 50%; for CFI, the bias increased by 0.040 unit as missing data increased from 20% to 50%. The degree of misfit also had an effect on the absolute bias. For RMSEA, holding all other variables constant, the absolute bias increased by 0.005 unit as the degree of misfit increased by one unit (i.e., as the factor correlation decreased by 0.1 unit; see Table 4.3); for CFI, the bias increased by 0.012 unit as the degree of misfit increased by one unit. These results imply that as the factor correlation decreased from 1 to 0.2, the RMSEA and CFI bias,

[4]Consistent with the regression analyses for Study 1, because the number of variables with missing data and the percentage of missing data both measure the amount of missing data, we only included the percentage of missing data in the analyses.

on average, increased by $0.005 \times 8 = 0.04$ and $0.012 \times 8 = 0.96$ unit, respectively; these changes were very substantial when considered on the metrics for RMSEA and CFI. In short, all these patterns of results by the regression analyses were perfectly consistent with those shown in Figure 4.6.

Overall, Study 2 further illustrated that the incomplete data AFIs are affected by many characteristics of missing data, such as the amount of missing data (in terms of the number of variables and the percentage of missing data per variable), missing data mechanism, and, new in this study, the number of missing data patterns. Because the misspecification was greater in this study, and because the type of misspecification (wrong number of factors) was such as to affect all variables, the patterns of results illustrating differences between complete and incomplete data AFIs were also more dramatic than they were in Study 1.

4.3 Discussion

The results from our simulation studies show that the impact of missing data on the values of the AFIs varies from trivial to quite dramatic. If the misfit in the hypothesized model is highly localized (e.g., an omitted correlated residual) and pertains to variables that are fully observed, the impact on the RMSEA can be almost zero. On the other hand, in the most extreme case when the hypothesized model is misspecified globally (1-factor model is fit to 2-factor data), the number of variables with missing data is high, and the missing mechanism is MAR, we have found that in some conditions the AFIs changed by as much as 40% when the missing data increased from 0% to 50%. Across all conditions, the minimum of the fit function for the hypothesized either stayed the same or decreased with the presence of missing data. While we are not yet able to offer an analytical proof to show our results pattern holds in general, the main pattern we have observed is the following: with more missing data, more information is lost about the misfit contained in the data

unless the model is correctly specified or the variables involved in the misspecification are distinct from variables with missing data. Since RMSEA is a direct function of the fit function minimum for the hypothesized model, RMSEA will generally stay the same or decrease with more missing data. We are not certain whether it is possible for RMSEA to increase with more missing data, but we have not been able to create an example where it does so.

The pattern for the CFI, which involves a comparison to the fit of the baseline model, was more complex. We found that CFI tended to indicate worse fit (decreased) when the location of misfit was localized to a few indicators of one factor, and variables with missing data loaded on a different factor. In this case, with more missing data, the fit function minimum for the baseline model decreased more relative to the fit function minimum for the hypothesized model, leading to an overall slight decrease in the CFI (see Figures 4.2 and 4.3). When misfit was entangled with the location of missing data, CFA, like RMSEA, always indicated better fit with missing data.

The impact of the percentage of missing data and the number of variables with missing data had a predictable effect on the AFIs: the differences relative to complete data AFIs became more dramatic. Increasing the number of missing data patterns tended to further distort the AFIs. The impact of the missing data mechanism was more nuanced: it seemed to primarily affect the initial distortion in the AFIs when moving from complete data to 20% missing data.

Our findings have direct implications for researchers who use FIML to handle missing data. While evaluating approximate model fit using AFIs is always nuanced and subjective, existing tentative cutoffs have all been developed with complete data [8, 17]. When researchers use FIML estimation to handle missing data, they should be cautious when using any firm cutoffs for AFIs. AFIs computed by the current software following FIML estimation do not necessarily reflect what researchers probably want to know: what the

amount of misfit would have been had the data been complete; in other words, the incomplete data AFIs are not consistent estimates of the population values of the complete data AFIs. Therefore, when compared to the existing cutoffs, the incomplete data AFIs may be inaccurate indicators of the amount of model misfit. In particular, the RMSEA, arguably the most popular index of approximate fit, may *under*estimate the amount of misfit with missing data. As shown in our simulation studies, the AFIs can cross the recommended cutoff points as missing data increase, and thus researchers could draw *opposite* conclusions about model fit depending on the amount of missing data present.

Chapter 5

Alternative Approaches for Computing AFIs

> An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
>
> *John Wilder Tukey*

As seen in the previous three chapters, the most popular FIML estimation can produce very distorted AFIs. Based on our simulation studies, in some cases, the population AFIs changed by as much as 40% when the percentage of missing data increased from 0% to 50%. In this chapter, we propose an alternative approach for computing AFIs following FIML estimation, which we refer to as the FIML-corrected or FIML-C approach. The FIML-C approach involves modifying the current computations of RMSEA and CFI so that they estimate what these AFIs would have been had the data been complete.

In addition to the FIML-C approach, we study another AFI computation approach, which involves the use of the two-stage (TS) estimation. In the typical TS procedure, the saturated model is fit to incomplete data in the first stage, and then the complete data fit function is minimized in the second stage, with the saturated model's estimates of means

and covariances replacing the sample means and covariances in this fit function [43, 46]. Because the complete data fit function is used to compute the AFIs, if the missing mechanism is ignorable, they should approach the same population values as if the data had been complete.

For both FIML-C and TS approaches, we also propose a series of small sample corrections that are meant to improve performance in small samples. These developments are based on the earlier work that proposed similar corrections to AFIs in the context of non-normal data [6, 7] and in the context of categorical data [34]. All of these corrections are derived through finding the expected value of the estimate of the population fit function minimum under a correctly specified model. In the last section of the chapter, we show the derivations of the small sample corrections for the FIML-C and TS approaches.

5.1 Alternative AFIs following FIML Estimation

The FIML-C approach allows us to approximate what the model fit would have been had there been no missing data; that is, FIML-C RMSEA and CFI approximate RMSEA_{ML} and CFI_{ML} (Equation 2.14) rather than $\text{RMSEA}_{\text{FIML}}$ and CFI_{FIML} (Equation 2.15). This approach to computing AFIs is an extension of the approach proposed by Savalei [34] for AFIs for categorical data. While it is an approximation rather than an exact approach, Savalei [34] found that the approximation works well for mild and moderate model specifications. We therefore also expect FIML-C AFIs to work well unless the model misspecifications are severe.

The FIML-C AFIs are computed according to the following equations:

$$\begin{aligned}
\widehat{\text{RMSEA}}_{\text{FIML-C}} &= \sqrt{\max\left(\frac{F_{\text{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{df}{n}}{df},0\right)} \\
\widehat{\text{CFI}}_{\text{FIML-C}} &= 1-\frac{\max\left(F_{\text{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{df}{n},0\right)}{\max\left(F_{\text{c}}(\hat{\mu}_B,\hat{\Sigma}_B|\tilde{\mu},\tilde{\Sigma})-\frac{df_B}{n},F_{\text{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{df}{n},0\right)},
\end{aligned} \tag{5.1}$$

where the subscript B stands for the baseline model, and $F_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})$ is the complete data ML fit function in Equation 2.2, with the saturated FIML estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ used in place of sample means and sample covariance matrix and evaluated at the structured FIML estimates $\hat{\mu}$ and $\hat{\Sigma}$. By evaluating the complete data ML fit function at the FIML parameter estimates, we approximate what the complete data ML fit function minimum would have been had there been no missing data.

However, Equation 5.1 may not suffice in small samples. Because the FIML parameter estimates are obtained using a different fit function than F_c, the degrees of freedom may no longer be the best estimate of the expected value of F_c evaluated at the FIML estimates. To correct for this bias, we can incorporate small sample corrections as follows:

$$\begin{aligned}
\widehat{\mathrm{RMSEA}}_{\text{FIML-C,s}} &= \sqrt{\max\left(\frac{F_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{k}{n}}{df},0\right)} \quad ; \\
\widehat{\mathrm{CFI}}_{\text{FIML-C,s}} &= 1-\frac{\max\left(F_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{k}{n},0\right)}{\max\left(F_{\mathrm{c}}(\hat{\mu}_B,\hat{\Sigma}_B|\tilde{\mu},\tilde{\Sigma})-\frac{k_B}{n},F_{\mathrm{c}}(\hat{\mu},\hat{\Sigma}|\tilde{\mu},\tilde{\Sigma})-\frac{k}{n},0\right)},
\end{aligned} \tag{5.2}$$

where k and k_B are the FIML-C correction terms for the hypothesized and baseline models, respectively [6, 7, 34]. In the most general form,

$$k = \mathrm{tr}(U_{\mathrm{m}}W_{\mathrm{m}}^{-1}W_{\mathrm{c}}W_{\mathrm{m}}^{-1}U_{\mathrm{m}}\Gamma), \tag{5.3}$$

where the subscript m indicates the matrix is related to the FIML fit function for the missing data and the subscript c indicates the matrix is related to the fit function for the complete data but evaluated at FIML estimates. Section 5.2.3 provides the derivation for this general equation for k. For each unique component in Equation 5.3, there are different options for estimating the matrix with sample data, resulting in different ways of estimating k. In this chapter, we examine six ways of estimating k. Table 5.1 lists the equations for these six versions of computations of k, and explains the components in the equations for

k. In the following paragraphs, we explain each component in Equation 5.3 as well as the six versions of k listed in Table 5.1. To make it easier to refer to the different FIML-C versions, we use FIML-C V0 to denote the FIML-C version without small sample correction, and use the FIML-C V1-V6 to denote the six FIML-C versions with small sample corrections.

The matrix W_{m} in Equation 5.3 is the population weight matrix used in the FIML estimation, or the "FIML weight matrix." The FIML weight matrix can be thought of as the information matrix for the saturated model. This information matrix can be observed or expected. However, for MAR data, the observed information matrix is the only asymptotically unbiased estimate [33]. Therefore, the observed information matrix is used for all versions of W_m (see "Estimate of FIML Weight Matrix" in Table 5.1). In addition, at the sample level, W_m can be evaluated at different sample estimates depending on which software you use. Some software such as EQS evaluate the weight matrix at the hypothesized model estimates ($\hat{\Sigma}$, $\hat{\mu}$) whereas others such as M*plus* evaluate it at the saturated model estimates ($\tilde{\Sigma}$, $\tilde{\mu}$) [42]. In *lavaan*, which is the software of our choice, by default, the weight matrix is evaluated with the hypothesized model estimates, but it can be changed to the saturated model estimates by setting mimic="Mplus". Although some researchers suggest the use of hypothesized model estimates over the saturated model estimates [42], there is no consensus which one is better; therefore, we vary this option across the versions. For FIML-C V1, V2, V4 and V5 in Table 5.1, the weight matrix is evaluated at the hypothesized model estimates (denoted as $\hat{W}_m$); for FIML-C V3 and V6, the weight matrix is evaluated at the saturated model estimates (denoted as $\tilde{W}_m$).

The matrix U_{m} is the residual weight matrix:

$$U_{\mathrm{m}} = W_{\mathrm{m}} - W_{\mathrm{m}}\hat{\Delta}(\hat{\Delta}' W_{\mathrm{m}}\hat{\Delta})^{-1}\hat{\Delta}' W_{\mathrm{m}}, \tag{5.4}$$

where $\hat{\Delta}$ is the matrix of model derivatives, always evaluated at the hypothesized model estimates. Since U_{m} is a function of W_m, we use $\hat{U}_m$ to denote the estimate of the residual weight matrix when $\hat{W}_m$ is substituted in Equation 5.4 (as in FIML-C V1, V2, V4, and V6), and use $\tilde{U}_m$ when $\tilde{W}_m$ is substituted (as in FIML-C V3 and V6; see "Estimate of Residual Weight Matrix" in Table 5.1).

The matrix Γ is the asymptotic covariance matrix of the saturated FIML estimates. Without normality assumption, Γ can be estimated using the fourth order moment of the data or it can be estimated using the sandwich method involving a triple product: $\Gamma = W_m^{-1} V_m W_m^{-1}$, where V_m is the FIML first order information matrix, and W_m and V_m are both evaluated at the saturated model estimates. In the special case when we assume normality *and* the hypothesized model is correctly specified, $\Gamma_0 = W_{\mathrm{m},0}^{-1}$ (i.e., asymptotically), and the correction will simplify. In this case, k is simplified to

$$k = \mathrm{tr}(W_c W_m^{-1} U_m W_m^{-1}). \tag{5.5}$$

Because we do not assume the model is correct when evaluating fit indices, this simplification is only an approximation even if we have normal data. We include this variation of Γ or k in our research. FIML-C V4-V6 in Table 5.1 assume normality and correctly specified model, and use Equation 5.5 for k, whereas FIML-C V1-V3 do not assume normality nor correctly specified model, and use Equation 5.3.

Lastly, the matrix W_c is the complete data weight matrix, which is the information matrix based on the complete data fit function. Similar to W_m, we can either evaluate W_c at the FIML hypothesized or saturated model estimates. Unlike W_m, both observed and expected information version of this matrix are asymptotically unbiased estimate for W_c. When the saturated model estimates are used for W_c, observed and expected versions are the same, therefore, there are only three options. For notations, we use $\hat{W}_c^{\mathrm{OBS}}$ (as in

FIML-C V1 and V4 in Table 5.1) to denote the observed information matrix evaluated at the hypothesized model estimates, $\hat{W}_c^{\text{EXP}}$ (as in FIML-C V2 and V5) to denote the expected information matrix evaluated at the hypothesized model estimates, and $\tilde{W}_c$ for the observed or expected information matrices evaluated at the saturated model estimates (as in FIML-C V3 and V6; see "Estimate of Complete Data Weight Matrix" in Table 5.1).

In the above paragraphs, we explained in detail the different computation versions for k. However, in order to compute the CFI, we also need to compute k_B, the correction factor for the baseline model. The variations for k_B are the same as the variations for k except that in the case of k_B, the "hypothesized model" is the baseline model (see Table 5.1) [1].

5.1.1 Population Limits for FIML-C AFIs

Because the correction terms k and k_B stay relatively the same as n increases, the population limits of these AFIs are the same regardless of whether small sample corrections are incorporated in the computation of sample FIML-C AFIs. The population limits of Equations 5.1 and 5.2 are given by:

$$\begin{aligned} \text{RMSEA}_{\text{FIML-C}} &= \sqrt{\frac{F_c(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma)}{df}}; \\ \text{CFI}_{\text{FIML-C}} &= 1 - \frac{F_c(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma)}{F_c(\mu_{B,0_m}, \Sigma_{B,0_m} | \mu, \Sigma)}, \end{aligned} \tag{5.6}$$

where μ_{0_m}, Σ_{0_m}, $\mu_{B,0_,}$ and $\Sigma_{B,0_m}$ are the population limits of the corresponding FIML model-implied means and covariance matrices. Comparing Equations 2.14 and 5.6, we can understand why the FIML-C approach is an approximation. In general, $\mu_0 \neq \mu_{0_m}$ and $\Sigma_0 \neq \Sigma_{0_m}$; that is, when the model is misspecified, the FIML parameter estimates may have different population limits from the corresponding ML estimates for the complete data.

[1] We note that for FIML, regardless of what the hypothesized model is, the saturated model estimates stay the same. This explains why for k_B, when W_m is evaluated at the saturated model estimates, there is no subscript B as shown in Tables 5.1.

Table 5.1: Equations for k and k_B for FIML-C versions

	Equation	
	Structured Model	Baseline Model
	Versions without the Normality Assumption	
FIML-C V1	$k = \text{tr}(\hat{U}_m \hat{W}_m^{-1} \hat{W}_c^{\text{OBS}} \hat{W}_m^{-1} \hat{U}_m \tilde{\Gamma})$	$k_B = \text{tr}(\hat{U}_{m,B} \hat{W}_{m,B}^{-1} \hat{W}_{c,B}^{\text{OBS}} \hat{W}_{m,B}^{-1} \hat{U}_{m,B} \tilde{\Gamma})$
FIML-C V2	$k = \text{tr}(\hat{U}_m \hat{W}_m^{-1} \hat{W}_c^{\text{EXP}} \hat{W}_m^{-1} \hat{U}_m \tilde{\Gamma})$	$k_B = \text{tr}(\hat{U}_{m,B} \hat{W}_{m,B}^{-1} \hat{W}_{c,B}^{\text{EXP}} \hat{W}_{m,B}^{-1} \hat{U}_{m,B} \tilde{\Gamma})$
FIML-C V3	$k = \text{tr}(\tilde{U}_m \tilde{W}_m^{-1} \tilde{W}_c \tilde{W}_m^{-1} \tilde{U}_m \tilde{\Gamma})$	$k_B = \text{tr}(\tilde{U}_{m,B} \tilde{W}_m^{-1} \tilde{W}_c \tilde{W}_m^{-1} \tilde{U}_{m,B} \tilde{\Gamma})$
	Versions with the Normality Assumption	
FIML-C V4	$k = \text{tr}(\hat{W}_c^{\text{OBS}} \hat{W}_m^{-1} \hat{U}_m \hat{W}_m^{-1})$	$k_B = \text{tr}(\hat{W}_{c,B}^{\text{OBS}} \hat{W}_{m,B}^{-1} \hat{U}_{m,B} \hat{W}_{m,B}^{-1})$
FIML-C V5	$k = \text{tr}(\hat{W}_c^{\text{EXP}} \hat{W}_m^{-1} \hat{U}_m \hat{W}_m^{-1})$	$k_B = \text{tr}(\hat{W}_{c,B}^{\text{EXP}} \hat{W}_{m,B}^{-1} \hat{U}_{m,B} \hat{W}_{m,B}^{-1})$
FIML-C V6	$k = \text{tr}(\tilde{W}_c \tilde{W}_m^{-1} \tilde{U}_m \tilde{W}_m^{-1})$	$k_B = \text{tr}(\tilde{W}_c \tilde{W}_m^{-1} \tilde{U}_{m,B} \tilde{W}_m^{-1})$
	Components of the Equation	
	Estimate of the Complete Data Weight Matrix	
$\hat{W}_c^{\text{OBS}}$	Observed information matrix, evaluated at hypothesized model estimates.	
$\hat{W}_c^{\text{EXP}}$	Expected information matrix, evaluated at hypothesized model estimates.	
$\tilde{W}_c$	Observed or expected information matrix, evaluated at saturated model estimates.	
$\hat{W}_{c,B}^{\text{OBS}}$	Observed information matrix, evaluated at baseline model estimates.	
$\hat{W}_{c,B}^{\text{EXP}}$	Expected information matrix, evaluated at baseline model estimates.	
	Estimate of the FIML Weight Matrix	
$\hat{W}_m$	Observed information matrix, evaluated at hypothesized model estimates.	
$\tilde{W}_m$	Observed information matrix, evaluated at saturated model estimates.	
$\hat{W}_{m,B}$	Observed information matrix, evaluated at baseline model estimates.	
	Estimate of the Residual Weight Matrix	
$\hat{U}_m$	$\hat{U}_m = \hat{W}_\text{m} - \hat{W}_m \hat{\Delta} (\hat{\Delta}' \hat{W}_m \hat{\Delta})^{-1} \hat{\Delta}' \hat{W}_m$ where $\hat{\Delta}$ is the matrix of hypothesized model derivatives, evaluated at the hypothesized model estimates.	
$\tilde{U}_m$	$\tilde{U}_\text{m} = \tilde{W}_\text{m} - \tilde{W}_\text{m} \hat{\Delta} (\hat{\Delta}' \tilde{W}_\text{m} \hat{\Delta})^{-1} \hat{\Delta}' \tilde{W}_\text{m}$.	
$\hat{U}_{m,B}$	$\hat{U}_{m,B} = \hat{W}_{m,B} - \hat{W}_{m,B} \hat{\Delta}_B (\hat{\Delta}_B' \hat{W}_{m,B} \hat{\Delta})^{-1} \hat{\Delta}_B' \hat{W}_{m,B}$ where Δ_B is the matrix of baseline model derivatives, evaluated at the baseline model estimates.	
$\tilde{U}_{m,B}$	$\tilde{U}_{m,B} = \tilde{W}_m - \tilde{W}_\text{m} \hat{\Delta}_B (\hat{\Delta}_B' \tilde{W}_m \hat{\Delta}_B)^{-1} \hat{\Delta}_B' \tilde{W}_m$.	
	Estimate of the Asymptotic Covariance Matrix of Saturated Model Estimates	
$\tilde{\Gamma}$	Without the normality and correct model assumptions, as in V1-V3, $\tilde{\Gamma} = \tilde{W}_m^{-1} \tilde{V}_m \tilde{W}_m^{-1}$ where $\tilde{V}_m$ is the FIML first order information matrix evaluated at the saturated model estimates. With these assumptions, $\tilde{\Gamma} = \tilde{W}_m^{-1}$, and thus V1-V3 are simplified to V4-V6.	

To put it in another way, recall that the original FIML approach has two problems that result in AFIs changing with missing data: 1) the population fit function minima between complete and incomplete data are different, 2) the "pseudo-parameters" between complete and incomplete data are different. The FIML-C approach can adjust for the differences in the fit function equations between incomplete and complete data but it cannot adjust for the differences in the "pseudo-parameters" between incomplete and complete data. However, when the model is only slightly misspecified, the "pseudo-parameters" between complete and incomplete data should be similar: $\mu_0 \approx \mu_{0_m}$ and $\Sigma_0 \approx \Sigma_{0_m}$. Therefore, the FIML-C approximation should work well in situations where the degree of model misfit is low.

5.1.2 Analytical Example for FIML-C Estimation

In this section, we use the same analytical examples we presented in Chapter 3 Section 3.2 to show how RMSEA is computed under FIML-C. Recall in the example in Section 3.2.1, the parameter values for incomplete data are the same as those for the complete value; that is, the fit function mimima and RMSEAs for complete and incomplete data are different solely due to the differences in their equations. As explained in the previous section, in this case, using the FIML-C approach, the $\text{RMSEA}_{\text{FIML-C}}$ value exactly equals to the RMSEA_{ML} for the complete data.

Now we explain in more detail the example in Section 3.2.2, where the parameter value changes with missing data. Recall in this example, the one model parameter, ψ. Under the FIML estimation, $\psi = 0.7378$. In addition, the complete data fit function equation is $F_c(\mu_0, \Sigma(\theta_0)|\mu, \Sigma) = \log(2\psi) + \frac{1}{\psi} - 1$ (see Equation 3.9). Therefore, by substituting $\psi = 0.7378$ into the complete data fit function equation, we can obtain the FIML-C RMSEA

as follows

$$\begin{aligned} \text{RMSEA}_{\text{FIML-C}} &= \sqrt{\frac{F_c(\mu_{0_m}, \Sigma_{0_m} | \mu, \Sigma)}{df}} \\ &= \sqrt{\frac{\log(2(0.7378)) + \frac{1}{0.7378} - 1}{2}} \\ &= 0.6101. \end{aligned} \tag{5.7}$$

Notice this FIML-C value is closer to the complete data $\text{RMSEA}_{\text{ML}} = 0.5887$ than the FIML $\text{RMSEA}_{\text{FIML}} = 0.5135$.

5.2 Alternative AFIs following TS Estimation

The TS approach is an alternative estimation method to FIML for incomplete data. The TS approach obtains parameter estimates by obtaining the saturated model estimates $\tilde{\mu}$, $\tilde{\Sigma}$ in the first stage, and then minimizes $F_c\big(\mu(\theta), \Sigma(\theta) | \tilde{\mu}, \tilde{\Sigma}\big)$, the complete data fit function with the "EM" estimates replacing $\bar{x}$ and S, in the second stage. Under the saturated model, the TS and FIML approaches obtain the same parameter estimates, but under the structured model, the two approaches obtain different parameter estimates.

While the TS approach is theoretically less efficient than FIML [43], it has been shown in simulation studies to perform very similarly to FIML [14, 36, 43], and it has some advantages. In this article, we focus on one such advantage: the TS approach naturally yields AFIs that approach desirable population values (i.e., those given by Equation 2.14) asymptotically. We will denote the TS estimates by $\breve{\theta}$, and the model-implied vector of means and covariance matrix obtained under this approach by $\breve{\mu}$ and $\breve{\Sigma}$. The TS AFIs

without a small sample correction (denoted as TS V0) are as follows:

$$\begin{aligned}\widehat{\text{RMSEA}}_{\text{TS}} &= \sqrt{\max\left(\frac{F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{df}{n}}{df},0\right)}\\ \widehat{\text{CFI}}_{\text{TS}} &= 1-\frac{\max\left(F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{df}{n},0\right)}{\max\left(F_{\text{c}}(\breve{\Sigma}_B,\breve{\mu}_B|\tilde{\Sigma},\tilde{\mu})-\frac{df_B}{n},F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{df}{n},0\right)}.\end{aligned} \tag{5.8}$$

The main difference between the FIML-C AFIs in Equation 5.1 and the TS AFIs in Equation 5.8 is that the TS approach involves minimizing the complete data fit function (which is minimized by $\breve{\theta}$) whereas the FIML-C approach uses an approximate minimum, obtained by evaluating the complete data fit function at the model-implied FIML estimates $\hat{\theta}$. This distinction has important theoretical consequences: as long as the missing data mechanism is ignorable, $\breve{\mu} \to \mu_0$ and $\breve{\Sigma} \to \Sigma_0$, so that the TS AFIs in Equation 5.8 have the population values given by Equation 2.14. Therefore, the TS AFIs naturally estimate what the fit would have been had the data been complete. In other words, at the population level, the TS approach provides an exact solution that is parallel to the complete data solution.

Although TS approach provides an exact solution at the population level, at the finite sample level, the TS approach still needs small-sample corrections to improve the estimate of AFIs. The corrections below parallel the corrections proposed for nonnormal data [6, 7]. We define small sample corrected TS AFIs as follows:

$$\begin{aligned}\widehat{\text{RMSEA}}_{\text{TS,s}} &= \sqrt{\max\left(\frac{F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{c}{n}}{df},0\right)}\\ \widehat{\text{CFI}}_{\text{TS,s}} &= 1-\frac{\max\left(F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{c}{n},0\right)}{\max\left(F_{\text{c}}(\breve{\Sigma}_B,\breve{\mu}_B|\tilde{\Sigma},\tilde{\mu})-\frac{c_B}{n},F_{\text{c}}(\breve{\Sigma},\breve{\mu}|\tilde{\Sigma},\tilde{\mu})-\frac{c}{n},0\right)},\end{aligned} \tag{5.9}$$

where c and c_B are the TS correction terms for the hypothesized and the baseline models, respectively. Here, $c = \text{tr}[U_{\text{c}}\Gamma]$, where U_c is the residual weight matrix obtained in the

second stage and Γ is the asymptotic covariance matrix of the saturated model (see Section 5.2.4 for detailed derivation). More specifically, $U_c = W_c - W_c\breve{\Delta}(\breve{\Delta}'W_c\breve{\Delta})^{-1}\breve{\Delta}'W_c$, where W_c and $\breve{\Delta}$ are the complete data weight matrix and model derivatives, respectively; $\Gamma = W_m^{-1}V_mW_m^{-1}$, where W_m and V_m are the FIML observed and first order information matrices obtained in the first stage of the TS method, respectively.

W_m is the weight matrix obtained in Stage 1 of the TS method, and V_m is the first order information matrix obtained in Stage 1; both W_m and V_m are evaluated at the saturated model estimates (denoted as $\tilde{W}_m$ and $\tilde{V}_m$, see Table 5.2) .

Similar to the FIML-C correction terms, TS correction terms can also be estimated in different ways. In our research, we examine two different computational versions of these corrections, which are shown in Table 5.2. In both versions, Γ is evaluated at the saturated model estimates. The difference between the two versions is that whether U_c is evaluated at the hypothesized or the saturated model estimates. In TS V1, U_c is evaluated at the hypothesized model estimates (denoted as $\hat{U}_c$), whereas in TS V2, U_c is evaluated at the saturated model estimates (denoted as $\tilde{U}_c$; see Table 5.2).

Finally, for CFI estimates, we need to compute c_B, the correction term for the baseline model. The two variations for the computations for c_B are the same as the variations for c except that for c_B, the "hypothesized model" is the baseline model (see Table 5.2).

5.2.1 Population Values for TS AFIs

Because the correction terms c and c_B stay relatively the same as n increases, the population limits of these AFIs are the same regardless of whether small sample corrections are incorporated in the computation of sample TS AFIs. As $n \to \infty$ in Equations 5.8 and 5.9,

the population limits of RMSEA and CFI are the following:

$$\begin{aligned} \text{RMSEA}_{\text{TS}} &= \sqrt{\frac{F_{\text{c}}(\mu_{0_{TS}}, \Sigma_{0_{TS}} | \mu, \Sigma)}{df}} \quad ; \\ \text{CFI}_{\text{TS}} &= 1 - \frac{F_{\text{c}}(\mu_{0_{TS}}, \Sigma_{0_{TS}} | \mu, \Sigma)}{F_{\text{c}}(\mu_{B,0_{TS}}, \Sigma_{B,0_{TS}} | \mu, \Sigma)}, \end{aligned} \tag{5.10}$$

where $\mu_{0_{TS}}$, $\Sigma_{0_{TS}}$, $\mu_{B,0_{TS}}$ and $\Sigma_{B,0_{TS}}$ are the population limits for $\breve{\mu}$, $\breve{\Sigma}$, $\breve{\mu}_B$ and $\breve{\Sigma}_B$, respectively. Theoretically, $\mu_{0_{TS}}$, $\Sigma_{0_{TS}}$, $\mu_{B,0_{TS}}$ and $\Sigma_{B,0_{TS}}$ equal to μ_0, Σ_0, $\mu_{B,0}$ and $\Sigma_{B,0}$ for the complete data; therefore, RMSEA_{TS} and CFI_{TS} should equal to RMSEA_{ML} and CFI_{ML} for the complete data in Equation 2.14.

5.2.2 Analytical Example for TS Estimation

In this section, we use the same analytical examples in Chapter 3 Section 3.2 to demonstrate why the TS parameter value, the fit function minimum and the RMSEA for incomplete data are the same as those for the complete data. Recall, in the two examples in Section 3.2 , the population covariance matrix and mean vector for the two random variables X and Y are given by (see Equation 3.3):

$$\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \text{ and } \mu = (0,0)'.$$

Now suppose there are 50% MCAR missing data on Y, the same as the example in Section 3.2.1. In the first stage of the TS approach, we need to find parameter values for the saturated model under the FIML estimation. It should be obvious that since the saturated model is always the correct model, the parameter values for the saturated model equal to the true parameter values (as shown in the above equation) for ignorable missing data under the FIML estimation. We can verify this by "manually" fitting a saturated

model. One way to specify a saturated model in the covariance structure is the following:

$$\Sigma_{\text{sat}}(\theta) = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \text{ and } \mu = (0,0)',$$

where the parameter vector is $\theta = (\alpha, \beta, \gamma)'$. In this case, the fit function we want to minimize becomes:

$$\begin{aligned} F_{\text{MCAR}}(\mu(\theta_{0_m}), \Sigma(\theta_{0_m}) | \mu, \Sigma, \phi) =& q_1 \Big(\log |\Sigma_1(\theta_{0_m}) \Sigma_1^{-1}| + \text{tr}(\Sigma_1 \Sigma_1^{-1}(\theta_{0_m})) \\ & + (\mu_1 - \mu_{0,1})' \Sigma_1^{-1}(\theta_{0_m})(\mu_1 - \mu_{0,1}) - p_1 \Big) \\ & + q_2 \Big(\log |\Sigma_2(\theta_{0_m}) \Sigma_2^{-1}| + \text{tr}(\Sigma_2 \Sigma_2^{-1}(\theta_{0_m})) \\ & + (\mu_2 - \mu_{0,2})' \Sigma_2^{-1}(\theta_{0_m})(\mu_2 - \mu_{0,2}) - p_2 \Big) \\ =& \frac{1}{2} \Big(\log(4\alpha\gamma - 4\beta^2) + \frac{\gamma}{2\alpha\gamma - 2\beta^2} + \frac{\alpha}{2\alpha\gamma - 2\beta^2} - 2 \Big) \\ & + \frac{1}{2} \big(\log(2\gamma) + \frac{0.5}{\gamma} - 1 \big). \end{aligned}$$

From here, it is obvious that if $\theta = (\alpha, \beta, \gamma)' = (0.5, 0, 0.5)'$ which are the true parameter values in Equation 3.3, then $F_{\text{MCAR}}(\mu(\theta_{0_m}), \Sigma(\theta_{0_m}) | \mu, \Sigma, \phi) = 0$. Therefore, in the first stage of the TS approach, we can obtain the true population covariance matrix for the complete data through fitting the saturated model. In the second stage, we use the true population covariance matrix to minimize the complete data fit function. It is obvious that by doing so, we obtain the complete data fit function minimum and thus the complete data RMSEA: $\text{RMSEA}_{\text{TS}} = \text{RMSEA}_{\text{ML}} = 0.5887$.

The TS RMSEA for the second example in Section 3.2 can be shown to be the same as the complete data RMSEA using the same logic. In short, for ignorable missing data, the TS approach allows us to first estimate the true population covariance matrix for the complete data, which can then be used to estimate the complete data fit function minima

Table 5.2: Equations for c and c_B for TS versions

Equation		
	Structured Model	Baseline Model
TS V1	$c = \text{tr}(\hat{U}_c\tilde{\Gamma})$	$c_B = \text{tr}(\hat{U}_{c,B}\tilde{\Gamma})$
TS V2	$c = \text{tr}(\tilde{U}_c\tilde{\Gamma})$	$c_B = \text{tr}(\tilde{U}_{c,B}\tilde{\Gamma})$
Components of the Equation		
Estimate of the Residual Weight Matrix		
$\hat{U}_c$	$\hat{U}_c = \hat{W}_c - \hat{W}_c\breve{\Delta}(\breve{\Delta}'\hat{W}_c\breve{\Delta})^{-1}\breve{\Delta}'\hat{W}_c$ where $\hat{W}_c$ is the complete data observed information matrix obtained in Stage 2, evaluated at the hypothesized model estimates from Stage 2, and $\breve{\Delta}$ is the matrix of the hypothesized model derivatives, also evaluated at the hypothesized model estimates.	
$\tilde{U}_c$	$\tilde{U}_c = \tilde{W}_c - \tilde{W}_c\breve{\Delta}(\breve{\Delta}'\tilde{W}_c\breve{\Delta})^{-1}\breve{\Delta}'\tilde{W}_c$ where $\tilde{W}_c$ is the complete data observed information matrix obtained in Stage 2, evaluated at the saturated model estimates.	
$\hat{U}_{c,B}$	$\hat{U}_{c,B} = \hat{W}_{c,B} - \hat{W}_{c,B}\breve{\Delta}_B(\breve{\Delta}_B'\hat{W}_{c,B}\breve{\Delta}_B)^{-1}\breve{\Delta}_B'\hat{W}_{c,B}$, where $\hat{W}_{c,B}$ is the complete data observed information matrix obtained in Stage 2, evaluated at the baseline model estimates from Stage 2, and $\breve{\Delta}_B$ is the matrix of the baseline model derivatives, also evaluated at the baseline model estimates.	
$\tilde{U}_{c,B}$	$\tilde{U}_{c,B} = \tilde{W}_c - \tilde{W}_c\breve{\Delta}_B(\breve{\Delta}_B'\tilde{W}_c\breve{\Delta}_B)^{-1}\breve{\Delta}_B'\tilde{W}_c$	
Estimate of the Asymptotic Covariance Matrix of Saturated Model Estimates		
$\tilde{\Gamma}$	$\tilde{\Gamma} = \tilde{W}_m^{-1}\tilde{V}_m\tilde{W}_m^{-1}$, where $\tilde{W}_m$ and $\tilde{V}_m$ are the FIML observed and first order information matrices, respectively, both obtained in Stage 1 and evaluated at the saturated model estimates.	

and AFIs.

5.2.3 Derivation of Small Sample Correction in FIML-C

Let $\tilde{\beta} = (\text{vech}\tilde{\Sigma}', \tilde{\mu}')'$ and let $\hat{\beta} = \beta(\hat{\theta}) = (\text{vech}\hat{\Sigma}', \hat{\mu}')'$. When we assume that the hypothesized model is true in the population (i.e., there exists a θ_0 such that $\beta_0 = \beta(\theta_0)$), the following approximation holds in the population (e.g., Shapiro [38], Yuan and Bentler [43]):

$$\sqrt{n}(\tilde{\beta} - \hat{\beta}) = \sqrt{n}W_{\text{m},0}^{-1}U_{\text{m},0}(\tilde{\beta} - \beta_0) + o_p(1),$$

where $W_{\text{m},0}$ is the FIML information matrix and $U_{\text{m},0} = W_{\text{m},0} - W_{\text{m},0}\Delta(\Delta' W_{\text{m},0}\Delta)^{-1}\Delta' W_{\text{m},0}$ is the FIML residual weight matrix, where Δ is the matrix of model derivatives. We also have $\sqrt{n}(\tilde{\beta} - \beta_0) \rightarrow N(0, \Gamma_0)$. When we further assume normality, $\Gamma_0 = W_{\text{m},0}^{-1}$. Even though $\hat{\theta}$ does not minimize F_{c}, when the hypothesized model is correct, we can approximate

$$\hat{F}_{\text{c}} = (\tilde{\beta} - \hat{\beta})' W_{\text{c},0}(\tilde{\beta} - \hat{\beta}) + o_p(1) \approx (\tilde{\beta} - \beta_0)' U_{\text{m},0} W_{\text{m},0}^{-1} W_{\text{c},0} W_{\text{m},0}^{-1} U_{\text{m},0}(\tilde{\beta} - \beta_0),$$

where $W_{\text{c},0}$ is the complete data information matrix, evaluated at the FIML parameter values. The distribution of $n\hat{F}_c$ can then be approximated by a mixture of independent 1 degree of freedom chi-square variates with weights given by the eigenvalues of $U_{\text{m},0}W_{\text{m},0}^{-1}W_{\text{c},0}W_{\text{m},0}^{-1}U_{\text{m},0}\Gamma_0$, and the approximate expected value of $n\hat{F}_c$ is given by the trace of this matrix product, which is k in Table 5.1.

5.2.4 Derivation of Small Sample Correction in TS

Assuming null hypothesis model is true, the chi-square test statistic in the second stage of the TS method is $T_{TS} = (n)F_c \approx (n)(\tilde{\beta} - \beta_0)' U_c (\tilde{\beta} - \beta_0)$, where F_c is the minimum of the complete data normal theory fit function in the second stage, and U_c is the residual weight matrix obtained in the second stage; that is, $U_{\text{c}} = W_{\text{c}} - W_{\text{c}}\breve{\Delta}(\breve{\Delta}' W_{\text{c}}\breve{\Delta})^{-1}\breve{\Delta}' W_{\text{c}}$, where W_c and $\breve{\Delta}$ are the complete data normal theory weight matrix and model derivatives, respectively.

Asymptotically, $\sqrt{n}(\tilde{\beta} - \beta_0) \sim N(0, \Gamma)$, where Γ is the asymptotic covariance matrix of the estimates of parameters for the saturated model.

Asymptotically, $\sqrt{n}(\tilde{\beta} - \beta_0) \sim N(0, \Gamma_0)$, where Γ_0 is the asymptotic covariance matrix of the estimates of parameters for the saturated model obtained in the first stage. Therefore, the distribution of $n\hat{F}_c$ can then be approximated by a mixture of independent one degree of freedom chi-square variates with weights given by the eigenvalues of $U_{c,0}\Gamma_0$, and the approximate expected value of $n\hat{F}_c$ is given by the trace of this matrix product. Therefore, $c = \mathrm{tr}(U_{c,0}\Gamma_0)$.

Chapter 6

SEM AFIs under FIML-C and TS Estimations: Simulation Studies

> Bias and efficiency are the two most important statistical concepts when considering a parameter estimator. These two concepts provide the rationale for comparing different estimation methods. If method A generates less biased and more efficient parameter estimates than method B, then A is better than B. Because there is usually a reason for a statistical method to be developed in the first place, it might be hard to have a best method under all circumstances. But we might prefer method A if it yields better estimators than method B in most circumstances.
>
> *Ke-Hai Yuan, Xin Tong, Zhiyong Zhang, 2015*

In this chapter, we present the results of two simulations studies that investigate the performance of the newly proposed FIML-C and TS AFIs and to compare them with the FIML AFIs currently in use. These simulation studies are a follow-up for the simulations conducted in Chapter 3. Through these simulation studies, we aim to show that relative to the complete data AFIs, FIML-C and TS AFIs are not biased or less biased than FIML

AFIs.

6.1 Design

The design of the two simulation studies are the same as the simulation studies in Chapter 3 except more sample sizes are studied (see Table 4.1). There are four levels of sample size: $n = 200, 500, 1000, 1000000$. The conditions with $n = 200, 500, 1000$ are meant to simulate sample data with small, medium or large sample size; for these sample data conditions, we generated 1000 replications of normally distributed observations using the simulData() function in the lavaan package [30] in *R*. The conditions with $n = 1000000$ are meant to mimic the population so that FIML, FIML-C, and TS AFIs can be compared without the influence of sampling fluctuations; for these population conditions, we generated a single dataset of normally distributed observations.

In total, Studies 1 and 2 had 4320 and 1944 conditions, respectively.[1] For each condition, we computed FIML RMSEA and CFI that are currently in use, the seven versions of the FIML-C AFIs, and the three versions of the TS AFIs.[2] For the population data conditions, we computed population bias by subtracting the complete data population values (i.e., RMSEA_{ML} or CFI_{ML} from Equation 2.14) from the corresponding incomplete data population values. For the sample data conditions ($n = 200, 500, 1000$), we computed empirical bias and empirical standard error of the RMSEA or CFI estimates across the replications. In addition, we calculated the root mean square error (RMSE) using the

[1] Table 4.1 shows that for the simulation studies in Chapter 4, there were 1080 and 486 conditions for Studies 1 and 2, respectively. In the current simulation studies, we added 4 levels for sample size; therefore, there were $1080 \times 4 = 4320$ and $486 \times 4 = 1944$ conditions for Studies 1 and 2, respectively.

[2] The TS V0 AFIs were computed by setting estimator="two-stage" inside the cfa() function in *lavaan* and then inspecting the model fit to get the AFIs. All versions of FIML-C and small sample corrections to TS AFIs were implemented using custom *R* code that made use of *lavaan*'s internal functions. Sample code for implementing different versions of the FIML-C and TS approaches can be found in the Supplementary Materials

following equation:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{1000}(\text{AFI}_{i,\text{simu}} - \text{AFI}_{\text{ML}})^2}{1000}},$$

where $\text{AFI}_{i,\text{simu}}$ is the ith simulated RMSEA or CFI value, and AFI_{ML} is either RMSEA_{ML} or CFI_{ML} from Equation 2.14. RMSE is a joint measure of bias and efficiency of an estimator and provides information regarding the bias-variance trade-off associated with an estimator.

6.2 Results

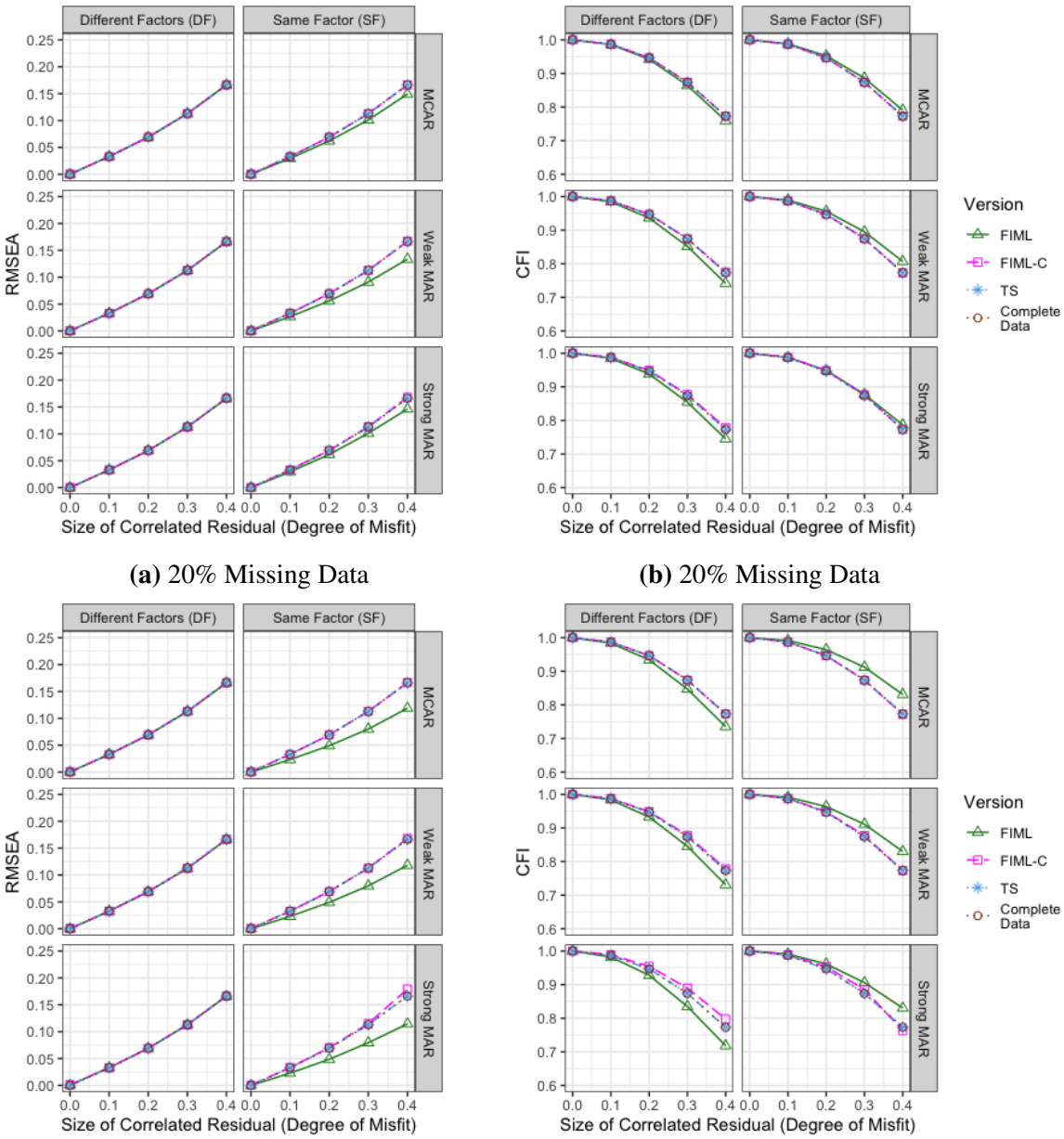
We summarized the results using figures and regression analyses to demonstrate the main patterns observed across the two studies. Since the patterns of results for FIML had been discussed extensively in Chapter 4, we mainly focused on the patterns of results for FIML-C and TS in this Chapter. We also provided the full simulation results for all conditions across the two studies in the table format in the Supplementary Materials.

6.2.1 Population Behavior

Figures for Population AFIs

Figures 6.1 and 6.2 show the population RMSEA and CFI values (estimated from $n = 1000000$) under the FIML, FIML-C (V0) and TS (V0) approaches in selected conditions of Studies 1 and 2. Small sample corrections disappear asymptotically, and thus the results for FIML-C V1-V6 and TS V1-V2 are not shown here because they are equal to FIML-C (V0) and TS (V0).

As shown in Figures 6.1 and 6.2, at the population level, the FIML AFIs that are currently implemented in popular software tended to exhibit a relatively large bias relative to the corresponding complete data AFIs unless the model had perfect fit. An exception



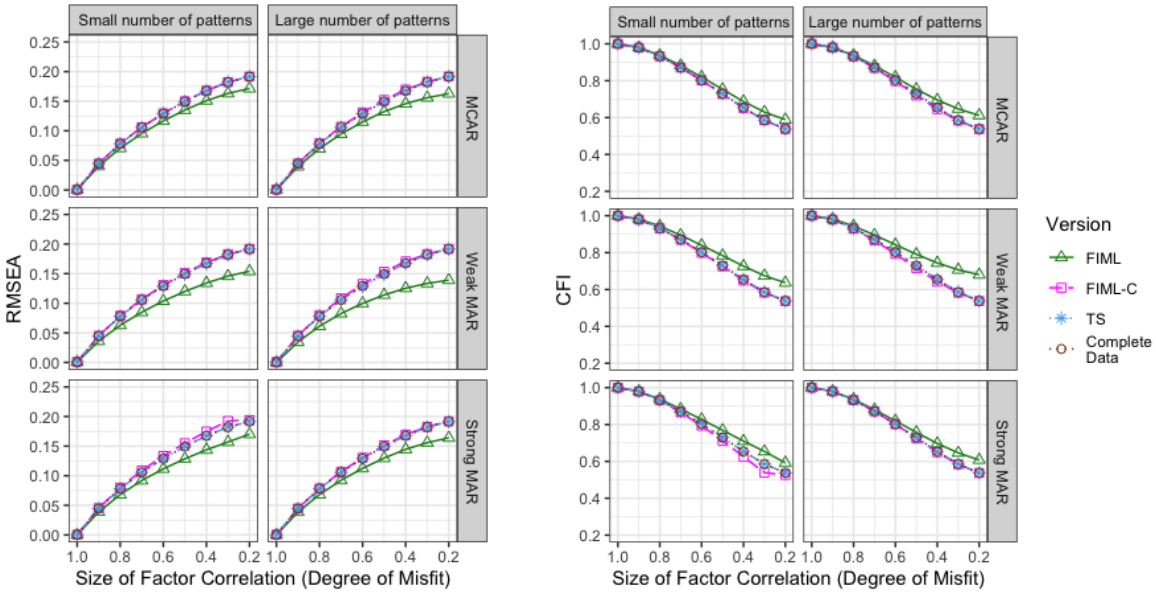
(a) 20% Missing Data

(b) 20% Missing Data

(c) 50% Missing Data

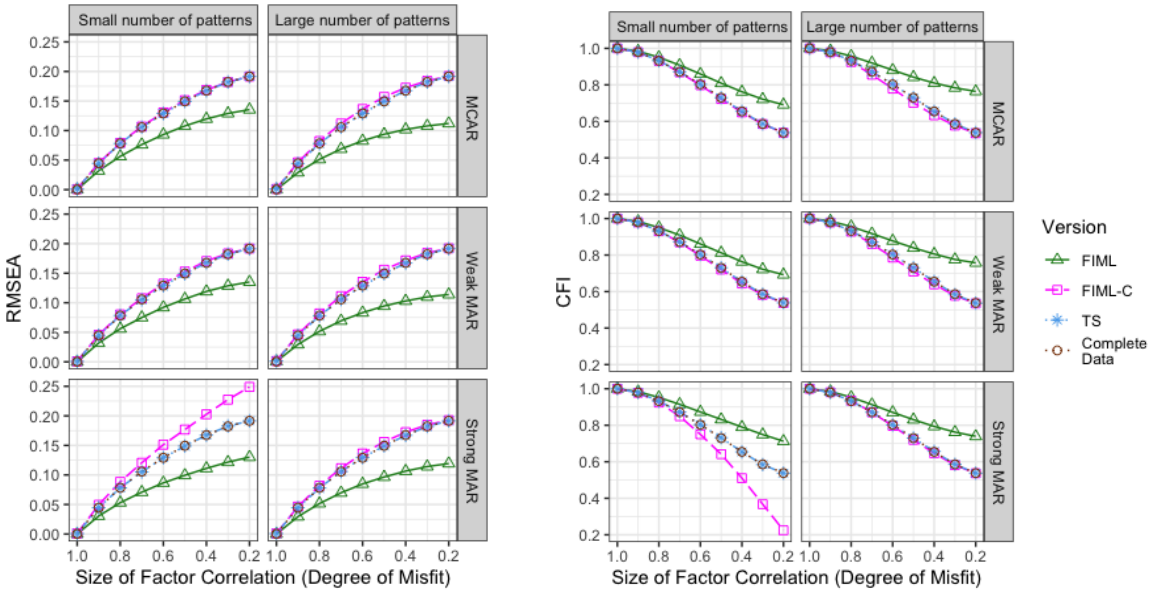
(d) 50% Missing Data

Figure 6.1: Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 1 comparing FIML, FIML-C and TS approaches. Complete data population RMSEA and CFI are also included for comparison. In these selected conditions, the number of variables with missing data is four, the number of correlated residuals is two, and the population factor correlation is zero. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals.



(a) 20% Missing Data

(b) 20% Missing Data



(c) 50% Missing Data

(d) 50% Missing Data

Figure 6.2: Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 2 comparing FIML, FIML-C and TS approaches. Complete data population RMSEA and CFI are also included for comparison. In these selected conditions, there are six variables that have missing data. The population model is a two-factor model with varying sizes for the factor correlation shown on the x-axis. The hypothesized model is a one-factor model.

was the RMSEA values in the conditions where the location of misfit was different from that of missing data (i.e., in the DF conditions). In contrast, the TS AFIs were the same as the complete data AFIs in all conditions; in other words, the TS AFIs asymptotically approached the complete data population AFI values, would be theoretically expected.

The FIML-C AFIs had very little bias in conditions with a smaller percentage of missing data (see Figures 6.1ab and 6.2ab). In conditions with a large percentage of missing data (i.e., 50% missing data), the population FIML-C AFIs were generally still a good approximation to the complete data AFIs, except when the mechanism was strong MAR and the degree of model misfit was high (see Figures 6.1cd and 6.2cd). There also appeared to be some dependence on the number of missing data patterns. For example, as shown in Figure 6.2cd, when the number of missing patterns was small and the missing data was strong MAR, at a high degree of model misfit (i.e., when the complete data RMSEA and CFI were 0.192 and 0.538, respectively), the FIML-C RMSEA was 0.057 higher than the complete data RMSEA and the FIML-C CFI was 0.313 lower than the complete data CFI. In other words, in these conditions, the FIML-C AFIs indicated worse model fit relative to the corresponding complete data AFIs. The reason for the poor performance of these FIML-C AFIs is due to the fact that in these conditions, the "pseudo-parameters" (estimated from $n = 1000000$) under FIML were very different from those for the corresponding complete data (see Table 6.1 for an example). Recall that the FIML-C approach uses parameter estimates from FIML, and, therefore, cannot correct for the differences in the FIML "pseudo-parameters" between complete and incomplete data.

Regression Analyses for Population AFIs

For the regression analyses of the population AFIs, we used the features of the missing data to predict the absolute bias of AFIs.[3] We coded the features of the missing data according

[3]The reason why we used the absolute bias instead of the raw bias is that with the raw bias, the negative regression coefficient may mean a decrease in the magnitude of the bias or may mean an increase in the

Table 6.1: "Pseudo-Parameters" for Complete and Incomplete Data under FIML

Factor Loadings for Complete Data	Factor Loadings for Incomplete Data
0.243	0.662
0.287	0.616
0.154	0.590
0.416	0.589
0.200	0.615
0.407	0.642
0.749	1.372
0.634	0.921
0.805	1.142
0.845	1.220
0.693	0.890
0.772	1.309

Note: The condition presented in the table involves strong MAR data with the minimum number of missing data patterns. The population model is a two-factor model (6 indicators loading on each factor) with a factor correlation of 0.2. The hypothesized model is a one-factor model with 12 indicators.

to Tables 4.3 and 6.2. For Study 1, the predictors included the estimation method, percentage of missing data, missing data mechanism, factor correlation in the population model and degree of misfit; for Study 2, the predictors included estimation method, percentage of missing data, missing data pattern, missing data mechanism and degree of misfit.[4]

The results of the regression analyses, presented in Table 6.3, showed that both FIML-C and TS methods had less absolute bias relative to FIML but TS method showed more reduction in bias. Specifically, for RMSEA, holding all other variables constant, the bias of FIML-C, on average, was 0.016 unit less than FIML in both studies; the bias of TS, on average, was 0.016 and 0.019 unit less than FIML in Studies 1 and 2, respectively. For CFI, the bias of FIML-C decreased by 0.014 and 0.029 unit in Studies 1 and 2, respectively; the

magnitude of the bias but in the negative direction.

[4]To simplify the analyses in Study 1, we held the number of correlated residuals at two, the number of variables with missing data at two and the location of missing data at SF. To simplify the analyses Study 2, we held the number of variables at six.

Table 6.2: Additional Variables in the Regression Analyses

Population AFIs	
Est	estimation method for AFIs. Est is considered a categorical variable that equals FIML, FIML-C or TS; FIML is the reference group.
Sample AFIs	
Est	estimation method for AFIs. Est is considered a categorical variable that equals FIML, FIML-C V3 or TS V2; FIML is the reference group.
Sample	sample size. Sample is considered a categorical variable that equals to 200 or 500; 200 is the reference group

Table 6.3: Results of the Regression Analyses for Bias in the Population

Study 1
$\text{BIAS}_{\text{RMSEA}} = 0.007 - 0.016\text{EstFIML-C} - 0.016\text{EstTS} + 0.004\text{Missing} + 0.001\text{Mechanism} + 0.000\text{FactorCor} + 0.003\text{Misfit}$
$\text{BIAS}_{\text{CFI}} = 0.005 - 0.014\text{EstFIML-C} - 0.016\text{EstTS} + 0.006\text{Missing} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.004\text{Misfit}$
Study 2
$\text{BIAS}_{\text{RMSEA}} = -0.004 - 0.016\text{EstFIML-C} - 0.019\text{EstTS} + 0.008\text{Missing} + 0.001\text{Pattern} + 0.001\text{Mechanism} + 0.002\text{Misfit}$
$\text{BIAS}_{\text{CFI}} = -0.001 - 0.029\text{EstFIML-C} - 0.038\text{EstTS} + 0.017\text{Missing} + 0.000\text{Pattern} + 0.003\text{Mechanism} + 0.005\text{Misfit}$

Note: The coding of the variables is explained in Tables 4.3 and 6.2.

bias of TS decreased by 0.016 and 0.038 unit in Studies 1 and 2, respectively. In addition, holding other variables constant, the percentage of missing and the degree of misfit had effects on the bias of AFIs, although their effects were relatively small. All of these results were consistent with those observed in Figures 6.1 and 6.2.

Summary for Population AFIs

In summary, at the population level, the TS approach performed very well in all conditions. The FIML-C approach performed similarly well in most conditions but for conditions with a large percentage of strong MAR data and a small number of patterns, the FIML-C

approximation began to fail, producing population AFIs with values that departed from the corresponding complete data AFI values. We note that in the conditions where FIML-C tended to fail, the complete data RMSEA ranged from the 0.149 to 0.191, and the complete data CFI ranged from 0.730 to 0.538.

6.2.2 Finite Sample Behavior

Figures for Sample AFIs

Figures 6.3-6.8 show selected results for the sample data conditions (i.e., $n = 200, 500$ and 1000). Figures 6.3-6.4 show the results for all computational versions of the proposed AFIs in selected conditions. Figures 5-8 provide additional results for the best performing versions. In order to present results from a variety of conditions, for each of Figures 6.3-8, we select different sets of conditions from either Study 1 or Study 2. The patterns of results presented in these figures generalize to those in the other conditions (see Supplementary Materials).

Figure 6.3 shows the empirical bias in selected sample data conditions for Study 1. Figure 6.4 shows the RMSE in selected sample data conditions in Study 2. Due to the large variability in both bias and RMSE values across the studied AFIs, the range of values on the y-axes in Figures 6.3 and 6.4 is very large. In these figures, we use color to indicate the versions of FIML-C and TS with large bias or RMSE in order to highlight the poorly performing versions; the well-performing versions are shown in grey (differences among the well-performing versions will be inspected more closely in Figures 5-8).

One noticeable pattern of results is that FIML-C V5 was one of the worst performing methods . The FIML-C V5 CFI estimates tend to be negatively biased (see Figure 6.3), and both RMSEA and CFI estimates tend to have large RMSE values, especially at small sample sizes (see Figure 6.4). Particularly, the bias of CFI under FIML-C V5 can be more than three times higher than that of other approaches, and the RMSE of both

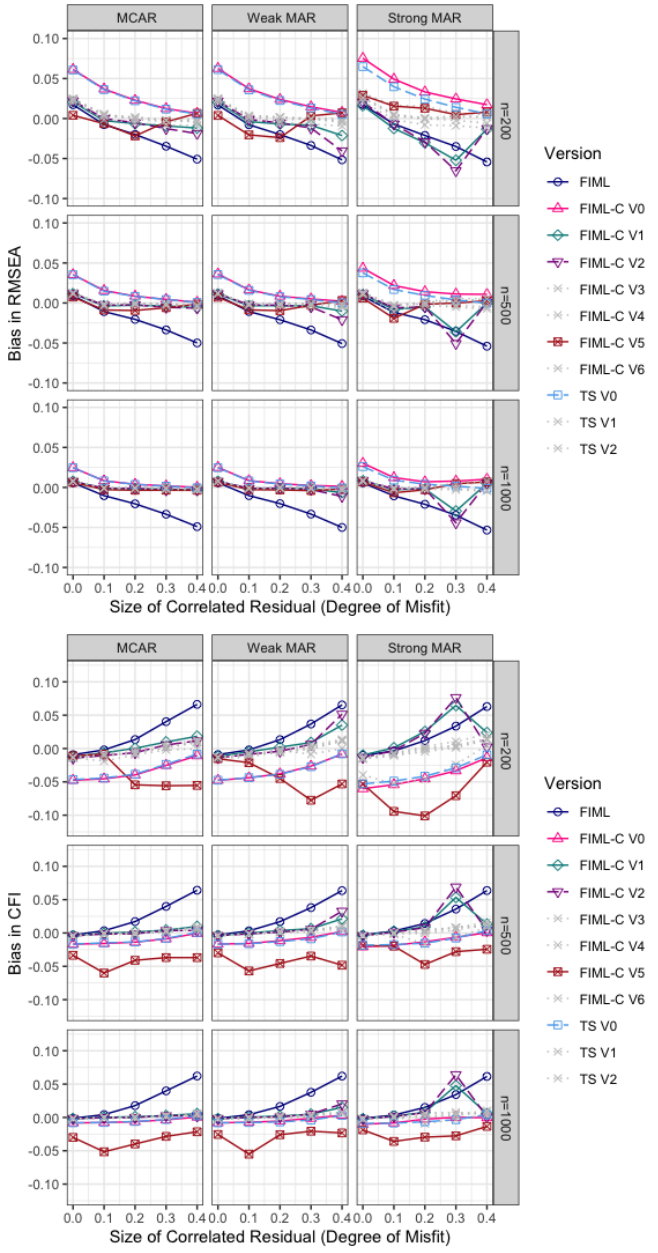


Figure 6.3: Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing FIML, FIML-C and TS approaches. In these conditions, the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, the percentage of missing is 50%, and the location of misfit is the same as the location of missing data. The population model is a two-factor model with varying sizes for the correlated residuals shown on the *x*-axis. The hypothesized model is a two-factor model without any correlated residuals.

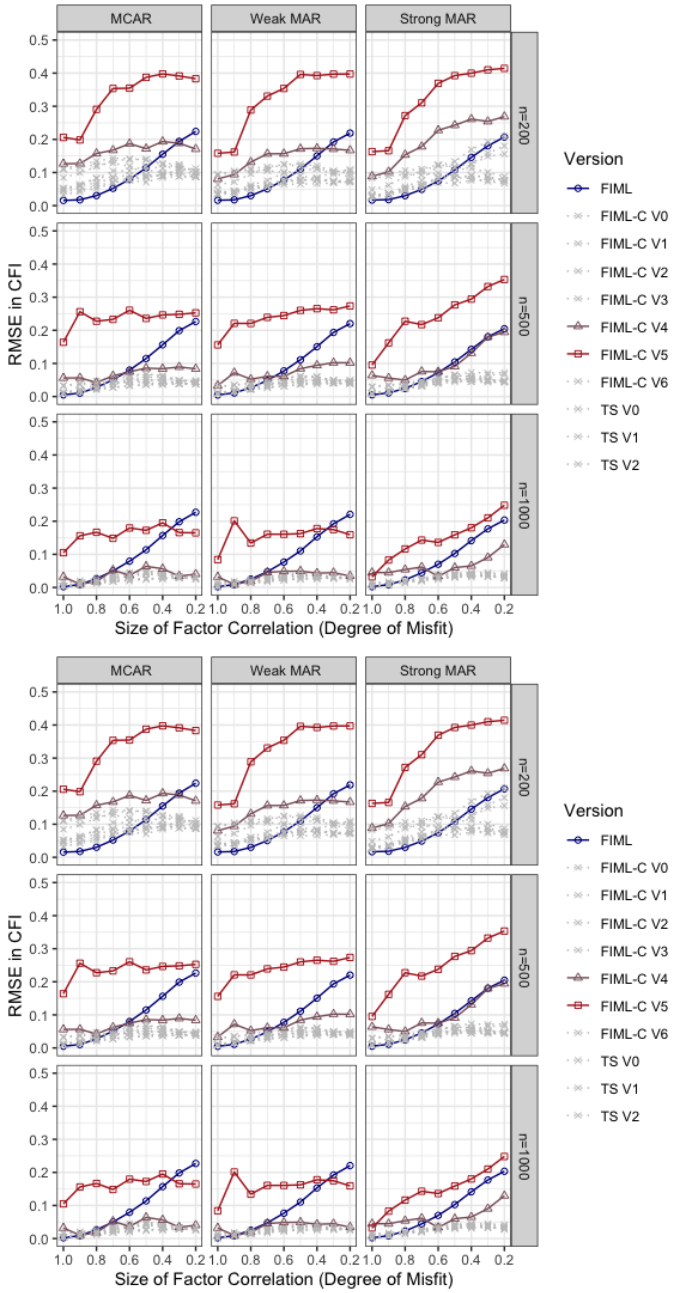


Figure 6.4: Root mean square error (RMSE) in the sample RMSEA and CFI estimates for selected conditions in Study 2 comparing FIML, FIML-C and TS approaches. In these conditions, the number of variables with missing data is six, the percentage of missing is 50%, and the number of patterns is large. The population model is a two-factor model with varying sizes for the factor correlation shown on the *x*-axis. The hypothesized model is a one-factor model.

RMSEA and CFI can be more than four times higher than those under other approaches. The large RMSE values were mainly due to the large standard errors of the RMSEA and CFI estimates under FIML-C V5, which can be three times higher than those under other approaches (see Supplementary Materials).

In addition to FIML-C V5, there were several other AFIs that did not perform very well. FIML-C V4 also produced AFIs with large RMSE values in some conditions. Although the RMSE values of FIML-C V4 were not as large as those of FIML-C V5, they were often larger than those of FIML AFIs, especially with small samples ($n = 200$; see Figure 4). Further, both FIML-C V0 and TS V0 tended to produce large and similar bias values in some conditions. Specifically, the AFIs under both FIML-C V0 and TS V0 had relatively large bias in the direction of indicating better fit when the sample size was small (i.e., $n = 200$) and the degree of model misfit was low (i.e., when complete data RMSEA was less than 0.08 and CFI greater than 0.95; see Figure 6.3). Both FIML-C V0 and TS V0 are estimation methods without small sample corrections, so their poor performances in small samples were expected. Finally, the AFIs under FIML-C V1 and V2 had noticeably large bias values in conditions where the degree of model misfit was relatively high. For example, in the strong MAR conditions where the complete data population RMSEA and CFI were around 0.12 and 0.87 respectively, the bias of the CFI and RMSEA estimates under FIML-C V1 and V2 was around 0.05, nearly as large as the bias under the original FIML approach. In conclusion, Figures 3 and 4 reveal that FIML-C V0, V1, V2, V4 and V5 as well as TS V0 methods did not perform well, and the top performing methods across all conditions were FIML-C V3, FIML-C V6, TS V1 and TS V2. FIML-C V3 and V6 were based on saturated model estimates of all matrices involved, either assuming normality (V6) or not assuming normality (V3; see Table 5.1).

The remaining figures compare bias and RMSE values among the top four best-performing AFIs in a variety of selected conditions in Studies 1 and 2. Figures 6.5 and 6.6 show the

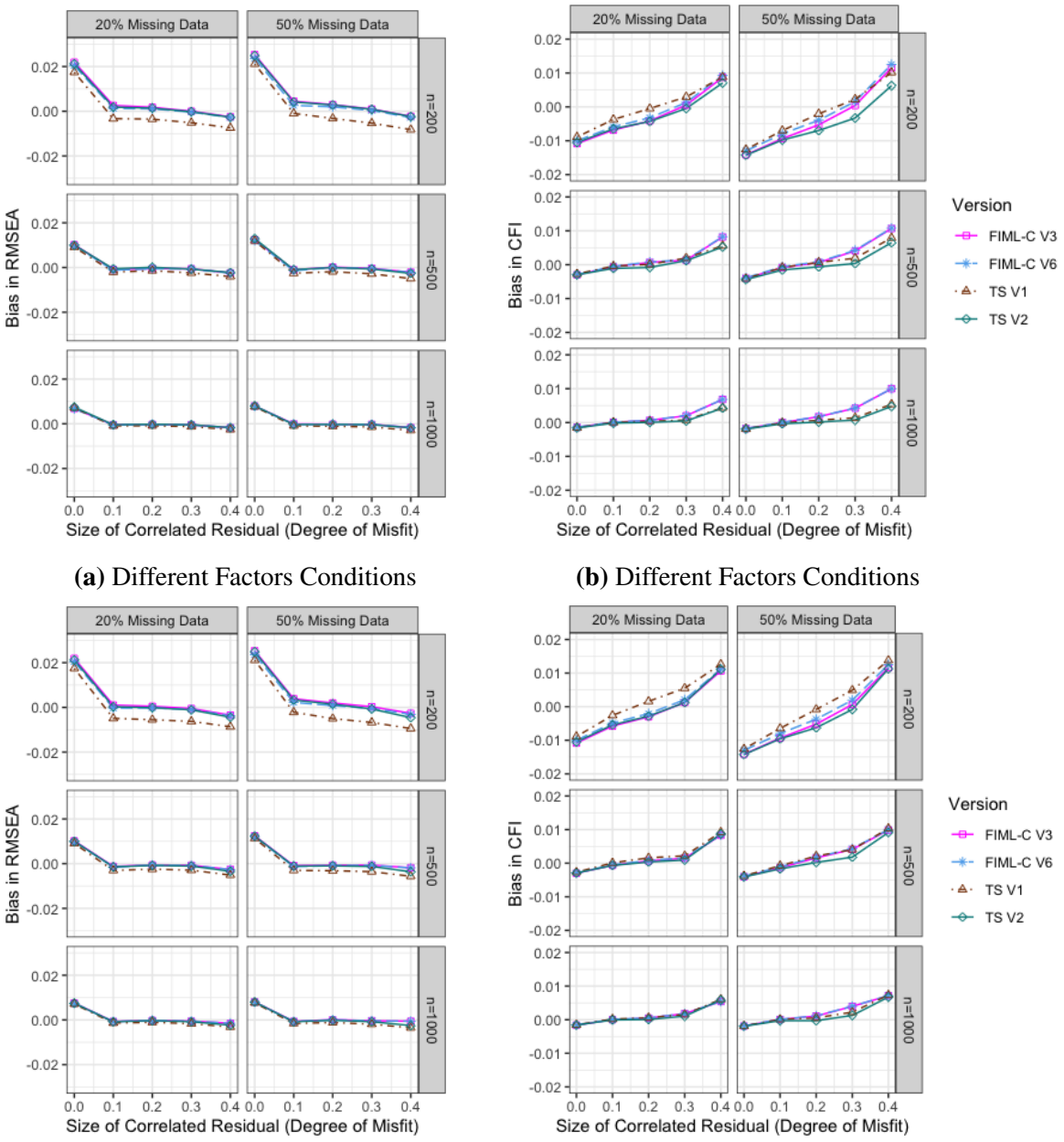
bias values for the top four performing AFIs in selected conditions whereas Figures 6.7 and 6.8 show the RMSE values in the same conditions. The *y*-axis range in these figures is smaller to allow for better discrimination among the AFI estimates.

Regression Analyses for Sample AFIs

For the regression analyses, we picked the best performing FIML-C and TS version (i.e., FIML-C V3 and TS V2) and compared them with FIML in terms of the absolute bias and the RMSE values. The predictors are the same as those for the population AFIs (see Section 6.2.1) except sample size was added as another predictor. Table 6.4 shows the results of the regression analyses. The results indicated that relative to FIML, AFIs under FIML-C V3 and TS V2, on average, had lower bias and RMSE; the amount of decrease in bias and RMSE for FIML-C V3 and TS V2 were very similar with TS V2 having slightly more decrease in bias and RMSE than FIML-C V3. For example, in Study 2, holding other variables constant, the RMSEA bias (RMSEA RMSE) for FIML-C V3 and TS V2 decreased by 0.025 (0.010) and 0.026 (0.011), respectively; the CFI bias (CFI RMSE) for FIML-C V3 and TS V2 decreased by 0.043 (0.012) and 0.050 (0.018), respectively. In addition, holding other variables constant, sample size and percentage of missing data had effects on the RMSE of AFIs but their effects were in the opposite directions. For example, in Study 2, as sample size increased from $n = 200$ to $n = 500$, the RMSE of RMSEA and CFI decreased by 0.007 and 0.016, respectively; on the other hand, as the percentage of missing data increased from 20% to 50% , the RMSE of RMSEA and CFI increased by 0.008 and 0.018, respectively. Overall, the patterns of results in the regression analyses were very consistent with those shown in the graph.

Summary for Sample AFIs

Overall, Figures 6.4-6.8 as well as the regression analyses showed that the top best-performing AFI estimation methods had similar bias and RMSE values across most con-



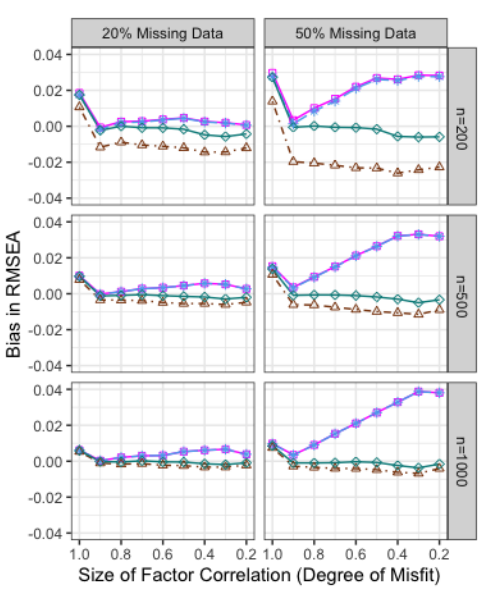
(a) Different Factors Conditions

(b) Different Factors Conditions

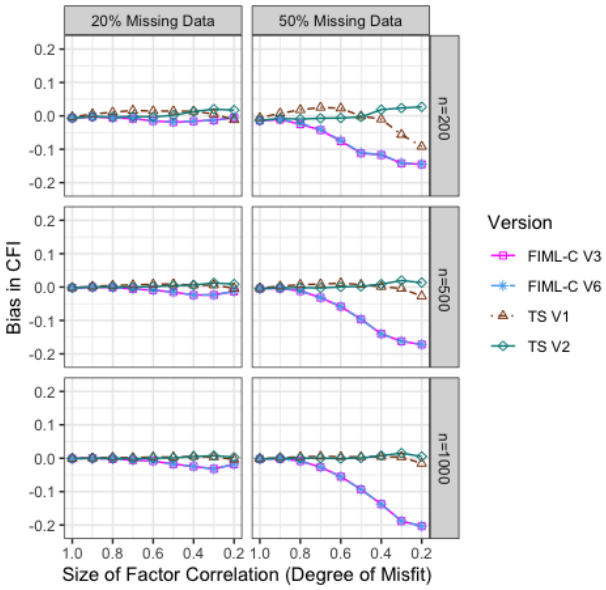
(c) Same Factor Conditions

(d) Same Factor Conditions

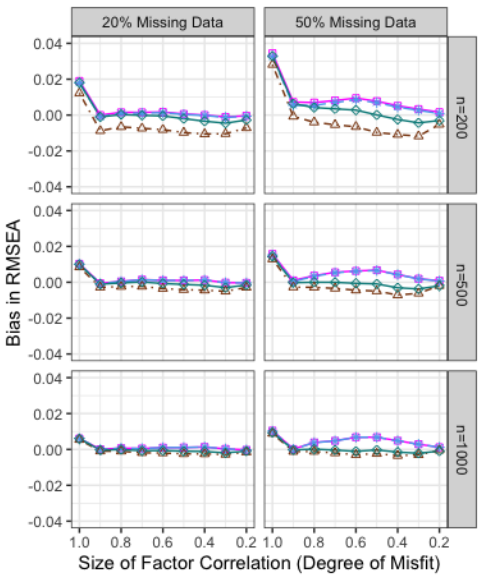
Figure 6.5: Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing among the best performing FIML-C and TS methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals.



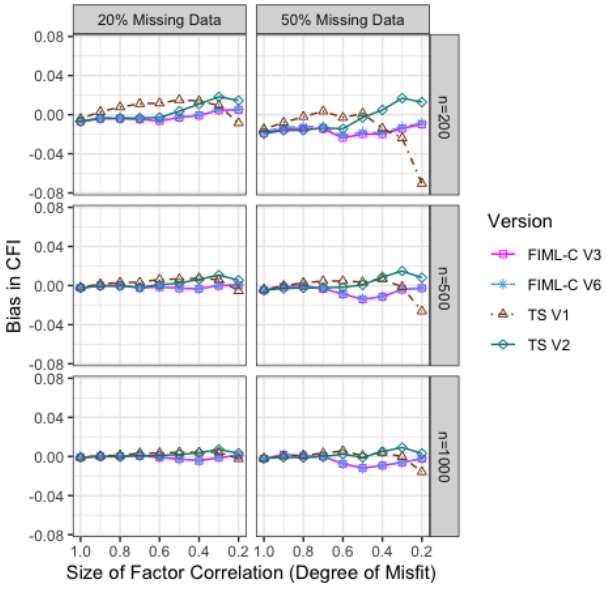
(a) Small Number of Patterns



(b) Small Number of Patterns

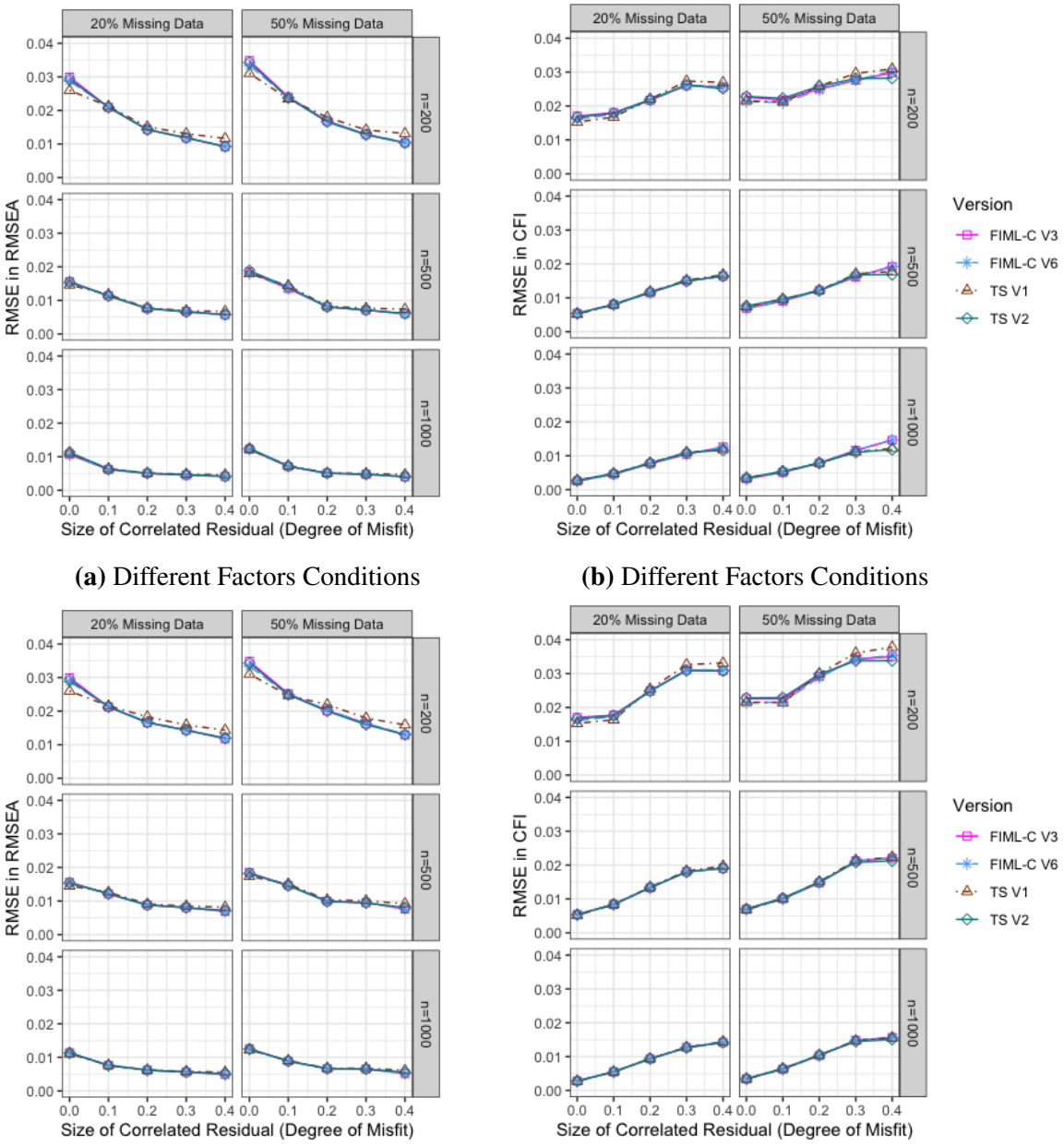


(c) Large Number of Patterns



(d) Large Number of Patterns

Figure 6.6: Bias in the sample RMSEA and CFI estimates for selected conditions in Study 2 comparing among the best performing FIML-C and TS methods. In these conditions, the number of variables with missing data is six and the missing mechanism is strong MAR. The population model is a two-factor model with varying sizes for the factor correlation shown on the *x*-axis. The hypothesized model is a one-factor model.



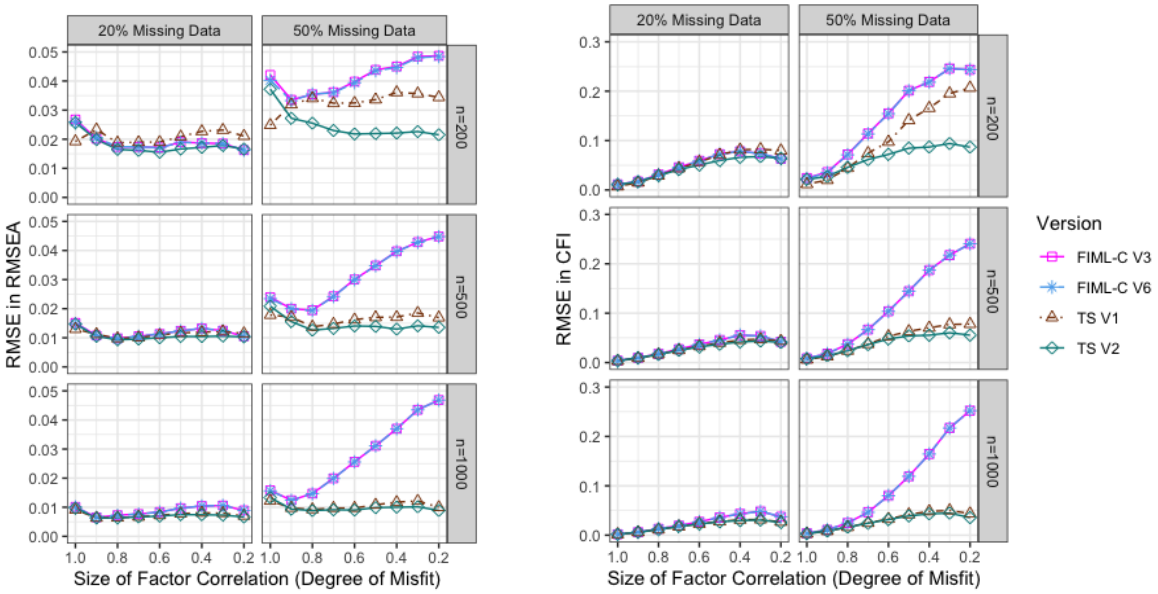
(a) Different Factors Conditions

(b) Different Factors Conditions

(c) Same Factor Conditions

(d) Same Factor Conditions

Figure 6.7: Root mean square error (RMSE) in sample RMSEA and CFI for selected conditions in Study 1 comparing among the best performing FIML-C and TS methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the *x*-axis. The hypothesized model is a two-factor model without any correlated residuals.



(a) Small Number of Patterns

(b) Small Number of Patterns

(c) Large Number of Patterns

(d) Large Number of Patterns

Figure 6.8: Root mean square error (RMSE) in sample RMSEA and CFI for selected conditions in Study 2 comparing among the best performing FIML-C and TS methods. In these conditions, the number of variables with missing data is six and the missing mechanism is strong MAR. The population model is a two-factor model with varying sizes for the factor correlation shown on the *x*-axis. The hypothesized model is a one-factor model.

Table 6.4: Results of the Regression Analyses for Bias in the Finite Samples

Study 1

$$\text{BIAS}_{\text{RMSEA}} = 0.017 - 0.015\text{EstFIML-C-V3} - 0.015\text{EstTS-V2} + 0.005\text{Missing} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.001\text{Misfit} - 0.002\text{Sample}$$

$$\text{BIAS}_{\text{CFI}} = 0.009 - 0.013\text{EstFIML-C-V3} - 0.013\text{EstTS-V2} + 0.006\text{Missing} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.004\text{Misfit} - 0.003\text{Sample}$$

$$\text{RMSE}_{\text{RMSEA}} = 0.021 - 0.006\text{EstFIML-C-V3} - 0.007\text{EstTS-V2} + 0.005\text{Missing} + 0.000\text{Mechanism} + 0.000\text{FactorCor} - 0.001\text{Misfit} - 0.006\text{Sample}$$

$$\text{RMSE}_{\text{CFI}} = 0.009 - 0.003\text{EstFIML-C-V3} - 0.005\text{EstTS-V2} + 0.005\text{Missing} + 0.000\text{Mechanism} + 0.000\text{FactorCor} + 0.006\text{Misfit} - 0.009\text{Sample}$$

Study 2

$$\text{BIAS}_{\text{RMSEA}} = 0.020 - 0.025\text{EstFIML-C-V3} - 0.026\text{EstTS-V2} + 0.009\text{Missing} + 0.001\text{Pattern} + 0.001\text{Mechanism} + 0.002\text{Misfit} - 0.002\text{samp}$$

$$\text{BIAS}_{\text{CFI}} = 0.016 - 0.043\text{EstFIML-C-V3} - 0.050\text{EstTS-V2} + 0.021\text{Missing} + 0.000\text{Pattern} + 0.004\text{Mechanism} + 0.007\text{Misfit} - 0.003\text{Sample}$$

$$\text{RMSE}_{\text{RMSEA}} = -0.019 - 0.010\text{EstFIML-C-V3} - 0.011\text{EstTS-V2} + 0.008\text{Missing} + 0.001\text{Pattern} + 0.001\text{Mechanism} + 0.001\text{Misfit} - 0.007\text{Sample}$$

$$\text{RMSE}_{\text{CFI}} = 0.006 - 0.012\text{EstFIML-C-V3} - 0.018\text{EstTS-V2} + 0.018\text{Missing} - 0.001\text{Pattern} + 0.004\text{Mechanism} + 0.009\text{Misfit} - 0.016\text{Sample}$$

Note: The coding of the variables is explained in Tables 4.3 and 6.2.

ditions. The two FIML-C AFIs (V3 and V6) produced almost identical bias and RMSE values across all conditions. However, FIML-C V3 and V6 performed considerably worse than TS V1 and V2 in some conditions. The most noticeable conditions for this pattern of results are the Study 2 conditions with a small number of patterns and 50% strong MAR missing data (see Figures 6.6ab and 6.8ab). Recall that in these conditions, at the population level, the FIML-C AFIs had large bias values, especially when the complete data AFIs had a high degree of misfit (see the bottom left panels in Figure 6.2cd); therefore, it is not surprising that at the sample level, the FIML-C AFIs also had large bias (see the right panels in Figure 6.6ab). Finally, TS V2, which is based on the saturated model estimates (see Table 5.2), tended to outperform TS V1 in finite samples. For example, in Study 2 conditions with a small sample size ($n = 200$; see Figures 6.6 and 6.8), the TS V1

estimates of AFIs tended to have larger bias and RMSE values than the TS V2 estimates of AFIs; in particular, the bias differences between TS V1 and V2 can be as large as 0.08 and the RMSE differences can be as large as 0.12.

6.3 Discussion

In these two simulation studies, we have compared the FIML, FIML-C and TS approaches for computing AFIs. Recall that the FIML-C AFIs are an approximation where the complete data fit function, with saturated FIML estimates as input ”data”, is evaluated at the FIML parameter estimates. This fit function value is then used in the AFI equations as if it were the minimum. On the other hand, the TS AFIs are based on the actual complete data fit function minimum, with saturated FIML estimates as input. The advantage of TS AFIs is that they have population values that are the same had there been no missing data, whereas the FIML-C AFIs approach these values only approximately, and can break down when misspecifications are large. Thus, the TS AFIs are theoretically superior. However, FIML is by far the most common estimation method for incomplete data, and having AFIs that are based on FIML estimates and work well is practically important.

At the population level, we found that in most conditions, both FIML-C and TS approaches performed very well, producing population AFIs with little or no bias relative to the complete data population AFIs. However, in conditions with a large percentage of strong MAR data and a small number of missing patterns, when the complete data AFIs showed a high degree of model misfit (i.e., RMSEA was greater than 0.15 and CFI less than 0.75), the FIML-C approach produced very biased AFIs whereas the TS approach still produced AFIs with no bias, as would be theoretically expected.

At the sample level, we can estimate FIML-C and TS AFIs with or without small sample corrections. Overall, we found that the FIML-C and TS AFIs without such corrections tended to have large bias in the direction of indicating worse fit when the sample size was

small (i.e., $n = 200$) and the degree of model misfit was low (i.e., complete data RMSEA lower than 0.08 and CFI higher than 0.95). We evaluated several different computational versions of the small sample corrections. For the FIML-C approach, we examined six different corrections; for the TS approach, we examined two different corrections. Four corrected versions of the FIML-C AFIs (V1, V2, V4 and V5) and one version of the TS AFIs (V1) evaluated the relevant matrices (such as the normal-theory weight matrix and the residual weight matrix, see Tables 1 and 2) at structured (model-implied) estimates of μ and Σ, whereas the remaining versions (FIML-C V3 and V6, and TS V2) evaluated these matrices at the saturated model estimates. We found that versions using the saturated model estimates (FIML-C V3 and V6, and TS V2) greatly outperformed the versions using the structured model estimates. A possible explanation for this pattern of results is that, when the model is misspecified, structured model estimates are quite off and may result in negative definitive estimates of the relevant matrices. In some replications the negative definite matrices can result in negative correction terms, creating bias in the resulting corrected AFIs. For example, based on our additional analyses, across replications, the baseline weight matrix $W_{m,B}$, when evaluated at the structured model estimates (i.e., $\hat{W}_{m,B}$), was negative definite in most conditions. As a result, the baseline correction term k_B under FIML-C V1, V2, V4 and V5 was negative in some replications. Especially for FIML-C V5, approximately 50% of k_B values across replications were negative in conditions with a small sample size (i.e., $n = 200$).

It is interesting to compare FIML-C (with best performing corrections, V3 or V6) and TS (with best performing corrections, V2). While these methods performed very similarly in most conditions, there were a small number of conditions where TS AFIs outperformed FIML-C AFIs. The reason why TS sometimes outperforms FIML-C is that FIML-C is an approximate method, and this approximation fails in a small number of conditions where the percentage of missing data is large (i.e., at least 50% missing data) and the degree of

model misfit is high (i.e., when the complete data population RMSEA was greater than 0.15 and CFI less than 0.75). However, in such conditions, model fit is already very bad, and arguably it is less important that the FIML-C AFIs are approximately unbiased, as long as they also reflect very poor model fit, so that researchers' conclusions about the model remain unchanged.

Chapter 7

Conclusion and Overall Discussion

> All generalizations are false, including this one.
>
> ---
>
> *Mark Twain*

Older missing data techniques such listwise delete and mean substitution mainly aim to get past the missing data so that at least some analyses could be done. This is in sharp contrast with modern missing data techniques, which goal is to effectively deal with missing data so that the results of data analyses are generally not affected by the missing data. In other words,'when researchers use modern techniques to handle missing data, they usually expect that their statistical analysis can estimate what the results would had been had there been no missing data. However, the first part of our dissertation work shows that this may not be case under certain circumstances.

In the first part of our dissertation (i.e., Chapters 2 and 4), we showed that using one of the most popular modern missing data techniques, the FIML estimation (which is usually the default method in current software for handling missing data), SEM AFIs such as the RMSEA and the CFI computed from incomplete data are often different from those for complete data. This discrepancy is not due to sampling fluctuations; that is, it occurs at the population level. We have provided a mathematical explanation for this phenomenon.

Specifically, we have shown that, as with complete data, maximizing the log-likelihood with incomplete data is equivalent to minimizing a certain "incomplete data ML fit function", but this function turns out to be different from the complete data ML fit function. Because the RMSEA and CFI rely on the fit function minimum in these equations, they approach different population values when data are complete versus incomplete. Furthermore, the incomplete data fit function, and hence the RMSEA and CFI values, differ with characteristics of missing data, such as missing data percentages, missing data patterns, and the exact missing data mechanism.

In addition to deriving the population fit function minima for different types of missing data, we have provided several small analytical examples and conducted two large sample simulation studies to illustrate how AFIs change with more missing data. From the analytical examples and simulation studies, we found that in addition to missing data percentage and mechanism, another characteristic of missing data that can largely affect AFIs is the the location of missing data relative to the location of misfit (i.e., whether the variables with missing data are those that are associated with model misspecifications). The general pattern of results was the following: when the location of misfit is relatively separate from the location of missing data, the fit function minimum does not change with more missing data, whereas when the location of misfit largely overlaps with the location of missing data, the fit function minimum decreases with missing data. Because RMSEA is a direct function of the fit function minimum for the hypothesized model, it stays the same when the location of misfit is separate from the location of missing data, and decreases (indicate better fit) when the two locations overlap. For CFI, how the location of missing data affects its value depends on the change in the fit function minimum for the hypothesized model versus that for the baseline model. When locations of misfit and missing data are separate, the fit function minimum for the baseline model decreases while the fit function minimum for the hypothesized model stays the same, leading to an overall decrease in

CFI (i.e., indicate worse fit). On the other hand, when the two locations overlap, both the fit function minima for the hypothesized and baseline models tend to decrease but the fit function minimum for the hypothesized model tends to decrease at a faster rate than that for the baseline model, leading to an overall increase in CFI (i.e., indicate better fit).

In short, the finding of the first part of the dissertation work is somewhat troublesome because it means that researchers using FIML AFIs as currently computed by popular software would tend to find better fit to the extent that there is missing data. However, most researchers expect that the FIML AFIs to estimate the same model fit (or misfit) as if there were no missing data, and then continue using the same cut-off guidelines developed for complete data AFIs.

To address the problem of how missing data affect AFIs under FIML, in the second part of the dissertation (i.e., Chapters 5 and 6), we have proposed and examined new estimates of AFIs following the FIML and TS estimations with incomplete data. The new estimates following FIML is called the FIML-C estimates. Theoretically, the FIML-C and TS approaches for computing AFIs allow researchers to estimate what the AFIs would have been had there been no missing data, thus placing the incomplete data AFIs' estimates on the same metric as complete data AFIs' estimates. For each of these new approaches, we have proposed several versions for calculating sample AFIs, one with no small sample correction and others with different ways of computing small sample corrections.

For the second part of the dissertation work, we conducted simulation studies to comparing the original FIML approach with different versions of the FIML-C and TS approaches. We found that at the population level, FIML-C and TS approaches, in most conditions, produced AFIs that are almost the same as complete data AFIs. However, in a small number of conditions with a high degree of misfit and a large percentage of missing data, the FIML-C approach produced AFIs that are very different from those for the complete data, whereas TS AFIs in the same conditions are the same as the complete

data AFIs. At the sample level, we found that relative to the complete data AFIs, the FIML-C and TS AFIs without small sample corrections tend to produce large bias in the direction of indicating worse fit when the sample size is small and the degree of misfit is low. For the FIML-C and TS AFIs with small sample corrections, the best performing versions (i.e., FIML-C V3 and V6, and TS V2) are those which relevant weight matrices are evaluated at the saturated model estimates rather than those evaluated at the structure estimates. Among these best performing versions, the TS approach outperforms the FIML-C approach in a small number of cases with a large percentage of missing data and high degree of misfit. The reason for TS' better performance over FIML-C is that FIML-C is an approximation method. FIML-C can only adjust for the differences in the fit function equations between incomplete and complete data; however, in the case where the model-implied means and covariances are very different between incomplete and complete data, FIML-C may fail to correct for the FIML AFIs. In summary, the finding for the second part of the dissertation work shows that the FIML-C and TS approaches can accurately estimate complete data AFIs. However, exceptions occur when there are a large amount of missing data and the hypothesized model is severely misspecified; in these cases, the FIML-C AFIs are no longer consistent estimates for the complete data population AFIs, and may deviate greatly from the complete data or TS AFIs.

Based on our study results, we recommend substantive researchers use FIML-C V3 and TS V2 AFIs when estimating model fit for incomplete data. Both FIML-C V3 and TS V2 showed very good performances in our simulation studies, and they theoretically work for both normal and non-normal data although their performances under non-normal data need to be examined in future research (see the following section). Whether researchers should use FIML-C V3 or TS V2 for estimating AFIs depends on whether they use FIML or TS to estimate the model parameters and standard errors. We recommend researchers use FIML-C AFIs if they use FIML to obtain parameter estimates, and use TS AFIs if

they use TS for parameter estimates. In other words, despite the superior performance of the TS AFIs, we would not recommend that researchers use TS AFIs just to evaluate model fit while using FIML to obtain parameter estimates and standard errors. The choice of estimation method should come first, and all computations, including AFIs, should be based on the same parameter estimates.

The next natural question that substantive researchers may ask is whether they should use FIML or TS for estimating parameters and standard errors. Under the correct model, both FIML and TS produce consistent estimates of the model parameters and standard error; in terms of efficiency, FIML tends to have higher efficiency than TS but the differences are small [35]. Therefore, under the correct model, FIML and TS methods will produce similar results. However, under a misspecified model, FIML and TS will estimate different "pseudo-parameters" in the population, with the TS "pseudo-parameters" correspond to the complete data "pseudo-parameters." In this case, there is no clear answer as to which "pseudo-parameters" are better. If the researchers are interested in the complete data "pseudo-parameters", then we would recommend the TS approach. On the other hand, if the researchers are interested in "pseudo-parameters" that are closest to the distribution of the population incomplete data, then we would recommend the FIML approach.

7.1 Limitations and Future Directions

As all research works, the current dissertation work is not without limitations. First, our stimulation studies only examined CFA models. Future studies should examine the effect of missing data on the AFIs using other SEM models. Second, the implementation of the FIML-C and TS methods require the use of internal functions in the *lavaan* package in *R* (see Supplementary Materials for sample code). The use of internal functions may make it harder for applied researchers to understand and implement the code. Therefore, in the

future, we hope to work with the developer of the *lavaan* package so that the FIML-C and TS methods can be added to the functions of the package. Third, in our simulation studies, we only examined normally distributed data; this may explain why the FIML-C versions that make the normality assumption performed similarly with the versions without the assumption. Future studies should include non-normal data conditions (e.g., data with large skewness and kurtosis) in order to confirm whether FIML-C version without the normality assumption can outperform the version with the assumption in the non-normal data conditions. Fourth, as shown in our simulation studies (see Table 6.1), with misspecified models, the "pseudo-parameters" under FIML and TS can be very different; such patterns of results were rarely discussed in previous research comparing FIML with TS because previous studies [e.g., 35, 47, 48] mainly focused on comparing the parameter estimates under FIML and TS for the correctly specified model. In a future study, we can conduct a simulation study that systematically examines the differences between "pseudo-parameters" under FIML and TS methods.

Another future direction for this line of research is to examine the MI approach for estimating AFIs. As mentioned in Section 1.3, MI is another popular missing data technique. In MI, we first create multiple "complete" datasets by using imputation under either the saturated or the structured model, fit the structured model to each dataset, and then pool the AFIs across these "complete" datasets. [1] When imputation is done under the saturated normal model, this MI method is conceptually equivalent to TS and should also produce AFIs that estimate complete data AFIs. It is interesting to note that FIML and MI are two most common methods for treating incomplete data in SEM, yet they produce AFIs with different population values, a problem that has not so far been noticed in the literature. We have done some preliminary simulation studies to confirm our expectation. Figure 7.1

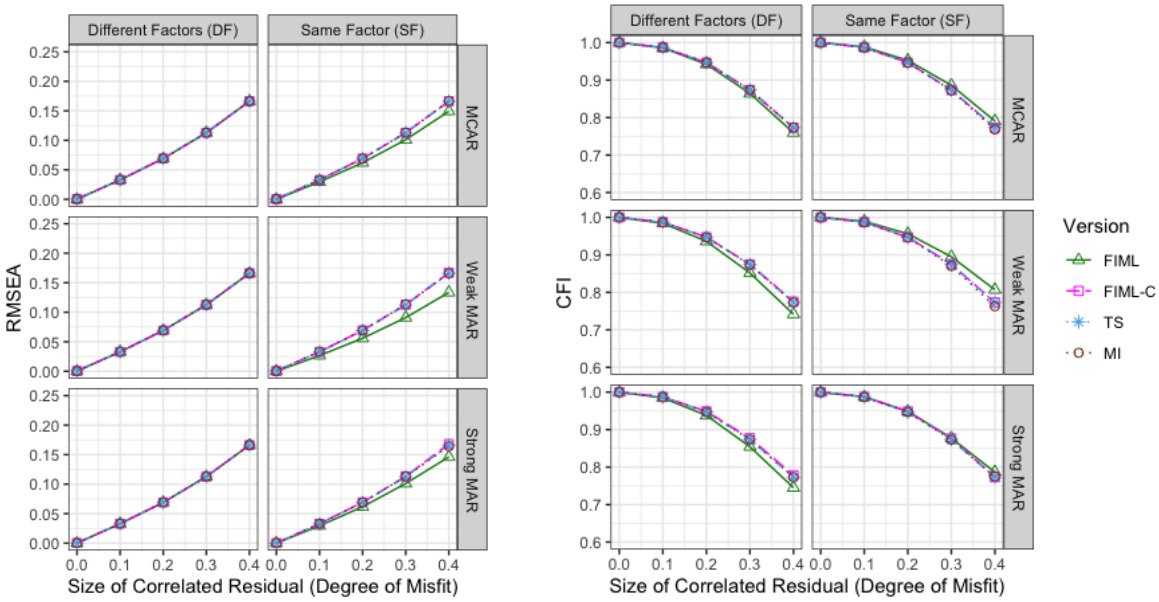
[1] The pooling stage of MI can be done in different ways. One way is to compute AFIs in each "complete" dataset, and then average across the computed AFIs. Another way is to compute the fit function minimum in each "complete" dataset, average across the fit function minima, and then computed the AFIs based on the pooled fit function minima. These different pooling methods also need to be studied in future research.

shows the population AFIs under the MI method. Consistent with our expectation, at the population level, the AFIs under MI are exactly the same as those under TS, which are the same as those under complete data (see Figure 6.1).

At the sample level, similar to the FIML-C and TS methods, small sample corrections may be needed to improve upon the MI method for estimating AFIs. Without a small sample correction, the MI AFIs should perform similarly as the FIML-C and TS AFIs without small sample corrections. Our preliminary simulation results show that for RMSEA, the MI without small sample correction indeed performed similarly as the FIML-C and TS versions without small sample corrections but for CFI, the results for MI were quite different from FIML-C and TS (see Figure 7.2). In our future work, we will attempt to find an explanation for this phenomenon.

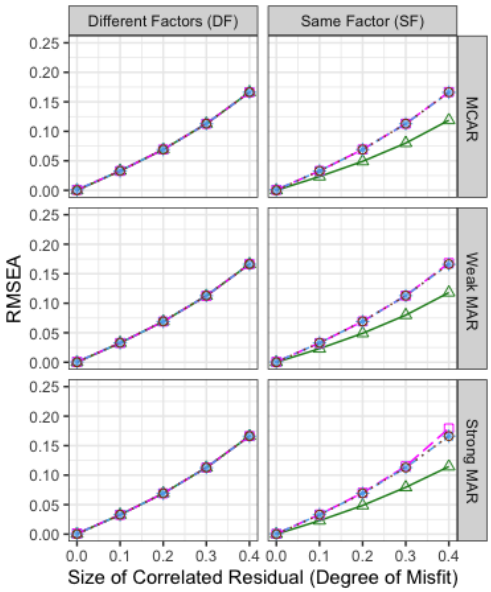
In addition, Enders and Mansolf [13] have attempted to develop a small sample correction for computing AFIs under MI. However, the main purpose of their work was to come up with a correction for the MI chi-square test statistic to account for missing data. They then used this corrected chi-square statistic in the equations for sample AFIs, but doing so distorted their population values (see Brosseau-Liard and Savalei [6], Brosseau-Liard et al. [7] for explanations). In our future research, we plan to develop appropriate small sample corrections for MI AFIs that do not distort their population values.

Finally, our finding that small sample corrections computed with weight matrices evaluated at the saturated model estimates were better than those computed at the structured model estimates has further implications. With the current SEM literature, there are very few research studies that investigate how computational options for estimating the weight matrices affect the estimates of other quantities, such as the estimates of the small sample correction for non-normal data and the estimate of missing information. In fact, many SEM users may not be aware of these computational options. In future studies, researchers should systematically examine these computational options.

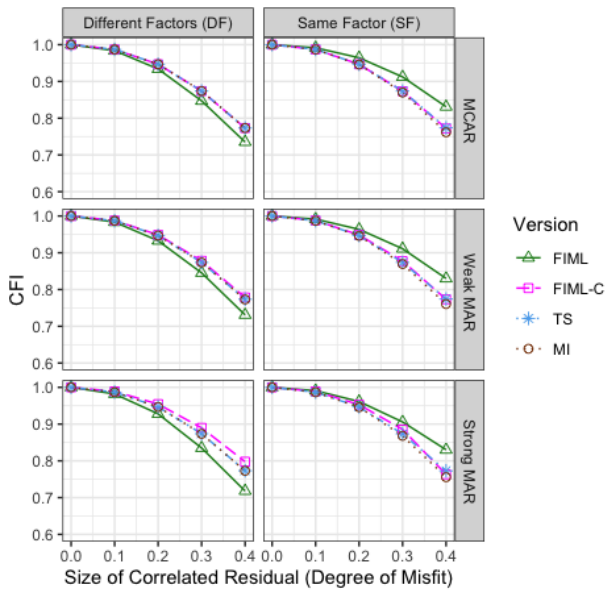


(a) 20% Missing Data

(b) 20% Missing Data

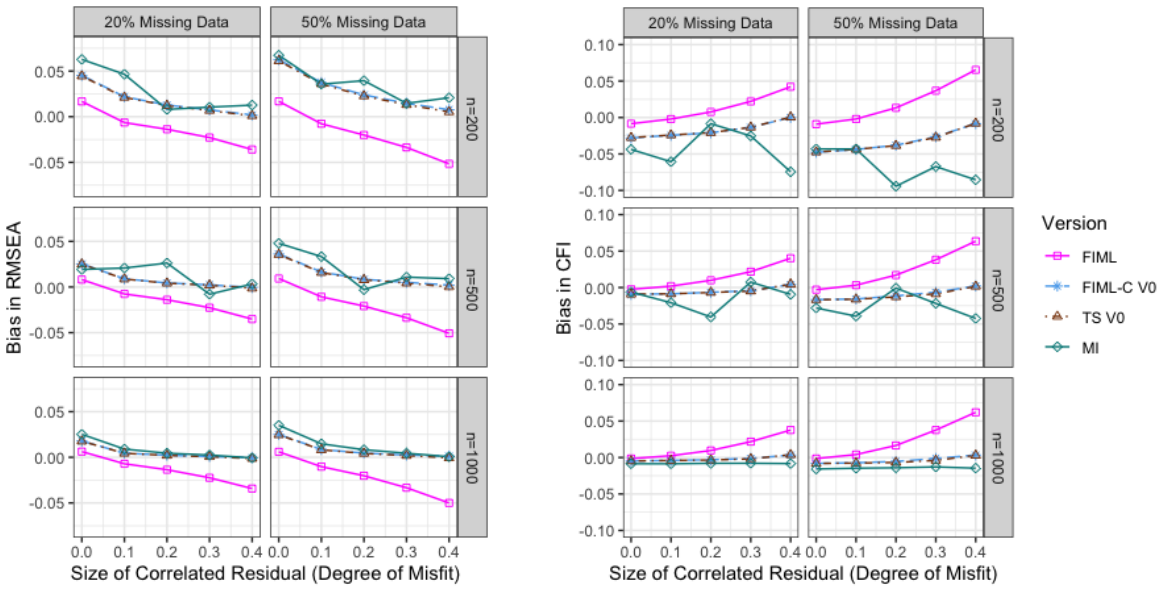


(c) 50% Missing Data



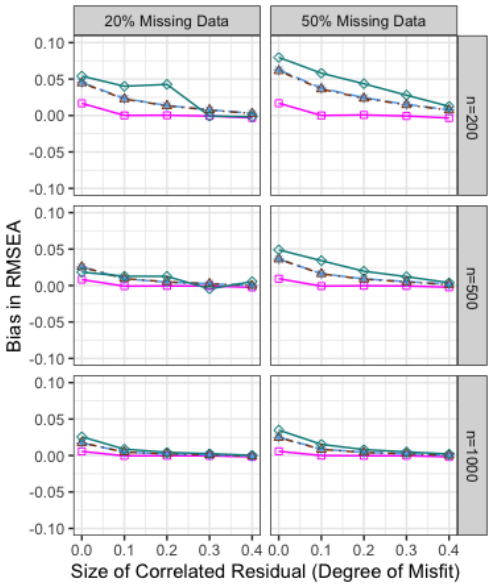
(d) 50% Missing Data

Figure 7.1: Population RMSEA and CFI (estimated from $n = 1000000$) for selected conditions in Study 1 comparing FIML, FIML-C, TS, MI approaches. In these selected conditions, the number of variables with missing data is four, the number of two correlated residuals is two, and the population factor correlation is zero. The population model is a two-factor model with varying sizes for the correlated residuals shown on the x-axis. The hypothesized model is a two-factor model without any correlated residuals.

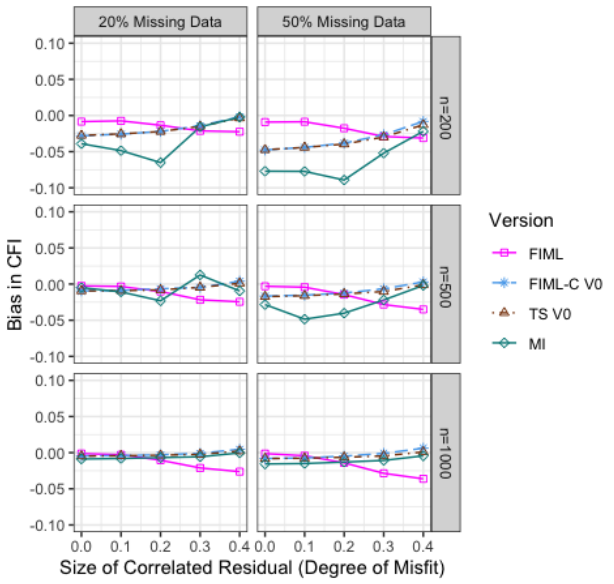


(a) 20% Missing Data

(b) 20% Missing Data



(c) 50% Missing Data



(d) 50% Missing Data

Figure 7.2: Bias in the sample RMSEA and CFI estimates for selected conditions in Study 1 comparing among FIML, FIML-C, TS and MI methods. In these conditions,the number of variables with missing data is four, the number of two correlated residuals is two, the population factor correlation is zero, and the missing data mechanism is weak MAR. The population model is a two-factor model with varying sizes for the correlated residuals shown on the *x*-axis. The hypothesized model is a two-factor model without any correlated residuals.

In conclusion, this dissertation makes a meaningful contribution to our understanding of how different missing data techniques affect the estimation of AFIs in SEM. Based on the results of this dissertation, we have offered practical advice to applied researchers who use SEM in their data analyses. Future research should continue investigating properties of different missing data techniques and exploring better methods for handling missing data in a variety of settings.

Bibliography

[1] Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112:545–557. 8

[2] Arbuckle (1999). *Full information estimation in the presence of incomplete data*. Lawrence Erlbaum, Mahwah, NJ. 8

[3] Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107:238–46. 12

[4] Bentler, P. M. and Wu, E. J. (1995). *EQS for Windows user's guide:[version 5]*. Multivariate Software, Encino, CA. 12

[5] Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, New York. 10, 23

[6] Brosseau-Liard, P. E. and Savalei, V. (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research*, 49(5):460–470. 68, 69, 76, 110

[7] Brosseau-Liard, P. E., Savalei, V., and Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, 47(6):904–930. 68, 69, 76, 110

[8] Browne, M. W. and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21:230–258. 12, 51, 65

[9] Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge. 19

[10] Davey, A., Salva, J., and Luo, Z. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling*, 12:578–597. 13, 17

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39:1–38. 21

[12] Enders, C. K. and Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1):35–54. 10

[13] Enders, C. K. and Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, 23:76–93. 14, 17, 110

[14] Enders, C. K. and 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*, 11:1–19. 21, 75

[15] Graham, J. W. (2010). *Missing data: Analysis and design*. Springer, New York. 6, 9

[16] Grigsby, T. J. and McLawhorn, J. (2019). Missing data techniques and the statistical conclusion validity of survey-based alcohol and drug use research studies: A review and comment on reproducibility. *Journal of Drug Issues*, 49(1):44–56. 6

[17] Hu, L. T. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternative. *Structural Equation Modeling*, 6:1–55. 12, 51, 65

[18] Jelicic, H., Phelps, E., and Lerner, R. (2009). Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Developmental Psychology*, 45:1195–1199. 1

[19] Joe, H. (2014). *Dependence modeling with copulas*. CRC press, Boca Raton. 20

[20] Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67:609–24. 27

[21] Lai, K. (2020). Correct estimation methods for RMSEA under missing data. *Structural Equation Modeling*, Advanced publication. 14

[22] Li, J. and Lomax, R. G. (2017). Effects of missing data methods in SEM under conditions of incomplete and nonnormal data. *The Journal of Experimental Education*, 85:231–58. 14, 17

[23] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 7

[24] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, New York. 20

[25] Maiti, S. S. and Mukherjee, B. N. (1990). A note on distributional properties of the jöreskog-sörbom fit indices. *Psychometrika*, 55(4):721–726. 12

[26] Muthén, B. O., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3):431–462. 23

[27] Muthén, B. O. and Muthén, L. K. (2010). Technical appendices. *Los Angeles, CA: Muthén & Muthén*. 23

[28] Nakagawa, S. (2015). Missing data: Mechanisms, methods, and messages. In Fox, G. A., Negrete-Yankelevich, S., and Sosa, V. J., editors, *Ecological statistics:Contemporary theory and application*, pages 81–105. Oxford Scholarship Online. 2

[29] Osborne, J. W. (2013). Is data cleaning and the testing of assumptions relevant in the 21st century? *Frontiers in Psychology*, 4:1–3. 1

[30] Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48:1–36. 46, 84

[31] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, New York. 2, 4, 8

[32] Rubin, J. B. (1976). Inference and missing data. *Biometrika*, 63:581–592. 2

[33] Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological methods*, 15(4):352. 70

[34] Savalei, V. (2020). Improving fit indices in structural equation modeling with categorical data. *Multivariate Behavioral Research*, pages 1–18. 68, 69

[35] Savalei, V. and Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling*, 12:183–214. 7, 10, 108, 109

[36] Savalei, V. and Bentler, P. M. (2009). A two-stage approach to missing data: theory and application to auxiliary variables. *Structural Equation Modeling*, 16:477–497. 6, 7, 9, 10, 75

[37] Savalei, V. and Rhemtulla, M. (2017). Normal theory two-stage ML estimator when data are missing at the item level. *Journal of Educational and Behavioral Statistics*, 42(4):405–431. 6, 10

[38] Shapiro, A. (2007). Statistical inference of moment structures. In Lee, S.-Y. L., editor, *Handbook of latent variable and related models*, pages 229–260. Elsevier, Amsterdam, Netherlands. 81

[39] Steiger, J. H. (1980). *Statistically based tests for the number of factors*. Paper presented at the annual meeting of Psychometric Society, Iowa City, IA. 12

[40] Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333. 19

[41] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25. 5

[42] Xia, Y., Yung, Y.-F., and Zhang, W. (2016). Evaluating the selection of normal-theory weight matrices in the satorra–bentler correction of chi-square and standard errors. *Structural Equation Modeling*, 23(4):585–594. 70

[43] Yuan, K. H. and Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30:165–200. 8, 9, 68, 75, 81

[44] Yuan, K. H., Jamshidian, M., and Kano, Y. (2018). Missing data mechanisms and homogeneity of means and variances-covariances. *Psychometika*, 83:425–42. 27

[45] Yuan, K. H. and Lu, L. (2008a). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, 43:621–52. 6

[46] Yuan, K.-H. and Lu, L. (2008b). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, 43(4):621–652. 9, 68

[47] Yuan, K.-H., Tong, X., and Zhang, Z. (2015). Bias and efficiency for SEM with missing data and auxiliary variables: Two-stage robust method versus two-stage ML. *Structural Equation Modeling*, 22(2):178–192. 109

[48] Yuan, K.-H. and Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, 77(4):803–826. 9, 10, 109