Title
Alternative Multiple Imputation Inference for Categorical Structural Equation Modeling

Author
Chung, Seungwon

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Alternative Multiple Imputation Inference for Categorical Structural Equation Modeling

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Seung Won Chung

2019

ABSTRACT OF THE THESIS

Alternative Multiple Imputation Inference for Categorical Structural Equation Modeling

by

Seung Won Chung
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Ying Nian Wu, Chair

The use of responses from questionnaires is ubiquitous in social and behavioral science research. One side effect of using such data is that researchers must often account for item level missingness. Multiple imputation is one of the most widely used missing data handling techniques, wherein missing data are replaced by plausible values from the their proper posterior distribution given the observed data. Instead of the standard procedure in structural equation modeling (SEM), which requires researchers to fit their model to imputed data sets as many times as the number of imputations and then combine parameter estimates and standard errors at the end, we propose a new and simpler approach that is computationally more convenient. It has a number of additional benefits such as the availability of fit indices. Motivated by Lee and Cai (2012), who proposed an alternative method for statistical inference under MI in SEM with continuous variables, we extend their approach to the case of categorical variables.

Within the context of ordered categorical data, the main idea is summarized as follows. Assume we have thresholds and polychoric correlations computed from M imputed data set. Our goal is to perform estimation and inference with these M different thresholds and polychoric correlations. We can easily average the

thresholds and polychoric correlations; however, the weight matrix for obtaining the correct statistic in CSEM requires reflecting the between-imputation variance on top of simple averaging of asymptotic covariance matrices of the thresholds and polychoric correlations. Finally, applying Browne (1984)'s Proposition 4 leads us to obtain the correct test statistic, $\tilde{T}_B$. We further consider $\tilde{T}_{YB}$, a small-sample adjustment of $\tilde{T}_B$ (Yuan & Bentler, 1997). We demonstrate our proposed statistics performance and their power to detect model misspecification via simulation studies. In addition, we illustrate our findings with two empirical data sets.

The thesis of Seung Won Chung is approved.

Li Cai

Mark Stephen Handcock

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The use of item response data is ubiquitous in social science research. Item responses, however, are rarely complete. Researchers must often account for missingness in their data. A growing body of research provides insight into the comparative performance of missing data techniques in structural equation modeling (SEM) (e.g., Allison, 2003; Arbuckle, 1996; Enders & Bandalos, 2001; Enders & Peugh, 2004; Olinsky, Chen, & Harlow, 2003; Takahashi & Wisenbaker, 1999; Wiggins & Sacker, 2002; Wang, 2007; Shin, D., & L., 2009; Li, 2010). Three methods of dealing with missing data in SEM are featured prominently in the literature: full-information maximum likelihood (FIML; Anderson, 1957; Arbuckle, 1996), multiple imputation (Schafer, 1997), and a two-stage procedure based on the Expectation-Maximization algorithm (EM2S; Allison, 2001; Cai & Lee, 2009; Enders & Peugh, 2004; Yuan & Bentler, 2000). While multiple imputation (Rubin, 1987) is one of the most widely used techniques for handling missing data, research on its use in the SEM context is surprisingly limited (e.g., Enders & Mansolf, 2018; Lee & Cai, 2012).

Unlike FIML, which generally requires the normality assumption, multiple imputation is considerably less restrictive in terms of distributional assumptions (Rubin, 1976; Little & Rubin, 1987; Schafer, 1997). Multiple imputation may be a better choice for researchers who must deal with categorical item level data, e.g., in educational testing. Furthermore, mixtures of continuous and categorical variables are encountered frequently in the practice of data analysis using SEM. Fully conditional specification (FCS), also known as multivariate imputation by chained

equations (MICE), is designed for these types of data. FCS imputes incomplete variables based on a series of conditional models, one for each incomplete variable. Accordingly, one advantage of the imputation approach is its flexibility because different distributions can be specified to model each variable (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren, 2006; Bouhlila & Sellaouti, 2013). Our approach can easily accommodate this extension. Therefore, the approach we propose in this paper will be broadly useful.

Before we introduce our alternative procedure, let us discuss the standard multiple imputation approach (see e.g., Schafer & Olsen, 1998) in SEM. In the standard approach, after multiple imputation, researchers must fit their models to all imputations and obtain final parameter estimates by averaging parameter estimates across the imputations. Standard errors are obtained by averaging and accounting for cross-imputation variability. However, this procedure has a number of limitations. First, the standard procedure of multiple imputation is computationally burdensome because model-fitting must be performed for each imputed dataset. Second, the commonly used fit statistics such as root mean square error of approximation (RMSEA; Browne & Cudeck, 1993) or Tucker-Lewis index (TLI; Tucker & Lewis, 1973) are not readily available in the standard multiple imputation approach. In a recent effort to resolve this issue, Enders and Mansolf (2018) defined commonly used SEM fit indices from Meng and Rubin (1992)'s pooling procedure for likelihood ratio statistics. We believe an even simpler procedure exists in our approach. Last but not least, the standard multiple imputation inference procedure only provides corrected point estimates and standard errors but not intermediate results such as the equivalent of the mean vector and covariance/correlation matrix, which are useful for replication and meta-analytic studies.

Motivated by Lee and Cai (2012)'s work on multiple imputation inference, who proposed a multiple imputation two stage (MI2S) estimator for continuous

and normally distributed observed variables, we extend their approach to the case of categorical variables or items. The guiding insight of the MI2S estimator is that the structural equation model is fitted after all multiple imputations have been combined as opposed to the traditional approach wherein researchers fit a structural equation model for each imputed data set and average the parameter estimate and standard errors at the end. In MI2S, researchers combine the imputations under the unrestricted multivariate normal model to obtain a single mean vector and covariance matrix (along with their asymptotic covariance matrix) that are corrected for missing data. The mean vector, covariance matrix, and their asymptotic covariance matrix become input into the second stage of business-as-usual estimation and statistical inference. Our new estimator follows the logic of the MI2S estimator and is applicable to categorical data.

The purpose of this paper is to introduce the use of the MI2S estimator for situations in which data are missing on categorical variables. We note that this paper is on the (inferential) procedure in SEM with the multiply imputed data after multiple imputation have been performed, and thus topics on imputation methods are beyond the scope of this paper. As an aside, we wish to address the relevance of the FIML estimator for categorical data.[1] It is also known as the marginal maximum likelihood (MML) estimator in the Item Response Theory (IRT) literature (e.g., Bock & Aitkin, 1981). Given the increasing availability of the FIML estimator for categorical data in software programs, it is tempting to ask why multiple imputation is still needed. We emphasize that multiple imputation is a general approach not dependent on particular formulations of the structural modeling framework. It more easily allows one to utilize the multi-stage estimation approach, which is described in the next section. The multi-stage approach itself possesses some advantages over FIML Forero et al. (2009). FIML is also compu-

[1]The multi-stage estimator is often referred to as a *limited information* method as opposed to *full information* maximum likelihood (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009) that relies on raw data.

tationally more intensive than our approach because it requires high-dimensional integration over a multivariate distribution with as many dimensions as there are observed variables. On the other hand, in the multi-stage approach, estimating the thresholds and polychoric correlations only requires one or two-dimensional integration (see Maydeu-Olivares & Joe, 2006; Wirth & Edwards, 2007; Forero et al., 2009). Furthermore, multi-stage estimation can easily incorporate auxiliary variables, while this is not true for FIML.

CHAPTER 2

Some Reviews on Missing Data and Structural Equation Model

2.1 Missing Data Mechanisms

Rubin (1976) have theorized missing data mechanisms. A missing data mechanism refers to a probability model for a missing data pattern, of which there are three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Following Schafer (1997), we provide a brief review on these three missing data mechanisms.

Let there be $i = 1, 2, ..., n$ observation units and $j = 1, 2, ..., p$ variables. Let $\mathbf{Y}$ be an $n \times p$ matrix of the complete data set. We label the incomplete observed data $\mathbf{O}$ and the missing data $\mathbf{X}$, so that $\mathbf{Y} = (\mathbf{O}, \mathbf{X})$. Let $\mathbf{R}$ denote an $n \times p$ matrix of missing data indicators such that its element

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{if } y_{ij} \text{ is missing.} \end{cases} \tag{2.1}$$

Now we formulate a probability model for $\mathbf{R}$, $f(\mathbf{R}|\mathbf{Y}, \phi)$, where ϕ is a set of unknown parameters. The data are MAR if

$$f(\mathbf{R}|\mathbf{M}, \mathbf{O}) = f(\mathbf{R}|\mathbf{O}, \phi), \tag{2.2}$$

which means the missing data mechanism does not depend on unobserved data

M. MCAR is a special case of MAR, where the missing data mechanism does not depend on observed data **O** in addition to unobserved data **M**. It can be expressed as

$$f(\mathbf{R}|\mathbf{M},\mathbf{O}) = f(\mathbf{R}|\phi). \tag{2.3}$$

If the missing data mechanism depends on unobserved data **M**, i.e.,

$$f(\mathbf{R}|\mathbf{M},\mathbf{O}) = f(\mathbf{R}|\mathbf{M},\mathbf{O},\phi), \tag{2.4}$$

then it is called NMAR. Note that MAR and MCAR are often called ignorable missing, whereas NMAR is called non-ignorable missing.[1]

Traditionally, there have been two approaches for missing data handling: 1) eliminating the cases with incomplete data and 2) filling in the missing values (Enders, 2010). However, none of the techniques is superior to maximum likelihood (ML) and multiple imputation (MI) under the MAR assumption (Allison, 2003; Schafer & Graham, 2002). In the next section, we briefly review conventional MI.

2.2 Multiple Imputation

The MI method is developed and elaborated by Rubin (1978, 1987). Rubin (1996) calls MI "the method of choice for addressing problems due to missing values." To simplify, MI refers to a missing data handling technique wherein missing data are replaced by $M > 1$ plausible values of the missing data as drawn from the posterior distribution of the missing data given the observed data.

In general, MI consists of three steps: 1) Imputation: the missing data are filled in with M values, creating M complete data sets; 2) Analysis: each com-

[1]More precisely, both MAR and *distinctness* should hold to be said ignorable (Little & Rubin, 1987; Rubin, 1987)

plete data set is analyzed using any statistical tools for complete cases; and 3) Pooling: the results from the M complete data sets are appropriately combined to form a single set of results. However, it is not readily applicable in a situation with more than one variable with missing data. Thus, the Markov chain Monte Carlo (MCMC) algorithm or data augmentation is customarily used to resolve the problem (Schafer, 1997; Allison, 2003).

The MCMC algorithm under the multivariate normal model takes the following steps: 1) a set of starting values for the means and the covariance matrix are chosen, which can be obtained from the EM algorithm; 2) the linear regression models that predict variables with missing data from variables with observed data are estimated for each missing data pattern using the current mean and covariance matrix; 3) the missing data values are predicted from the regression coefficients, and these values are augmented with random draws from the estimated distribution of residuals; 4) the means and the covariance matrix are recalculated after filling in all the data; and 5) random draws of means and covariances from the posterior distribution (a normal, inverted Wishart distribution) are made. This procedure is repeated until convergence, and the imputed values obtained in Step 3 are saved for each iteration for later analysis (Allison, 2003). Steps 1 through 3 may overall be called the imputation step (I-step) and Steps 4 through 5 the posterior step (P-step) (Enders, 2010).

An alternative to this so-called joint modeling approach is fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). It is commonly used for data sets containing categorical variables. This method similarly follows the aforementioned I-step and P-step, but with different mechanics. It imputes incomplete variables in a sequence. That is, missing values are drawn from a series of univariate distributions. It is important to note that different regression models (not necessarily linear) may be applied to match the scale of the incomplete variable (Enders, 2010; Raghunathan, Lepkowski, Hoewyk, &

Solenberger, 2001; van Buuren, 2006; van Buuren et al., 2006).

2.3 Structural Equation Model

In this section, we provide a brief overview on SEM (Bollen, 2010; Browne & Arminger, 1995; Yuan & Bentler, 2007). In general, a full model of SEM consists of two pieces: measurement model and structural model.

Let us begin with the measurement model. Confirmatory factor analysis (CFA) model is used as the measurement model in accordance with the psychometric convention. A factor analysis can be understood as a multivariate regression analysis without (observable) predictors. Specifically, latent variables, i.e., unobserved variables or unmeasured variables, are served as predictors. Let $\mathbf{y}_i$ be an $n \times 1$ vector of observed variables. Let $\boldsymbol{\eta}_i$ be a $p \times 1$ vector of common factors (latent variables). The measurement model can be written as

$$\mathbf{y}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \tag{2.5}$$

where $\boldsymbol{\tau}$ is an $n \times 1$ vector of intercept terms, $\boldsymbol{\Lambda}$ is an $n \times p$ common factor loading matrix, and $\boldsymbol{\epsilon}_i$ is an $n \times 1$ vector of error terms with zero means and covariance matrix of $\boldsymbol{\Phi}$. Note that the unique factors $\boldsymbol{\epsilon}_i$ and the common factors $\boldsymbol{\eta}_i$ are uncorrelated.

Next, the structural model is introduced to describe the relationship among latent variables. Recall $\boldsymbol{\eta}_i$ is a $p \times 1$ vector of common factors (latent variables). Let $\mathbf{x}_i$ be an $m \times 1$ vector of observed, fixed covariates. The latent structural equations is

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{2.6}$$

where $\mathbf{B}$ is a $p \times p$ matrix of (fixed) latent variable regression coefficients, $\boldsymbol{\Gamma}$ is a $p \times m$ matrix of (fixed) observed variable regression coefficients, and $\boldsymbol{\zeta}_i$ is a $p \times 1$

random vector of error terms with zero means and covariance matrix of $\mathbf{\Psi}$.

Rearranging the above equation gives us the following:

$$\boldsymbol{\eta}_i = \mathbf{A}(\mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i), \tag{2.7}$$

where $\mathbf{A}$ is an invertible matrix and equals to $(\mathbf{I}_p - \mathbf{B})^{-1}$. Here, $\mathbf{I}_p$ is a $p \times p$ identity matrix.

By combining the measurement model and the structural model, we obtain the full model equation, which is also known as the LISREL all-Y side model, as follows:

$$\mathbf{y}_i = \boldsymbol{\tau} + \mathbf{\Lambda}\mathbf{A}(\mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i) + \boldsymbol{\epsilon}_i. \tag{2.8}$$

From the full model, we can obtain the mean structure and the covariance structure of the model. Given that both $\boldsymbol{\zeta}_i$ and $\boldsymbol{\epsilon}_i$ are disturbance terms with zero means, that is, $E(\boldsymbol{\zeta}_i) = 0$ and $E(\boldsymbol{\epsilon}_i) = 0$, the mean structure is

$$\boldsymbol{\mu} = E(\mathbf{y}_i) = \boldsymbol{\tau}. \tag{2.9}$$

Given that $\boldsymbol{\zeta}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated, and $Var(\boldsymbol{\zeta}_i) = \mathbf{\Psi}$ and $Var(\epsilon_i) = \mathbf{\Phi}$, the covariance structure can be written as

$$\mathbf{\Sigma} = Var(\mathbf{y}_i) = \mathbf{\Lambda}\mathbf{A}\mathbf{\Psi}\mathbf{A}'\mathbf{\Lambda}' + \mathbf{\Phi}. \tag{2.10}$$

CHAPTER 3

Multi-Stage Estimation of Structural Equation Models with Categorical Variables

3.1 The Underlying Variables Formulation

In the SEM tradition, categorical observed variables can be viewed as the result of discretization of underlying continuous response variables. Without loss of generality, let us consider the case of n observed variables each having K ordered categories ($k = 0, 1, 2, , K-1$). Let $\mathbf{y}^* = (y_1^*, y_2^*, , y_n^*)'$ be a vector of n underlying continuous response variables. The observed categorical response $\mathbf{y} = (y_1, y_2, , y_n)'$ is formed by the discretization of $\mathbf{y}^*$ via a set of thresholds, $\boldsymbol{\tau}$. The relation between y_i and y_i^* for item i is given by

$$\begin{cases} y_i = 0, & \text{if} \quad \tau_{i,0} < y_i^* \leq \tau_{i,1} \\ y_i = 1, & \text{if} \quad \tau_{i,1} < y_i^* \leq \tau_{i,2} \\ \quad \vdots & \\ y_i = K-1, & \text{if} \quad \tau_{i,K-1} < y_i^* \leq \tau_{i,K} \end{cases} \tag{3.1}$$

where $-\infty = \tau_{i,0} < \tau_{i,1} < \tau_{i,2} ... \tau_{i,K} = \infty$. When there are K categories, there are $K-1$ well-defined thresholds.

This connection between categorical variables and continuous underlying response variables allows us to work with the underlying continuous variables $\mathbf{y}^*$ in SEM instead of the observed categorical variables $\mathbf{y}$. Let the covariance matrix of

$\mathbf{y}^*$ be denoted $\boldsymbol{\Sigma}$. One may impose a covariance structure model on $\boldsymbol{\Sigma}$ by introducing its dependence on a vector of free parameters $\boldsymbol{\theta}$. We consider a LISREL-type linear covariance structure model introduced in Chapter 2.

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'\boldsymbol{\Lambda}' + \boldsymbol{\Phi} \tag{3.2}$$

where $\boldsymbol{\Lambda}$ is an $n \times p$ common factor loading matrix, $\boldsymbol{\Psi}$ is a $p \times p$ common factor covariance matrix, and $\boldsymbol{\Phi}$ is an $n \times n$ covariance matrix of the unique factors. Matrix $\mathbf{A}$ is an invertible matrix and equals to $(\mathbf{I}_p - \mathbf{B})^{-1}$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\mathbf{B}$ is a matrix of regression coefficients describing the linear structural relationship among the common factors. Because the underlying variables can have arbitrary scaling, one method to identify the model is by setting $\boldsymbol{\Phi} = \mathbf{I}_p - \text{diag}(\boldsymbol{\Lambda}\mathbf{A}\boldsymbol{\Psi}\mathbf{A}'\boldsymbol{\Lambda}')$, such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{P}(\boldsymbol{\theta})$, where $\mathbf{P}$ has unit diagonals (a correlation matrix). Estimating the polychoric correlations among the observed variables is a critical aspect of categorical structural equation modeling.

3.2 Thresholds and Polychoric Correlation Estimation

The full item-by-item cross-classifications generate a contingency table with $C = K^n$ cells. Let $\boldsymbol{\pi}$ be the $C \times 1$ vector of true (population) probabilities, with the corresponding sample proportions $\mathbf{p}$. We know from the standard theory of discrete multivariate analysis that $\mathbf{p}$ converges in distribution to $\boldsymbol{\pi}$

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Xi}), \tag{3.3}$$

when the sample size N tends to infinity and $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$. We also know from work on limited-information goodness-of-fit estimation and testing (e.g., Maydeu-Olivares & Joe, 2005) that for each pair of observed variables there exist $(K-1)^2$ unique marginal probabilities. These marginal probabilities are

full-rank linear transformations of the cell probabilities. Specifically, let $\mathbf{L}_{ij}$ be an operator matrix of order $K^2 \times C$ that combines the cell probabilities into the marginal probabilities for item pair (i, j):

$$\mathbf{p}_{ij} = \mathbf{L}_{ij}\mathbf{p}, \quad \boldsymbol{\pi}_{ij} = \mathbf{L}_{ij}\boldsymbol{\pi}, \tag{3.4}$$

where $\mathbf{p}_{ij}$ and $\boldsymbol{\pi}_{ij}$ denote the unique marginal moments for the sample and for the population. We can see that the asymptotic distribution of $\mathbf{p}_{ij}$ is

$$\sqrt{N}(\mathbf{p}_{ij} - \boldsymbol{\pi}_{ij}) \xrightarrow{D} \mathcal{N}_{K^2}(\mathbf{0}, \boldsymbol{\Xi}_{ij}), \tag{3.5}$$

where $\boldsymbol{\Xi}_{ij} = \mathbf{L}_{ij}\boldsymbol{\Xi}\mathbf{L}'_{ij}$ i.e.,

$$\boldsymbol{\Xi}_{ij} = \mathbf{L}_{ij}\text{diag}(\boldsymbol{\pi})\mathbf{L}'_{ij} - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}'_{ij}. \tag{3.6}$$

It is important to note that $\boldsymbol{\Xi}_{ij}$ can be estimated consistently by plugging in sample proportions.

We are now ready to discuss thresholds and polychoric correlation estimation. Thresholds and polychoric correlations are determined implicitly from the maximized pairwise likelihood for each item pair (i, j). The following description of estimating thresholds and polychoric correlations follows Olinsky, Chen, and Harlow (1979) and Jöreskog (1994). Assuming we are given observed frequencies, n_{kl}, in category k for item i and category l on an item j, where $k = 0, 1, 2, ..., K-1$ and $l = 0, 1, 2, ..., K-1$. The pairwise likelihood is

$$L \propto \prod_{k=1}^{K-1} \prod_{l=1}^{K-1} \pi_{kl}^{n_{kl}}. \tag{3.7}$$

The model-implied probability π_{kl} that an observation falls into the category

k and l for an item pair (i, j) is the following double integral

$$\pi_{kl} = \int_{\tau_{i,k-1}}^{\tau_{i,k}} \int_{\tau_{j,l-1}}^{\tau_{j,l}} \phi(x, y; \rho_{ij}) dx dy, \tag{3.8}$$

where

$$\phi(x, y; \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left(- \frac{(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)} \right)$$

is the standard bivariate normal density with (polychoric) correlation ρ. The maximization of the pairwise likelihood leads to estimates of thresholds and the polychoric correlation.

In practice, the maximization is often done in two stages. First, the thresholds are considered fixed upon estimation. They are computed directly from the inverse normal cumulative distribution function, $\hat{\tau}_k = \Phi^{-1}(\dot{p}_k)$, where $\dot{p}_k$ is the observed cumulative category proportion for item i up to category k. We use Φ to denote the univariate normal cumulative distribution function. In the second stage, the polychoric correlation is estimated by differentiating the log-likelihood and finding the zero of the log-likelihood gradient. While the two-step procedure is theoretically not optimal, it is computationally far less burdensome than the simultaneous estimation of all parameters. The resulting estimates are usually close to the simultaneous solution (Olinsky et al., 1979).

3.3 Estimation of the Asymptotic Covariance Matrix

Let $\boldsymbol{\sigma}_{ij} = (\boldsymbol{\tau}_i, \boldsymbol{\tau}_j, \rho_{ij})$ be the $2(K-1)+1$ vector of thresholds and polychoric correlation for item pair (i, j). Let $\mathbf{G}(\hat{\boldsymbol{\sigma}}_{ij}, \mathbf{p}_{ij}) = \mathbf{0}$ be the nonlinear implicit equations derived from the pairwise likelihood for item pair (i, j), where the pairwise maximum likelihood solution is $\hat{\boldsymbol{\sigma}}_{ij}$. One can show (e.g., Christoffersson & Gunsjö, 1996, Equation 2) with the help of the mean value theorem and implicit differen-

tiation that

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\sigma}}_{ij} - \boldsymbol{\sigma}_{ij}) &= -\Big(\frac{\partial \mathbf{G}(\hat{\boldsymbol{\sigma}}_{ij}, \mathbf{p}_{ij})}{\partial \boldsymbol{\sigma}_{ij}}\Big)^{-1}\Big(\frac{\partial \mathbf{G}(\hat{\boldsymbol{\sigma}}_{ij}, \mathbf{p}_{ij})}{\partial \boldsymbol{\sigma}_{ij}}\Big)\sqrt{N}(\boldsymbol{\pi}_{ij} - \boldsymbol{\pi}_{ij}) \\ &= \Big(\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}\Big)\sqrt{N}(\mathbf{p}_{ij} - \boldsymbol{\pi}_{ij}).\end{aligned} \tag{3.9}$$

Thus the asymptotic distribution of $\hat{\boldsymbol{\sigma}}_{ij}$ is

$$\sqrt{N}(\hat{\boldsymbol{\sigma}}_{ij} - \boldsymbol{\sigma}_{ij}) = \Big(\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}\Big)\sqrt{N}(\mathbf{p}_{ij} - \boldsymbol{\pi}_{ij}) \xrightarrow{D} \mathcal{N}_{2(K-1)+1}\Big(\mathbf{0}, \frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}\boldsymbol{\Xi}_{ij}\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}'\Big). \tag{3.10}$$

Because $\mathbf{p}_{ij}$ and $\mathbf{p}_{kl}$ are different linear transformations of the same underlying multinomial cell probabilities, their asymptotic covariance matrix is equal to $\boldsymbol{\Xi}_{ijkl} = \mathbf{L}_{ij}\boldsymbol{\Xi}\mathbf{L}_{kl}'$. This implies that the asymptotic covariances between $\hat{\boldsymbol{\sigma}}_{ij}$ and $\hat{\boldsymbol{\sigma}}_{kl}$ can be approximated as

$$\boldsymbol{\Gamma}_{ij,kl} = \Big(\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}\Big)\boldsymbol{\Xi}_{ij}\Big(\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}}\Big)'. \tag{3.11}$$

For technical details of computing $\frac{\partial \boldsymbol{\sigma}_{ij}}{\partial \boldsymbol{\pi}_{ij}} = -\Big(\frac{\partial \mathbf{G}(\hat{\boldsymbol{\sigma}}_{ij}, \mathbf{p}_{ij})}{\partial \boldsymbol{\sigma}_{ij}}\Big)^{-1}\Big(\frac{\partial \mathbf{G}(\hat{\boldsymbol{\sigma}}_{ij}, \mathbf{p}_{ij})}{\partial \mathbf{p}_{ij}}\Big)$, please refer to Olinsky et al. (1979).

3.4 Estimation of Structural Parameters

Let us denote the estimated polychoric correlation matrix as $\hat{\mathbf{P}}$, and let the unique elements of the matrix be denoted $\hat{\boldsymbol{\rho}} = \text{vech}(\hat{\mathbf{P}})$, where vech$(\cdot)$ stands for the half vectorization operator that returns the lower-half of a correlation matrix. From the previous section, we see that

$$\sqrt{N}(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}) \xrightarrow{D} \mathcal{N}_{\frac{n(n-1)}{2}}(\mathbf{0}, \boldsymbol{\Gamma}), \tag{3.12}$$

where $\boldsymbol{\rho} = \text{vech}(\mathbf{P})$ and the asymptotic covariance matrix of unique polychoric correlations is $\boldsymbol{\Gamma}$. The elements of $\boldsymbol{\Gamma}$ can be consistently estimated by repeated application of the formula in Equation (3.11).

Estimation of the structural parameters in $\boldsymbol{\theta}$ is typically accomplished as a final stage of estimation by minimizing a quadratic form discrepancy function of the following form (Browne, 1984, Equation 2.7) over $\boldsymbol{\theta}$

$$F(\boldsymbol{\theta}) = [\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})]' \mathbf{V} [\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})], \tag{3.13}$$

where $\mathbf{V}$ is a weight matrix. If one chooses to use weighted least squares (WLS) estimation, then $\mathbf{V} = \boldsymbol{\Gamma}^{-1}$. If ones choice is unweighted least squares (ULS), then $\mathbf{V}$ is an identity matrix. If diagonally weighted least squares (DWLS) is used, then $\mathbf{V} = [\text{diag}(\boldsymbol{\Gamma})]^{-1}$. Except for WLS, other estimators are not asymptotically optimal (not minimum variance estimators), but they may be more stable for smaller and more realistic sample sizes encountered in empirical research. For WLS estimation, $(N-1)$ times the minimized discrepancy function value is distributed in large samples as a central chi-square variable under correct model specification. For ULS or DWLS, corrections to standard errors and fit statistics are generally needed.

CHAPTER 4

Alternative Multiple Imputation Inferential Procedure

Up to this point we have discussed the foundational aspects of SEM for categorical variables without consideration of the issue of missing data. Let us now discuss the issue of missing data in this context. We label the incomplete observed data as $\mathbf{O}$, and the missing data as $\mathbf{X}$. In multiple imputation, we draw M sets of imputations. For imputation m, the complete data set as a result of multiple imputation is $\mathbf{Y}^{(m)} = (\mathbf{O}, \mathbf{X}^{(m)})$. We may estimate the polychoric correlations from imputation m. Let the polychoric correlations be denoted $\boldsymbol{\rho}^{(m)}$, and the corresponding asymptotic covariance matrix be $\boldsymbol{\Sigma}^{(m)}$.

4.1 The Standard Approach

Let us first introduce the typical approach of estimating a structural equation model under multiple imputation. For each imputation, we obtain parameter estimate $\hat{\boldsymbol{\theta}}^{(m)}$ again by minimizing the general quadratic form discrepancy function (Browne, 1984, Equation 2.7)

$$F(\boldsymbol{\theta}) = [\boldsymbol{\rho}^{(m)} - \boldsymbol{\rho}(\boldsymbol{\theta})]' \mathbf{V}^{(m)} [\boldsymbol{\rho}^{(m)} - \boldsymbol{\rho}(\boldsymbol{\theta})], \tag{4.1}$$

where $\mathbf{V}^{(m)}$ is the weight matrix associated with imputation m. To obtain a single set of parameter estimates, the parameter estimates are averaged over the M imputations.

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)} \tag{4.2}$$

where $\hat{\theta}^{(m)}$ is a generic component of $\hat{\boldsymbol{\theta}}^{(m)}$. Standard errors can also be computed in a straightforward manner with the standard formula for multiple imputation Rubin (1987), combining the within-imputation variance and the between-imputation variance. The within-imputation variance, V_W, is the average of the squared standard errors over the M imputations,

$$V_W = \frac{1}{M} \sum_{m=1}^{M} \left(SE(\hat{\theta}^{(m)}) \right)^2, \tag{4.3}$$

where $SE(\hat{\theta}^{(m)})$ refers to the standard error estimate from imputation m. The between-imputation variance, V_B, is

$$V_B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \bar{\theta})^2, \tag{4.4}$$

which accounts for uncertainty in parameter estimates due to missing data. The total error variance is obtained by combining the within-imputation variance and the between-imputation variance as follows:

$$V_T = V_W + (1 + M^{-1}) V_B. \tag{4.5}$$

4.2 The New Approach

In the new approach, the structural equation model is fitted after all multiple imputations have been combined. Specifically, researchers combine the imputations to obtain a single matrix of polycoric correlations along with its asymptotic covariance matrix that are corrected for missing data. These polycoric correlation matrix and its asymptotic covariance matrix become the components of business-as-usual estimation in the second stage and statistical inference.

First, we average the polychoric correlations as

$$\bar{\boldsymbol{\rho}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\rho}^{(m)}. \tag{4.6}$$

We now have a single discrepancy function to minimize:

$$F(\boldsymbol{\theta}) = [\bar{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})]' \mathbf{V} [\bar{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})]. \tag{4.7}$$

However, simply averaging the asymptotic covariance matrix as

$$\bar{\boldsymbol{\Gamma}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\Gamma}^{(m)}. \tag{4.8}$$

will not lead to the correct weights for either WLS estimation or subsequent corrections to test statistics or standard errors if ULS or DWLS are used. This is because $\bar{\boldsymbol{\Gamma}}$ does not take into account the added uncertainty due to the missing data. Specifically, $\bar{\boldsymbol{\Gamma}}$ only captures uncertainty based on complete data, and uncertainty about the averaged polychoric correlations $\bar{\boldsymbol{\rho}}$ is not fully accounted for under missing data.

Fortunately, to obtain the corrected weight matrix, one only needs to add to $\bar{\boldsymbol{\Gamma}}$ a component that reflects the between-imputation variance in the estimated polychoric correlations $\boldsymbol{\rho}^{(m)}$:

$$\tilde{\boldsymbol{\Gamma}} = \left[\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\Gamma}^{(m)} \right] + \frac{M+1}{(M-1)M} \left[\sum_{m=1}^{M} (\boldsymbol{\rho}^{(m)} - \bar{\boldsymbol{\rho}})(\boldsymbol{\rho}^{(m)} - \bar{\boldsymbol{\rho}})' \right]. \tag{4.9}$$

The inverse of $\tilde{\boldsymbol{\Gamma}}$ will be the correct weight matrix to use in estimation or inference for the structural parameters in $\boldsymbol{\theta}$. Note that assuming proper imputations and infinite M, the resulting repeated-imputation inference is valid. That is to say, with a large sample size, $\bar{\boldsymbol{\rho}}$ is a consistent estimate of $\boldsymbol{\rho}$, and $\sqrt{N}(\bar{\boldsymbol{\rho}} - \boldsymbol{\rho})$ is normally distributed with zero means and asymptotic covariance matrix $\boldsymbol{\Gamma}$, which

is consistently estimated by $\tilde{\mathbf{\Gamma}}$ (Rubin, 1987).

The parameters that minimize the quadratic form discrepancy function in Equation (4.7) is referred to as $\tilde{\boldsymbol{\theta}}$. Under broad conditions, the minimizer $\tilde{\boldsymbol{\theta}}$ of the quadratic form discrepancy function in Equation (4.7) is consistent and asymptotically normal. As is typical, we have a choice of WLS, ULS or DWLS estimation. In WLS, the inverse of $\tilde{\mathbf{\Gamma}}$ is plugged into $\mathbf{V}$. In DWLS, the diagonal elements from the inverse of $\tilde{\mathbf{\Gamma}}$ are used as weights. In ULS, the identity matrix serves as the weight. In ULS and DWLS, a subsequent step of correcting the test statistic is required because the weight matrix is not correctly specified. To obtain the correct test statistic, we apply Browne (1984) Proposition 4.

Given model-implied moments, the residual moments are $\mathbf{e} = \bar{\boldsymbol{\rho}} - \boldsymbol{\rho}(\bar{\boldsymbol{\theta}})$. We define a residual-based test statistic

$$\tilde{T}_B = N\mathbf{e}'\tilde{\mathbf{\Omega}}\mathbf{e}, \tag{4.10}$$

where $\mathbf{\Omega} = \mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{-1}\tilde{\mathbf{\Delta}}(\tilde{\mathbf{\Delta}}'\mathbf{\Gamma}^{-1}\tilde{\mathbf{\Delta}})^{-1}\tilde{\mathbf{\Delta}}')^{-1}\tilde{\mathbf{\Delta}}'\mathbf{\Gamma}^{-1}$, and

$$\tilde{\mathbf{\Delta}} = \mathbf{\Delta}(\tilde{\boldsymbol{\theta}}) = \frac{\partial \boldsymbol{\rho}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

is the Jacobian matrix of the structural model evaluated at the parameter estimate $\tilde{\boldsymbol{\theta}}$. Under Browne (1984)'s Proposition 4, this residual-based test statistic is asymptotically chi-squared for any consistent and asymptotically normal estimator. This test statistic can be further extended to yield a statistic that may be better suited for smaller sample size, following the logic of T_{YB}, originally proposed by Yuan and Bentler (1997). T_{YB} is an adjustment of T_B while retaining the asymptotic chi-square distribution of T_B. T_{YB} tends to perform well for a small sample size (Maydeu-Olivares, Cai, & Hernndez, 2011; Yuan & Bentler, 1997, 2000). Our

corrected statistic $\tilde{T}_{YB}$, can be computed as

$$\tilde{T}_{YB} = \frac{\tilde{T}_B}{1 + N\tilde{T}_B/(N-1)}. \tag{4.11}$$

CHAPTER 5

Simulation Study 1: Calibration

The goal of the first simulation study is to show that the test statistic $\tilde{T}_B$ and $\tilde{T}_{YB}$ are asymptotically chi-square distributed under the null hypothesis that the model fits exactly, with the latter exhibiting better finite sample behavior. The simulation is carried out in four steps: 1) generation of complete and missing data, 2) multiple imputation for missing data, 3) combining multiple imputation, and 4) model fitting. The simulation conditions include the following four aspects: 1) the missing data mechanism, 2) missing data rate, 3) sample size, and 4) number of categories. In a fully crossed design, 500 replications were attempted for each of the conditions.

5.1 Data Generation

5.1.1 Generation of Complete and Missing data

The data generating model is a confirmatory factor analysis (CFA) model with 9 items and 3 factors. The covariance structure is $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Phi}$. The population factor loading matrix is

$$\boldsymbol{\Lambda}' = \begin{pmatrix} 0.7 & 0.8 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0.8 & 0.9 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0.8 & 0.8 \end{pmatrix} \tag{5.1}$$

and the factor correlation matrix is

$$\mathbf{\Psi} = \begin{pmatrix} 1.0 & & \\ 0.4 & 1.0 & \\ 0.3 & 0.5 & 1.0 \end{pmatrix} \tag{5.2}$$

For identification we let $\mathbf{\Psi} = \mathbf{I} - \text{diag}(\mathbf{\Lambda\Psi\Lambda}')$. It follows that $\mathbf{\Phi} = \text{diag}(0.51, 0.36, 0.19, 0.51, 0.36, 0.19, 0.36, 0.36, 0.36)$ and $\mathbf{\Sigma}$ is a correlation matrix. There are 12 free parameters, and the models degrees of freedom is 24.

We generated multivariate normal underlying response variables corresponding to the data generating model above. We examined 4 different sample sizes ($N = 250, 500, 1000, 2500$). The data were generated using R (R Core Team, 2017). The continuous underlying variables were discretized. We considered 3 cases for the number of categories $K = 2$, 3, and 5. Table 5.1 presents those thresholds. These thresholds were systematically chosen to provide coverage of possible ranges of thresholds commonly seen in practical settings.

Table 5.1: Generating Thresholds

Item	$K = 2$	$K = 3$		$K = 5$			
	$\tau_{i,1}$	$\tau_{i,1}$	$\tau_{i,2}$	$\tau_{i,1}$	$\tau_{i,2}$	$\tau_{i,3}$	$\tau_{i,4}$
1	-0.5	-1.0	0.0	-1.5	-0.7	0.2	1.0
2	0.0	-0.5	0.5	-1.0	-0.3	0.3	1.0
3	0.5	0.0	1.0	-1.0	-0.2	0.7	1.5
4	0.5	0.0	1.0	-1.0	-0.2	0.7	1.5
5	0.0	-0.5	0.5	-1.0	-0.3	0.3	1.0
6	-0.5	-1.0	0.0	-1.5	-0.7	0.2	1.0
7	-0.5	-1.0	0.0	-1.5	-0.7	0.2	1.0
8	0.0	-0.5	0.5	-1.0	-0.3	0.3	1.0
9	0.5	0.0	1.0	-1.0	-0.2	0.7	1.5

Note. K: number of categories

We examined 3 missing data conditions (NOMISS, MCAR, and MAR). Note that we included the no missing data case (NOMISS) purely as a benchmark.

Missing data were simulated using a variant of the procedures described by Lee and Cai (2012). We first describe the low missing data rate condition. For MCAR, each row of complete data was tested by a fair dice (1/6th chance) to determine whether missing values should be present or not. Once a row was chosen, we set the values of the last three items to missing. For MAR, the probabilities of missingness of the last three items depend on the mean of the first six items (Z). This was accomplished by dividing the distribution of Z into quartiles and setting the missingness probabilities of the four quartiles to (.50, .20, .075, .025). Implementation of this set of procedures in R (R Core Team, 2017) resulted in 17% and 20% of all observations missing, respectively for MCAR and MAR conditions. For the high missing data rate condition, the procedure remains the same, but we doubled the missing data probabilities. For MCAR, instead of rolling a fair dice, a 3-sided dice was tossed. For MAR, the missingness probabilities of the four quartiles were changed to (1.0, .40, .15, .05). Implementation in R (R Core Team, 2017) gave us about 33% and 40% of observations missing, for MCAR and MAR respectively. As we doubled the missing data probabilities, the missing rates are about twice those of the low missing data condition.

5.1.2 Multiple Imputation for Missing Data

For the missing data (MCAR, MAR) conditions, multiple imputation was performed with FCS (or MICE) using the software program BLImP (Keller & Enders, 2017). The details of the categorical variable imputation implemented in BLImP can be found in Enders, Keller, and Levy (2018). Burn-in interval and thinning interval were both set to 1,000.

We imputed 20 times for the low missing data rate condition and 60 for the high missing data rate condition. The decision on the number of imputations was based on the relative efficiency (RE) of imputations. The larger number of imputations is consistent with recent research that recommended more imputations

than the traditional recommendation of three to five (e.g., Bodner, 2008; Graham, Olchowski, & Gilreath, 2007; von Hippel, 2009; White, Royston, & Wood, 2009). We computed RE as follows.

First, given within-imputation variance V_W, between-imputation variance V_B, and total sample variance V_T, computed after M imputations, the fraction of missing information (FMI) adjusting the finite number of imputations can be expressed as

$$\text{FMI} = \frac{(1 + M^{-1})V_B + \frac{2V_w}{v+3}}{V_T}, \tag{5.3}$$

where $v = (M-1)(1 + V_W/(1 + M^{-1})V_B)^2$ is a degrees of freedom value. This represents the proportion of the total variance due to missing data (Enders, 2010). Since we consider combining the polychoric correlations, the V_w, V_B, and V_T terms are in matrix forms. Recall that V_w and V_B correspond to the first term and the second term in Equation (4.9). Hence, we need to summarize each matrix as a scalar. This we accomplish with the trace operator, though of course other operators may be used (e.g., the log-determinant). Now that we have FMI, RE is computed as

$$\text{RE} = \left(1 + \frac{\text{FMI}}{M}\right)^{-1}. \tag{5.4}$$

Tables 5.2 and 5.3 display the FMI and RE as a function of the number of imputations for the low and high missing data conditions. No noticeable difference was found across sample sizes and number of categories. The higher FMI for MAR compared to MCAR is a result of the slightly higher missing rates. A desirable level of RE may differ depending on the purpose of research. For example, Bodner (2008) pointed out that inferential procedures such as hypothesis testing with p-values and confidence intervals require more imputations. Our interest is on statistical inference, so we set the number of imputations to achieve RE close to or higher than .990, resulting in M of 20 and 60, for the low and high missing data rates.

Table 5.2: Fraction of Missing Information (FMI) and Relative Efficiency (RE) for the Low Missing Data Rate Condition

			$M = 5$		$M = 10$		$M = 20$		$M = 25$	
	K	N	FMI	RE	FMI	RE	FMI	RE	FMI	RE
MCAR	2	250	0.143	0.972	0.122	0.988	0.116	0.994	0.143	0.994
		500	0.134	0.974	0.12	0.988	0.112	0.994	0.135	0.995
		1000	0.123	0.976	0.117	0.988	0.114	0.994	0.136	0.995
		2500	0.132	0.974	0.12	0.988	0.114	0.994	0.135	0.995
	3	250	0.132	0.974	0.115	0.989	0.11	0.995	0.11	0.996
		500	0.13	0.975	0.114	0.989	0.109	0.995	0.106	0.996
		1000	0.135	0.974	0.115	0.989	0.108	0.995	0.109	0.996
		2500	0.132	0.974	0.118	0.988	0.113	0.994	0.111	0.996
	5	250	0.128	0.975	0.12	0.988	0.11	0.995	0.109	0.996
		500	0.135	0.974	0.124	0.988	0.114	0.994	0.112	0.996
		1000	0.136	0.973	0.118	0.988	0.113	0.994	0.113	0.995
		2500	0.115	0.977	0.108	0.989	0.104	0.995	0.104	0.996
MAR	2	250	0.16	0.969	0.149	0.985	0.147	0.993	0.144	0.994
		500	0.182	0.965	0.167	0.984	0.154	0.992	0.154	0.994
		1000	0.193	0.963	0.164	0.984	0.157	0.992	0.155	0.994
		2500	0.171	0.967	0.163	0.984	0.158	0.992	0.158	0.994
	3	250	0.217	0.958	0.195	0.981	0.183	0.991	0.181	0.993
		500	0.232	0.956	0.199	0.98	0.19	0.991	0.185	0.993
		1000	0.205	0.961	0.182	0.982	0.172	0.991	0.172	0.993
		2500	0.211	0.959	0.187	0.982	0.178	0.991	0.178	0.993
	5	250	0.262	0.95	0.239	0.977	0.219	0.989	0.221	0.991
		500	0.218	0.958	0.193	0.981	0.191	0.991	0.19	0.992
		1000	0.244	0.953	0.207	0.98	0.205	0.99	0.202	0.992
		2500	0.251	0.952	0.216	0.979	0.202	0.99	0.199	0.992

Note. M: number of imputations; K: number of categories; N: sample size; FMI: fraction of missing information; RE: relative efficiency

Table 5.3: Fraction of Missing Information (FMI) and Relative Efficiency (RE) for the High Missing Data Rate Condition

			$M = 5$		$M = 10$		$M = 20$		$M = 40$		$M = 60$	
	K	N	FMI	RE	FMI	RE	FMI	RE	FMI	RE	FMI	RE
MCAR	2	250	0.308	0.942	0.274	0.973	0.256	0.987	0.245	0.994	0.244	0.996
		500	0.294	0.944	0.279	0.973	0.269	0.987	0.256	0.994	0.256	0.996
		1000	0.318	0.94	0.285	0.972	0.273	0.987	0.263	0.993	0.26	0.996
		2500	0.297	0.943	0.271	0.974	0.26	0.987	0.248	0.994	0.249	0.996
	3	250	0.278	0.947	0.25	0.976	0.245	0.988	0.238	0.994	0.233	0.996
		500	0.297	0.944	0.261	0.975	0.24	0.988	0.238	0.994	0.239	0.996
		1000	0.297	0.944	0.251	0.976	0.246	0.988	0.244	0.994	0.238	0.996
		2500	0.299	0.944	0.262	0.974	0.245	0.988	0.244	0.994	0.241	0.996
	5	250	0.276	0.948	0.258	0.975	0.258	0.987	0.248	0.994	0.245	0.996
		500	0.281	0.947	0.249	0.976	0.243	0.988	0.237	0.994	0.236	0.996
		1000	0.286	0.946	0.258	0.975	0.251	0.988	0.238	0.994	0.237	0.996
		2500	0.279	0.947	0.243	0.976	0.237	0.988	0.229	0.994	0.227	0.996
MAR	2	250	0.459	0.916	0.424	0.959	0.41	0.98	0.401	0.99	0.397	0.993
		500	0.437	0.92	0.418	0.96	0.388	0.981	0.382	0.991	0.379	0.994
		1000	0.416	0.923	0.419	0.96	0.392	0.981	0.388	0.99	0.384	0.994
		2500	0.461	0.916	0.434	0.958	0.42	0.979	0.407	0.99	0.405	0.993
	3	250	0.568	0.898	0.521	0.951	0.489	0.976	0.48	0.988	0.478	0.992
		500	0.577	0.897	0.529	0.95	0.498	0.976	0.486	0.988	0.481	0.992
		1000	0.546	0.902	0.485	0.954	0.48	0.977	0.468	0.988	0.463	0.992
		2500	0.5	0.909	0.468	0.955	0.466	0.978	0.471	0.988	0.464	0.992
	5	250	0.526	0.905	0.51	0.951	0.502	0.976	0.497	0.988	0.495	0.992
		500	0.586	0.895	0.535	0.949	0.507	0.975	0.497	0.988	0.492	0.992
		1000	0.562	0.899	0.52	0.951	0.498	0.976	0.496	0.988	0.496	0.992
		2500	0.551	0.901	0.523	0.95	0.501	0.976	0.508	0.987	0.5	0.992

Note. M: number of imputations; K: number of categories; N: sample size; FMI: fraction of missing information; RE: relative efficiency

5.2 Model Fitting

For each complete or imputed data set, polychoric correlations and the associated asymptotic covariance matrix of polychoric correlations were computed using the lavaan package in R (R Core Team, 2017). The correlations and the asymptotic covariance matrix were further combined in R.

The combined correlations matrix and the associated asymptotic covariance matrix were used as the inputs for fitting the CFA model. We used ULS estimation for each replication. The reason that we opted for ULS over WLS or DWLS is that ULS provides more accurate and less variable parameter estimates as well as more precise standard errors (Forero et al., 2009). The corrected test statistics, $\tilde{T}_B$ and $\tilde{T}_{YB}$, were computed at the end.

5.3 Simulation Results

Tables 5.4 - 5.6 show the Type I error rates at the .01, .05, and .10 α-levels for T_B when $K = 2$, 3, and 5, respectively. We removed invalid replications having zero or negative unique variances. Tables 5.7 - 5.9 present the same information for $\tilde{T}_{YB}$ when $K = 2$, 3, and 5, respectively. We expect that the statistic would be chi-square distributed, and that the observed means calculated across the valid replications would be close to 24, the degrees of freedom of the model, and that the variances would be twice the degrees of freedom. Furthermore, the empirical rejection rates of $\tilde{T}_B$ and $\tilde{T}_{YB}$ should be close to the nominal α-level.

Examining $\tilde{T}_B$ for the NOMISS condition, the statistics are better calibrated as the sample size increases. For $N = 250$ and $N = 500$, the empirical rejection rates are much higher than the nominal levels. From $N = 1,000$, the chi-square approximation begins to improve. This is consistent with prior research (Maydeu-Olivares et al., 2011). $\tilde{T}_{YB}$, on the other hand, does perform better for smaller

sample sizes, though it tends to be more conservative than liberal (see Table 7 - 9). This mirrors the results from (Yuan & Bentler, 1997). Also, it is unsurprising that the number of valid replications (Reps) also increases as the sample size increases.

Turning to the MCAR and MAR conditions, it appears that the results are comparable to those for the NOMISS condition. Very little difference is found across missing data mechanisms (MCAR or MAR) and number of categories ($K = 2, 3, 5$). In addition, the statistic is well calibrated regardless of missing data rates (low or high).

Table 5.4: Type I Error Rates for T_B (K=2)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.80	26.828	67.753	8.766	55.738	0.047	0.127	0.184
	500	0.92	25.343	52.326	8.456	53.900	0.013	0.078	0.155
	1000	0.98	24.641	54.884	7.637	56.359	0.024	0.057	0.128
	2500	1.00	24.636	49.209	10.039	50.045	0.008	0.062	0.124
MCAR				Low missing rates (%)					
	250	0.79	26.492	68.812	10.727	59.096	0.050	0.116	0.176
	500	0.92	25.446	51.713	10.077	55.554	0.011	0.072	0.158
	1000	0.98	24.611	56.456	8.649	54.345	0.020	0.061	0.129
	2500	1.00	24.777	52.333	10.045	53.884	0.008	0.070	0.130
				High missing rates (%)					
	250	0.76	26.117	66.645	9.179	58.383	0.042	0.103	0.177
	500	0.90	25.690	48.531	8.270	52.612	0.020	0.058	0.125
	1000	0.98	24.848	53.980	9.431	55.190	0.025	0.065	0.125
	2500	1.00	25.467	55.961	8.342	50.906	0.020	0.082	0.150
MAR				Low missing rates (%)					
	250	0.78	26.556	72.634	7.770	59.750	0.054	0.121	0.175
	500	0.92	25.397	55.152	8.069	51.749	0.020	0.092	0.159
	1000	0.98	24.961	59.349	9.473	56.269	0.029	0.070	0.125
	2500	1.00	24.988	49.992	8.756	48.961	0.006	0.054	0.124
				High missing rates (%)					
	250	0.67	26.953	82.179	8.299	73.923	0.063	0.123	0.183
	500	0.89	25.646	58.018	11.057	60.946	0.038	0.074	0.132
	1000	0.98	25.218	55.682	8.631	55.324	0.025	0.072	0.133
	2500	1.00	25.049	54.689	9.190	51.514	0.016	0.084	0.140

Note. N: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

Table 5.5: Type I Error Rates for T_B (K=3)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.93	26.174	57.853	9.305	52.754	0.026	0.094	0.176
	500	0.99	25.274	55.903	6.643	55.377	0.020	0.083	0.137
	1000	1.00	24.107	48.548	9.928	45.496	0.016	0.056	0.094
	2500	1.00	24.130	43.010	9.200	45.084	0.008	0.050	0.092
MCAR				Low missing rates (%)					
	250	0.92	25.939	62.743	8.727	56.380	0.024	0.098	0.172
	500	0.98	25.202	55.592	9.190	52.287	0.026	0.086	0.126
	1000	1.00	24.329	50.226	9.108	51.622	0.010	0.064	0.114
	2500	1.00	24.439	45.262	8.683	45.551	0.008	0.042	0.122
				High missing rates (%)					
	250	0.91	24.905	51.421	8.123	48.818	0.024	0.066	0.126
	500	0.99	24.766	56.061	8.978	57.520	0.026	0.081	0.124
	1000	1.00	24.320	48.035	10.410	51.181	0.014	0.056	0.108
	2500	1.00	24.060	48.738	10.238	49.666	0.014	0.054	0.096
MAR				Low missing rates (%)					
	250	0.91	25.924	59.342	9.682	49.198	0.015	0.096	0.173
	500	0.99	25.268	50.114	6.154	53.016	0.018	0.063	0.125
	1000	1.00	24.372	51.035	6.863	51.785	0.012	0.054	0.110
	2500	1.00	24.634	48.075	9.176	50.984	0.010	0.066	0.130
				High missing rates (%)					
	250	0.84	24.478	50.967	8.580	51.083	0.014	0.062	0.115
	500	0.98	24.506	51.234	8.649	51.461	0.012	0.070	0.131
	1000	1.00	24.381	49.342	10.241	50.949	0.016	0.054	0.126
	2500	1.00	24.213	45.850	8.867	51.593	0.008	0.054	0.108

Note. N: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

Table 5.6: Type I Error Rates for T_B (K=5)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.99	26.814	68.066	5.873	54.964	0.045	0.123	0.194
	500	1.00	25.301	61.149	8.767	61.707	0.024	0.098	0.142
	1000	1.00	24.765	56.563	10.094	54.608	0.020	0.068	0.120
	2500	1.00	24.148	48.057	9.698	48.026	0.016	0.064	0.102
MCAR				Low missing rates (%)					
	250	0.98	26.101	61.195	7.372	54.561	0.026	0.098	0.169
	500	1.00	25.005	61.914	8.362	61.195	0.022	0.092	0.146
	1000	1.00	24.736	51.309	9.138	51.220	0.020	0.068	0.114
	2500	1.00	24.397	46.283	10.257	52.780	0.008	0.050	0.110
				High missing rates (%)					
	250	0.98	25.464	60.419	8.987	51.461	0.024	0.102	0.161
	500	1.00	24.552	55.988	10.089	53.845	0.020	0.056	0.116
	1000	1.00	24.730	53.993	10.197	53.590	0.016	0.070	0.132
	2500	1.00	24.399	52.262	10.256	62.217	0.014	0.058	0.108
MAR				Low missing rates (%)					
	250	0.99	25.670	57.918	8.509	55.030	0.024	0.079	0.162
	500	1.00	25.060	62.603	8.713	65.138	0.030	0.080	0.128
	1000	1.00	24.953	53.460	10.203	49.920	0.026	0.080	0.116
	2500	1.00	24.573	52.308	9.266	58.615	0.018	0.066	0.120
				High missing rates (%)					
	250	0.95	24.582	51.662	10.005	47.970	0.019	0.063	0.122
	500	1.00	24.291	53.335	8.000	53.566	0.016	0.068	0.120
	1000	1.00	24.567	56.180	7.433	51.216	0.020	0.074	0.130
	2500	1.00	24.817	51.686	8.281	52.476	0.020	0.070	0.120

Note. N: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

Table 5.7: Type I Error Rates for T_{YB} (K=2)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.80	24.013	43.167	8.467	45.510	0.010	0.047	0.100
	500	0.92	24.026	42.332	8.315	48.636	0.007	0.041	0.083
	1000	0.98	23.996	49.196	7.579	53.346	0.016	0.047	0.102
	2500	1.00	24.376	47.191	9.998	49.062	0.006	0.054	0.104
MCAR				Low missing rates (%)					
	250	0.79	23.736	43.873	10.282	47.724	0.008	0.050	0.088
	500	0.92	24.120	41.871	9.878	49.979	0.004	0.028	0.095
	1000	0.98	23.966	50.678	8.575	51.538	0.014	0.047	0.111
	2500	1.00	24.514	50.151	10.004	52.747	0.008	0.064	0.122
				High missing rates (%)					
	250	0.76	23.434	42.979	8.851	47.258	0.005	0.045	0.074
	500	0.90	24.347	39.211	8.135	47.584	0.004	0.040	0.071
	1000	0.98	24.194	48.388	9.343	52.298	0.016	0.047	0.098
	2500	1.00	25.188	53.565	8.314	49.890	0.016	0.076	0.134
MAR				Low missing rates (%)					
	250	0.78	23.778	45.970	7.534	48.150	0.010	0.054	0.100
	500	0.92	24.070	44.575	7.941	46.878	0.004	0.046	0.102
	1000	0.98	24.297	53.007	9.384	53.266	0.029	0.055	0.107
	2500	1.00	24.721	47.953	8.725	48.020	0.006	0.052	0.122
				High missing rates (%)					
	250	0.67	24.076	51.113	8.030	56.949	0.015	0.063	0.102
	500	0.89	24.291	46.242	10.817	54.301	0.018	0.049	0.090
	1000	0.98	24.546	49.742	8.557	52.419	0.018	0.059	0.098
	2500	1.00	24.780	52.386	9.156	50.474	0.016	0.072	0.132

Note. N: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

Table 5.8: Type I Error Rates for T_{YB} (K=3)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.93	23.504	38.079	8.969	43.501	0.002	0.028	0.062
	500	0.99	23.957	45.121	6.555	49.835	0.014	0.048	0.095
	1000	1.00	23.493	43.791	9.830	43.513	0.002	0.044	0.080
	2500	1.00	23.882	41.294	9.167	44.285	0.006	0.042	0.086
MCAR				Low missing rates (%)					
	250	0.92	23.298	41.216	8.431	45.937	0.004	0.028	0.072
	500	0.98	23.893	44.803	9.024	47.319	0.008	0.053	0.096
	1000	1.00	23.704	45.233	9.025	49.083	0.004	0.050	0.096
	2500	1.00	24.185	43.446	8.653	44.735	0.008	0.036	0.106
				High missing rates (%)					
	250	0.91	22.479	34.187	7.866	40.789	0.000	0.026	0.049
	500	0.99	23.496	45.265	8.819	51.564	0.008	0.047	0.091
	1000	1.00	23.697	43.260	10.302	48.685	0.008	0.042	0.080
	2500	1.00	23.811	46.740	10.196	48.698	0.010	0.044	0.086
MAR				Low missing rates (%)					
	250	0.91	23.295	39.412	9.318	41.054	0.000	0.018	0.074
	500	0.99	23.962	40.605	6.078	47.915	0.008	0.028	0.071
	1000	1.00	23.743	46.027	6.816	49.231	0.004	0.042	0.090
	2500	1.00	24.375	46.095	9.142	49.964	0.006	0.056	0.122
				High missing rates (%)					
	250	0.84	22.127	34.105	8.293	42.358	0.000	0.014	0.038
	500	0.98	23.269	41.637	8.501	46.641	0.004	0.033	0.080
	1000	1.00	23.754	44.427	10.137	48.475	0.006	0.040	0.092
	2500	1.00	23.963	43.986	8.835	50.549	0.006	0.042	0.104

Note. *N*: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

Table 5.9: Type I Error Rates for T_{YB} (K=5)

							Significance Level		
	N	Reps	Mean	Var	Min	Max	0.01	0.05	0.1
NOMISS	250	0.99	23.999	43.859	5.738	44.993	0.002	0.045	0.095
	500	1.00	23.974	49.231	9.616	54.904	0.010	0.054	0.106
	1000	1.00	24.113	50.673	9.993	51.775	0.018	0.060	0.100
	2500	1.00	23.898	46.087	9.660	47.120	0.006	0.048	0.096
MCAR				Low missing rates (%)					
	250	0.98	23.436	39.941	7.159	44.722	0.002	0.026	0.073
	500	1.00	23.704	50.004	8.224	54.498	0.008	0.048	0.104
	1000	1.00	24.090	46.049	9.055	48.719	0.014	0.046	0.094
	2500	1.00	24.143	44.373	10.215	51.688	0.008	0.040	0.096
				High missing rates (%)					
	250	0.98	22.914	39.796	8.673	42.618	0.000	0.024	0.073
	500	1.00	23.303	45.322	9.888	48.591	0.012	0.032	0.068
	1000	1.00	24.082	48.452	10.094	50.859	0.012	0.048	0.108
	2500	1.00	24.143	50.005	10.214	60.705	0.014	0.048	0.094
MAR				Low missing rates (%)					
	250	0.99	23.092	37.823	8.227	45.037	0.002	0.026	0.055
	500	1.00	23.753	50.044	8.564	57.604	0.020	0.040	0.098
	1000	1.00	24.295	47.981	10.100	47.542	0.014	0.062	0.100
	2500	1.00	24.314	50.074	9.232	57.271	0.014	0.066	0.106
				High missing rates (%)					
	250	0.95	22.211	34.501	9.617	40.195	0.000	0.019	0.044
	500	1.00	23.070	43.306	7.874	48.364	0.006	0.036	0.072
	1000	1.00	23.925	50.475	7.378	48.716	0.014	0.062	0.106
	2500	1.00	24.553	49.503	8.253	51.397	0.012	0.066	0.116

Note. N: sample size; Reps: proportion of valid replications; Var: variance of test statistic; Min: minimum value of test statistic; Max: maximum value of test statistic

CHAPTER 6

Simulation Study 2: Power

We conducted a second, smaller simulation to investigate the power of the proposed statistics, $\tilde{T}_B$ and $\tilde{T}_{YB}$ to detect model misspecification. To generate model misspecification, we utilized Tucker, Koopman, and Linn (1969) procedure. Specifically, we introduced 50 minor common factors that account for 10% of unique variance. This is a very mild degree of misspecification that cannot be easily accounted for with the confirmatory factor model.

Table 6.1 presents the results of the empirical rejection rates at the .05 nominal α-level (from the low missing rate condition). The results from the null conditions were added for comparison. Consistent with the simulation results of Type I error rate calibration, the powers of $\tilde{T}_B$ and $\tilde{T}_{YB}$ under MCAR and MAR are comparable to those under NOMISS. Note that it is not a surprise to see that $\tilde{T}_B$ is more powerful than $\tilde{T}_{YB}$, and the difference in power is reduced as K and N increase.

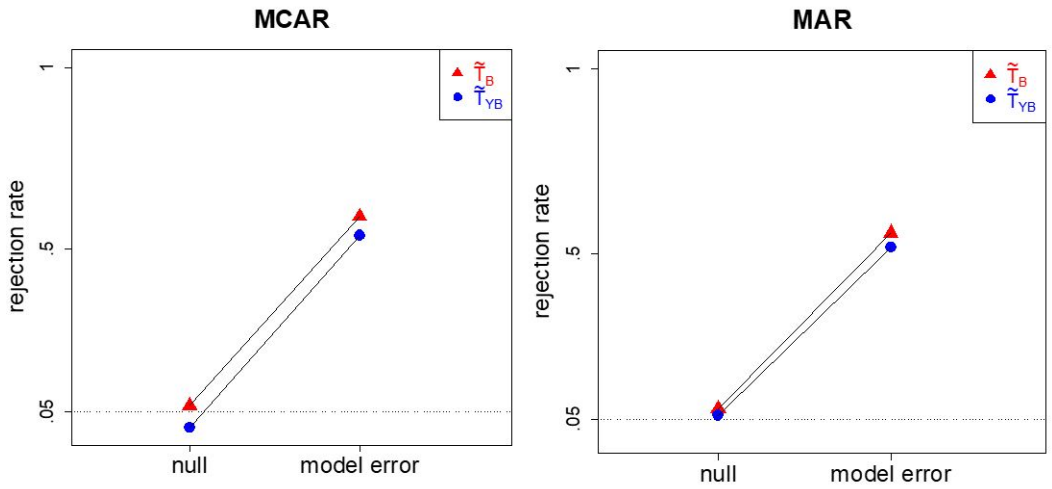
For the condition with low missing data rates, $K = 5$, and $N = 1,000$, we plotted the results of the empirical rejection rates at the .05 nominal α-level in Figure 1. The results from the null conditions were added for comparison. They clearly show that $\tilde{T}_B$ and $\tilde{T}_{YB}$ have power ($> .50$) to detect model misspecification with small degree of model error in the conditions examined.

Table 6.1: Power at $\alpha = .05$ level

		T_B					T_{YB}		
K	N	NOMISS	MCAR	MAR	K	N	NOMISS	MCAR	MAR
2	250	0.148	0.134	0.144	2	250	0.042	0.038	0.054
	500	0.126	0.136	0.148		500	0.066	0.072	0.094
	1000	0.144	0.156	0.164		1000	0.106	0.116	0.134
	2500	0.388	0.348	0.362		2500	0.358	0.330	0.336
3	250	0.190	0.170	0.144	3	250	0.082	0.062	0.066
	500	0.212	0.202	0.202		500	0.144	0.136	0.126
	1000	0.346	0.292	0.312		1000	0.306	0.246	0.266
	2500	0.728	0.690	0.684		2500	0.716	0.676	0.668
5	250	0.252	0.250	0.198	5	250	0.112	0.126	0.084
	500	0.364	0.356	0.300		500	0.272	0.280	0.216
	1000	0.630	0.590	0.556		1000	0.586	0.536	0.516
	2500	0.958	0.946	0.922		2500	0.954	0.940	0.902

Note. K: number of categories; N: sample size

Figure 6.1: Empirical Rejection Rates at $\alpha = .05$ of $\tilde{T}_B$ and $\tilde{T}_{YB}$ for MCAR and MAR



CHAPTER 7

Empirical Application

We utilize a data set to demonstrate how the proposed procedure works in practice. In the example, the data set had no missing observations originally, but we artificially created missingness, as in the simulation. For the empirical analysis, we used LISREL (Jöreskog & Sörbom, 2006). The procedure described in this paper can be implemented using LISREL or other commercial packages as long as one could combine polychoric correlations and the asymptotic covariance matrix after multiple imputation.

After multiple imputation, we ran PRELIS to obtain polychoric correlations and the associated asymptotic covariance matrix. As a side note, PRELIS produces the asymptotic covariance matrix in binary format. We use the BIN2ASC utility to convert the binary file into ASCII format. Then we combine the correlations and the asymptotic covariance matrix using R or any software that can manipulate ASCII input data. Finally, we run WLS, ULS, or DWLS estimation in LISREL and collect the residual-based statistic directly from the output, which corresponds to $\tilde{T}_B$.

7.1 Example 1

The source of the first data set is from the Korea Youth Panel Survey (KYPS), conducted by the National Youth Policy Institute (NYPI). Specifically, we used 4th grade elementary school students of the first wave in 2003 as our sample. The panel survey originally contains a variety of items on career choice, career plan,

academic performance and career preparation, leisure, daily life, etc. We chose a subset from a section on Attachment and used 12 items: 6 items measuring Parental attachment, 3 items on Teacher attachment, and 3 items on Friend attachment. The ratings were on a 5-point ordinal scale, from very untrue to very true. The example items for each construct are presented in Table 7.1

Table 7.1: Example Items for Attachment from KYPS

Construct	Exmaple Items
Parental Attachment	My parents and I have frequent conversations.
Teacher Attachment	Teachers treat me with love and affection.
Friend Attachment	I get along well with friends at school.

The model under consideration is a CFA model with the following factor loading matrix,

$$\mathbf{\Lambda}' = \begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & \lambda_{41} & \lambda_{51} & \lambda_{61} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{72} & \lambda_{82} & \lambda_{92} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{10,3} & \lambda_{11,3} & \lambda_{12,3} \end{pmatrix}, \quad (7.1)$$

and the following factor correlation matrix,

$$\mathbf{\Psi} = \begin{pmatrix} 1.0 & & \\ \psi_{21} & 1.0 & \\ \psi_{31} & \psi_{32} & 1.0 \end{pmatrix} \quad (7.2)$$

Also, we specified $\mathbf{\Phi} = \mathbf{I} - \text{diag}(\mathbf{\Lambda\Psi\Lambda}')$for the identification.

For illustrative purposes, we used $N = 2,800$ complete cases and created MAR using a similar method by (Yuan, Lambert, & Fouladi, 2004). Specifically, in this research, we set the last three variables to missing if the sum of the first nine variables is greater than its sample median. This yields the data having about 44% of all observations missing. The next steps follow the same procedures as in the

simulation study: multiple imputation for missing data, combining multiple imputation, and model fitting. Using BLImP (Keller & Enders, 2017), we generated M=100 imputations. For analysis, the imputations were combined in a summary of one polychoric correlation and one correct weight matrix. Then, we first fitted the three-factor model with ULS estimation.

As a comparison, the statistic from complete cases and the statistic from MAR data but with the incorrect weight matrix ($\tilde{\mathbf{\Gamma}}$), ignoring the between-imputation variability, was obtained as well. Let us denote the latter statistic, $\tilde{T}_B$, because we simply average the polychoric correlations and the associated asymptotic covariance matrix, which does not lead to the correct weight matrix. With this additional analysis and comparison of $\tilde{T}_B$ with $\bar{T}_B$, we wish to stress the fact that the statistic without the correction should not be used. This is because it does not account for the between-imputation variance in the estimated polychoric correlations and thus may lead to erroneous conclusions. We present it here only to illustrate its biasing effect.

$\tilde{T}_B$ for the MAR data is 357.93 ($df = 51, p < .001$). The RMSEA based on $\tilde{T}_B$ can also be calculated, which is 0.052 with 90% confidence interval (CI) [0.047, 0.056]. From complete data analysis, $\tilde{T}_B$ is 398.14 ($df = 51, p < .001$), and the RMSEA is 0.053 with 90% CI [0.049, 0.058]. The corrected statistic ($\tilde{T}_B$) value is close to the statistic ($\tilde{T}_B$) computed from the complete cases. The RMSEA estimates are also close, indicating acceptably close fit. However, $\bar{T}_B$, the model fit statistic without proper correction for missing data is much larger at 493.87 ($df = 51, p < .001$), and the RMSEA is 0.061 with 90% [0.057, 0,066]. In consequence, the RMSEA and its CI give qualitatively different conclusions about model fit.

7.2 Example 2

In the second example, we used the data from PROMIS Smoking Initiative. A total of 277 smoking items were administered to a sample of daily and non-daily smokers. To minimize respondent burden, blocks of items were constructed so that each respondent was administered two blocks randomly (see Edelen, Tucker, Shadel, Stucky, & Cai, 2012; Hansen et al., 2014). This creates a natural MCAR design (missing by design). The item ratings were on a 5-point ordinal scale (e.g., not at all, a little bit, somewhat, quite a bit, very much). The researchers identified six domains: nicotine dependence, coping expectancies, emotional and sensory expectancies, health expectancies, psychosocial expectancies, and social motivations. The example items for each domain are presented in Table 7.2

Table 7.2: Examples Items for PROMIS Smoking Initiative

Construct	Example Items
Nicotine Dependence	My desire to smoke seems overpowering.
Coping Expectancies	I rely on smoking to deal with stress.
Emotional and Sensory Expectancies	I enjoy the steps I take to light up a cigarette.
Health Expectancies	Smoking is taking years off my life.
Psychosocial Expectancies	I feel embarrassed when I smoke.
Social Motivations	I enjoy the social aspect of smoking with other smokers.

Here we only used data from daily smokers ($N = 4,201$) and randomly selected 5 items from the six domains, yielding 30 items in total. The fitting model is a CFA model with 6 correlated factors for a data set with a substantial amount of missing data.

Based on $M = 200$ imputations, we computed $\tilde{T}_B$, which is 5192.11 ($df = 390, p < .001$), and the RMSEA based on $\tilde{T}_B$ is 0.050 with a 90% confidence interval of [0.049, 0.051]. The result is consistent with and improves upon prior factor

analytic and item response theory based analyses reported by Hansen et al. (2014).

CHAPTER 8

Discussion

In this research, we propose an alternative inferential multiple imputation procedure for CSEM by extending the work by Lee and Cai (2012) on multiple imputation for continuous variables. As they previously addressed, one benefit of this alternative approach is that it lessens the burden of fitting the model as many times as the number of imputations, thereby making the inferential procedure more efficient. Moreover, common fit statistics in SEM as well as intermediate results become available. This study introduces test statistics, $\tilde{T}_B$ and $\tilde{T}_{YB}$, that researchers may obtain using the new inferential multiple imputation procedure in CSEM.

The guiding insight is to average polychoric correlations computed from M imputed data sets before fitting the structural model. We can easily average the polychoric correlations. The weight matrix for least squares based parameter estimation, however, requires proper accounting of the between-imputation variance due to missing data. Thresholds can be averaged in the same manner as the correlations if they are of interest to the data analyst. Finally, applying Browne (1984) Proposition 4 leads us to obtain a new and corrected test statistic. Yuan and Bentler (1997)'s adjustment can also be used to improve the small sample performance of the new test statistic.

In our simulation studies, we demonstrate that the new test statistics, particularly Yuan and Bentler (1997)'s adjustment, are well calibrated under MCAR and MAR conditions. In addition, both tests show reasonable statistical power under

mild model misspecification. Furthermore, an empirical example illustrates our findings in the simulation studies. Additional fit indices such as RMSEA can be computed based on the proposed test statistics.

This work, of course, has limitations. The limitations of our research partly stem from the assumptions inherent in our approach. When these assumptions are violated, our new estimator may no longer be applicable. We discuss the assumptions more clearly here, and compare our estimator to the standard multiple imputation approach. First, an approach based on multiple imputation in general is restricted to certain missing data mechanisms, namely MCAR and MAR. Model-based imputation imputes missing data values from the posterior predictive distribution of the missing data given the observed data. If data are not missing at random (NMAR), our approach is only as good as the attempt to multiply impute from an inadequate imputation model. Second, our new approach assumes underlying multivariate normality. We generated data to be consistent with this assumption, and they were discretized to obtain categorical observed variables. The probit model in FCS imputation also views discrete responses as arising from latent variables that are normally distributed. Violation of this assumption could adversely influence the results. In addition, our new approach is derived from asymptotic (in sample size) results, with a number of simplifying assumptions (e.g., regularity of the structural model) and ensuing linearization arguments that lead to asymptotic normality. Finally, our estimator also relies on having a larger than usual number of imputations in order to obtain improved estimates of the average polychoric correlations and weight matrices.

Our new estimator can be more practical than FIML in certain respects. FIML is advocated by many due to ease of use through implementations in software programs, though it is by no means infallible. For instance, it is not easy to accommodate auxiliary variables in the FIML approach. In contrast, multi-stage estimation incorporates auxiliary variables easily. FIML is also more computa-

tionally demanding, especially for categorical variables. For detailed discussions of FIML with categorical data, we refer the reader to the literature on IRT parameter estimation. Forero et al. (2009) provides an overview and extensive simulation studies on the estimation of IRT models by comparing full information and limited information methods. They essentially conclude that there is no clear benefit of using FIML over the limited information multi-stage approach. The limited information methods (WLS/DWLS/ULS) are substantially faster than FIML. And FIML does not provide clear advantages in terms of parameter estimation accuracy or standard error accuracy.

Overall, the contribution of this paper is to two research areas: CSEM and multiple imputation in SEM. First, while a number of researchers have focused mainly on estimation and test statistic for CSEM (e.g., Forero et al., 2009; Maydeu-Olivares & Joe, 2014; Monroe & Cai, 2015), issues regarding missing data have not been explored solely for ordinal indicators in the SEM literature. Second, as Lee and Cai (2012) and Enders and Mansolf (2018) have pointed out, multiple imputation inference in SEM is an area that has rarely received attention despite the prevalent usage of multiple imputation in practice. Given that the necessity of multiple imputation is much greater and its advantage much more valuable for categorical data, we believe that this research not only contributes to the literature but also will meet practical needs.

Before we end our discussion, we suggest some potential directions for future studies. First, we adapted Lee and Cai (2012)'s MI2S approach to the case of dichotomous and ordered polytomous data. Though this study is limited to the discussion of ordered categorical data, one can apply the same logic further to data having mixtures of categorical and continuous variables. An example is the use of plausible values from institutionally-generated imputation procedures such as those in large-scale educational surveys. Second, the simulation study needs to be extended. Other models beyond the simple CFA model should be examined.

The distributions of the observed variables could be more varied. Third, the corrected statistics, $\tilde{T}_B$ and $\tilde{T}_{YB}$, could be compared with other test statistics and the traditional multiple imputation inferential approach in terms of performance. In addition, now that Meng and Rubin (1992)'s likelihood ratio statistic has been examined by Enders and Mansolf (2016), their statistic once applied to CSEM may serve as a comparison to $\tilde{T}_B$ and $\tilde{T}_{YB}$. Finally, Wu, Jia, and Enders (2015) showed that the FCS multiple imputation approach did not perform better than multiple imputation with the normality assumption, even when data were binary. It would be interesting to compare different multiple imputation approaches in our context.

APPENDIX A

Example Code

We provide a simple example of the alternative multiple imputation procedure using LISREL (Jöreskog & Sörbom, 2006). We use the LSAT6 data set, which is built in R (R Core Team, 2017). It consists of responses to 5 dichotomous items for 1,000 examinees. Let us assume researchers are equipped with M imputed data files after creating missing values in the LSAT6 data set.

First, M number of polychoric correlations and the asymptotic covariance matrices (ACM) are obtained using PRELIS (Jöreskog & Sörbom, 1996). The following is an example PRELIS code for $m = 1$.

```
PRELIS CODE FOR TETRACHORIC CORRELATIONS
LSAT6 DATA, 5 VARIABLES, 1000 CASES
DA NI=5 NO=1000
RA FI=LSAT6_1.DAT
OR ALL
OU PM SM=LSAT6_1.PCM AC=LSAT6_1.ACM
```

Next, we combine those polychoric correlations and the asymptotic covariance matrices (ACM) using R (R Core Team, 2017).

```
# read the PCM file
readPCM <- function(n,PCMF) {
  q <- n*(n+1)/2 # number of elements in PCMF
  fortran.PCM <- readLines(PCMF)
```

```
  PCM <- ""
for (i in 1:length(fortran.PCM)) {
    PCM <- paste(PCM,fortran.PCM[i],sep="")
  }
  PCM <- gsub("D","E",PCM)
  v <- rep(0,q)

  # read and convert to float
  for (i in 1:q) {
    startpos <- (i-1)*13+1
    endpos <- i*13
    v[i] <- as.numeric(substr(PCM,startpos,endpos))
  }
  # returns the full correlation matrix
  V <- diag(n)
  V[upper.tri(V,diag=TRUE)] <- v
  V <- V + t(V) - diag(diag(V))
  return(V)
}

# read the ASCII ACM file
readACM <- function(n,ACMF) {
  q <- n*(n-1)/2 # number of unique correlations
  s <- q*(q+1)/2 # number of unique elements in the ACM
  fortran.ACM <- readLines(ACMF)
  fortran.ACM <- fortran.ACM[2:length(fortran.ACM)]
  ACM <- ""
for (i in 1:length(fortran.ACM)) {
    ACM <- paste(ACM,fortran.ACM[i],sep="")
```

```
  }
  ACM <- gsub("D","E",ACM)
  v <- rep(0,s)
  # read and convert to float
  for (i in 1:s) {
    startpos <- (i-1)*23+1
    endpos <- i*23
    v[i] <- as.numeric(substr(ACM,startpos,endpos))
    v[i] <- v[i]/N
  }
  # returns the full covariance matrix
  V <- diag(q)
  V[upper.tri(V,diag=TRUE)] <- v
  V <- V + t(V) - diag(diag(V))
  return(V)
}

# initialize
Pcorr.sum <- 0
AC.sum <- 0
between.sum <- 0

for (m in 1:M){

PCMF <- paste("LSAT6_m=",m,".PCM",sep="")

  # read the PCM file
  Pcorr <- readPCM(n,PCMF)
```

```
  # combine the polychoric correlations
Pcorr.vechs <- Pcorr[upper.tri(Pcorr,diag=FALSE)]
Pcorr.sum <- Pcorr.sum + Pcorr.vechs

}

Pcorr.mean = Pcorr.sum/M

# save the correlation matrix in the form LISREL can read
V <- diag(n)
V[upper.tri(V,diag=FALSE)] <- Pcorr.mean
V <- V + t(V) - diag(diag(V))
PCMF.combined <- paste("LSAT6.PCM",sep="")
write.table(V[upper.tri(V,diag=TRUE)],PCMF.combined,col.names=F,row.names=F)


for (m in 1:M){

  fromF <- paste("LSAT6_m=",m,".ACM",sep="")
  toF <- paste("LSAT6_m=",m,".ACM.TXT",sep="")

  # call BIN2ASC utility to convert ACM binary file to ASCII ACM file
  writeLines(c(fromF,toF),paste("BIN2ASCcontrol.txt"))
  shell("BIN2ASC < BIN2ASCcontrol.txt")

  # read the ASCII ACM file
  AC <- readACM(n,toF)
  AC.sum <- AC.sum + AC
```

```
  # between-imputation variance
  PCMF <- paste("LSAT6_m=",m,".PCM",sep="")
  Pcorr <- readPCM(n,PCMF)
  Pcorr.vechs <- Pcorr[upper.tri(Pcorr,diag=FALSE)]
  between.sum <- between.sum + (Pcorr.vechs -Pcorr.mean)%*
%t(Pcorr.vechs - Pcorr.mean)

}

AC.mean <- AC.sum/M
between.mean <- between.sum*(M+1)/(M-1)/M
ACM.correct <- AC.mean + between.mean
ACM.correct <- N*ACM.correct
W.correct <- solve(ACM.correct) # the inverse of the ACM
w.correct <- W.correct[upper.tri(W.correct,diag=TRUE)]
                  # the unique elements from W in the order LISREL wants
WMF <- "correct_LSAT6.WM"
write.table(format(w.correct,digits=9),WMF,col.names=F,row.names=F,quote=F)
WMFlines <- readLines(WMF)
WMFlines <- c("(F15.9)",WMFlines)
writeLines(WMFlines,WMF)
```

Last, we fit a one-factor model with WLS estimation using LISREL to obtain the parameter estimates and fit indices.

```
ULS ITEM FACTOR ANALYSIS OF LSAT6 DATA
USING TETRACHORIC CORRELATIONS AND THE ACM
DA NI=5 NO=1000 MA=PM
PM FI=LSAT6.PCM
WM FI=correct_LSAT6.WM
```

```
MO NX=5 NK=1 LX=FR
OU ME=WLS PV=LSAT6.PV SV=LSAT6.SV
```

If one uses ULS or DWLS estimation, the input file (the asymptotic covariance matrix) in LISREL (Jöreskog & Sörbom, 2006) should be in binary format. Upon request, the authors can provide the ASC2BIN utility and the relevant code for reformatting the corrected weight matrix to be plugged into the ASCII2BIN.

BIBLIOGRAPHY

Allison, P. D. (2001). *Missing data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07–136. Thousand Oaks, CA: Sage.

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

Anderson, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200-203.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 697–715). Mahwah, NJ: Lawrence Erlbaum Associates.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, *15*, 651–675.

Bollen, K. A. (2010). *Structural equations with latent variables*. New York, NY: Wiley.

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in timss: A case study. *Large-Scale Assessments in Education*, *1*.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York, NY: Plenum.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cai, L., & Lee, T. (2009). Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. *Multivariate Behavioral Research*, *44*, 281?-304.

Christoffersson, A., & Gunsjö, A. (1996). A short note on the estimation of the asymptotic covariance matrix for polychoric correlations. *Psychometrika*, *61*, 171–75.

Edelen, M. O., Tucker, J. S., Shadel, W. G., Stucky, B. D., & Cai, L. (2012). Toward a more systematic assessment of smoking: development of a smoking module for PROMIS®. *Addictive Behaviors*, *37*, 1287–1284.

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

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430–57.

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, *23*, 76–93.

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*, 76–93.

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, *11*, 1–19.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 625–641.

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

Hansen, M., Cai, L., D., S. B., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS®smoking item banks. *Nicotine & Tobacco Research*, *16*, S17589.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: Users reference guide*. Chicago, IL: SSI International.

Jöreskog, K. G., & Sörbom, D. (2006). Lisrel 8.80 for windows [Computer software manual]. Lincolnwood, IL.

Keller, B. T., & Enders, C. K. (2017). Blimp users manual (Version 1.0) [Computer software manual]. Los Angeles, CA.

Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics*, *37*, 675–702.

Li, J. (2010). *Effects of full information maximum likelihood, expectation maximization, multiple imputation, and similar response pattern imputation on structural equation modeling with incomplete and multivariate nonnormal data* (Unpublished doctoral dissertation). The Ohio State University.

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

Maydeu-Olivares, A., Cai, L., & Hernndez, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 333–56.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 305–328.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*, 103-111.

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*(6), 569-583.

Olinsky, A., Chen, S., & Harlow, L. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, *151*, 53–79.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85–95.

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

Rubin, D. B. (1978). Multiple imputation in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–34.

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

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall/CRC.

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, *33*, 545–571.

Shin, T., D., M. L., & L., J. D. (2009). Effects of missing data methods in structural equation modeling with nonnormal longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 70–98.

Takahashi, Y., & Wisenbaker, J. (1999). *A comparison of the effectiveness of missing data treatments in the context of structural equation modeling.* Paper presented at the 1999 annual meeting of the American Educational Research Association, Montreal, Canada.

Tucker, L., Koopman, R., & Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459.

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

van Buuren, S. (2006). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219–242.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064.

von Hippel, P. T. (2009). Multiple imputation of discrete and continuous data by fully conditional specification. *Sociological Methodology*, *39*, 265–291.

Wang, H. (2007). *Missing data analysis in structural equation modeling: Expectation maximization and multiple imputation methods* (Unpublished doctoral dissertation). The University of Alabama.

White, I. R., Royston, P., & Wood, A. M. (2009). Multiple imputation using

chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377–399.

Wiggins, R. D., & Sacker, A. (2002). Strategies for handling missing data in sem: A users perspective. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (p. 105-120). Mahwah, NJ: Lawrence Erlbaum Publishers.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research*, *50*, 484–503.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.

Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 297–358). Amsterdam: North-Holland.

Yuan, K.-H., Lambert, P., & Fouladi, R. (2004). Mardias multivariate kurtosis with missing data. *Multivariate Behavioral Research*, *39*, 413–37.