Original Research Article

Educational and Psychological
Measurement
1–32
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644241261271
journals.sagepub.com/home/epm
Sage

Evaluating Imputation-Based Fit Statistics in Structural Equation Modeling With Ordinal Data: The MI2S Approach

Suppanut Sriutaisuk[1], Yu Liu[2], Seungwon Chung[3], Hanjoe Kim[4], and Fei Gu[1]

Abstract

The multiple imputation two-stage (MI2S) approach holds promise for evaluating the model fit of structural equation models for ordinal variables with multiply imputed data. However, previous studies only examined the performance of MI2S-based residual-based test statistics. This study extends previous research by examining the performance of two alternative test statistics: the mean-adjusted test statistic (T_M) and the mean- and variance-adjusted test statistic (T_{MV}). Our results showed that the MI2S-based T_{MV} generally outperformed other test statistics examined in a wide range of conditions. The MI2S-based root mean square error of approximation also exhibited good performance. This article demonstrates the MI2S approach with an empirical data set and provides Mplus and R code for its implementation.

Keywords

missing data, ordinal data, multiple imputation, model fit, structural equation modeling

[1] Faculty of Psychology, Chulalongkorn University, Bangkok, Thailand
[2] University of Houston, Houston, TX, USA
[3] U.S. Food and Drug Administration, Washington, DC, USA
[4] Yonsei University, Seoul, Seoul Korea

Corresponding Author:
Suppanut Sriutaisuk, Faculty of Psychology, Chulalongkorn University, Rama 1 Road, Wangmai, Pathumwan, Bangkok 10330, Thailand.
Email: suppanut.s@chula.ac.th

In structural equation modeling (SEM), a primary challenge is the handling of incomplete data of ordinal variables, a situation frequently encountered in educational and psychological sciences. Two widely used modern methods for handling missing data are full information maximum likelihood (FIML) and multiple imputation (Enders, 2023; Schafer & Graham, 2002). For researchers dealing with ordinal data, multiple imputation may be a preferable choice over FIML, partly because the distributional assumptions required by multiple imputation are considerably less restrictive (Little & Rubin, 2020; Rubin, 1987). In addition, multiple imputation facilitates the incorporation of auxiliary variables, thus making the missing at random (MAR) assumption more plausible (Collins et al., 2001). Furthermore, multiple imputation can provide commonly reported model fit statistics, such as chi-square test statistics and root mean square error of approximation (RMSEA; Browne & Cudeck, 1993). However, very limited research has been done on evaluating the model fit in SEM that utilizes multiple imputation to handle incomplete data of ordinal variables. Existing studies have predominantly focused on the standard multiple imputation approach (e.g., Asparouhov & Muthén, 2022; Liu & Sriutaisuk, 2020; Shi et al., 2020, 2023; Teman, 2012). In our paper, we focus on the multiple imputation two-stage (MI2S) approach proposed by Chung and Cai (2019; see also Lee & Cai, 2012), which has shown promise for its efficiency.

Before we introduce the MI2S approach, it is helpful to review the three phases of the standard multiple imputation approach in SEM (Enders, 2022; Rubin, 1987). The first phase is the imputation phase, where missing data are imputed m times, leading to m imputed data sets. The second phase is the analysis phase, where the hypothesized model is fitted to each of these m imputed data sets, resulting in m distinct sets of results, including parameter estimates, standard errors, and model fit statistics. Finally, the third phase is the pooling phase, where the m sets of results are combined into a single set of imputation-based results. Although the standard multiple imputation approach generally performs well under MAR (Asparouhov & Muthén, 2010b, 2022; Shi et al., 2020; Teman, 2012), it is computationally intensive due to the repeated model fitting. Moreover, it is not straightforward to derive a single set of fit statistics, such as the chi-square test statistic, from these m imputations (see Liu & Sriutaisuk, 2020; Shi et al., 2020).

The MI2S approach simplifies the process of multiple imputation in SEM by eliminating the need for repeated model fitting. Specifically, it requires fitting the hypothesized model only once to the average polychoric correlations and thresholds, along with a corrected asymptotic covariance matrix that accounts for the uncertainty from missing data. This approach is considerably more efficient than the standard multiple imputation approach, especially when dealing with numerous imputed data sets or many hypothesized models or both. Moreover, a single package of fit statistics is readily available for model evaluation. This includes the residual-based test statistics (Browne, 1984; Yuan & Bentler, 1998) recommended by Chung and Cai (2019). However, a recent study found that these test statistics may not perform well in many practical scenarios, such as when $m \leq 100$ imputed data sets (Liu et al., 2021). Consequently, the obtained results may not be accurate, highlighting the need for alternative test statistics.

To address this issue, our paper assesses the performance of two scaled test statistics from the MI2S approach: the mean-adjusted test statistic (Satorra & Bentler, 1994) and the mean- and variance-adjusted test statistic (Asparouhov & Muthén, 2010a), both of which will be discussed in detail below. In addition, this study explores the feasibility of employing the MI2S-based RMSEA to evaluate model fit.

The remaining sections are organized as follows. We begin with a brief review of SEM with ordinal data, focusing on model estimation and model fit statistics. Then, we describe the MI2S approach for ordinal data and discuss the potential of scaled test statistics within the MI2S framework. This is followed by a simulation study to assess the performance of MI2S-based fit statistics. Finally, we present an empirical example and a discussion of our findings.

SEM With Ordinal Data

Observed ordinal data can be viewed as a result of categorizing a theorized underlying continuous response variable by thresholds, and the correlation between a pair of theorized normally distributed continuous latent variables is measured by the polychoric correlation. Once the polychoric correlations, thresholds, and their asymptotic covariance matrix (which contains the information about the asymptotic distribution of parameter estimates) are estimated from the data (Olsson, 1979), the hypothesized model is fitted to these summary statistics. In this so-called limited information method (Forero & Maydeu-Olivares, 2009; Wirth & Edwards, 2007), model parameters are estimated by minimizing the least squares fit function, $F_{LS} = (\boldsymbol{r} - \hat{\boldsymbol{\rho}})' \mathbf{W}^{-1}(\mathbf{r} - \hat{\boldsymbol{\rho}})$, where $\mathbf{r}$ is a vector of nonredundant elements of the sample polychoric correlation matrix and thresholds, $\hat{\boldsymbol{\rho}}$ is the corresponding vector of nonredundant elements of the model-implied polychoric correlation matrix and thresholds, and $\mathbf{W}$ is a weight matrix, of which the specific form needs to be determined by the researcher. There are three common choices available for the matrix $\mathbf{W}$, leading to three different least squares estimators. In weighted least squares (WLS), $\mathbf{W}$ is specified as the entire asymptotic covariance matrix. In diagonally weighted least squares (DWLS), $\mathbf{W}$ is specified as a diagonal matrix whose diagonal elements are taken from the diagonal of the asymptotic covariance matrix. In unweighted least squares (ULS), $\mathbf{W}$ is specified as an identity matrix.

After parameter estimation, one way to test whether the hypothesized model fits the data is to use the conventional chi-square test statistic, calculated as the product of sample size (N) and the minimum of the fit function, $T = NF_{LS}$. The T statistic is referenced to a chi-square distribution with degrees of freedom equal to the difference between the number of nonredundant sample polychoric correlations and thresholds and the number of estimated parameters in the model. Among the three estimators, only WLS provides T that asymptotically follows a chi-square distribution with the model degrees of freedom. However, the sample size required to accurately estimate the asymptotic covariance matrix is exceedingly large. As such, WLS often leads to convergence failures, biased parameter estimates, and inflated test statistics unless the sample size is sufficiently large relative to the model size (e.g., N = 500 for a model with a total of five items; Flora & Curran, 2004; Yang-Wallentin et al., 2010). Therefore, WLS is rarely recommended in current practice.

To reduce the computational burden of WLS, DWLS/ULS avoids the inversion of the entire asymptotic covariance matrix. As a result, the chance of encountering convergence failure or improper solution, particularly when the sample size is small (Flora & Curran, 2004). However, the conventional test statistic T from DWLS/USL does not follow a chi-square distribution because the weight matrix is intentionally misspecified (i.e., $\mathbf{W}$ is not the entire asymptotic covariance matrix), although the parameter estimates remain asymptotically unbiased (Savalei, 2014).

The scaled test statistics have been developed to rescale the conventional test statistic to better approximate a chi-square distribution with the model degrees of freedom. These statistics are arguably the most popular in the context of SEM with categorical variables. A certain variant of these statistics has been implemented as the default in major SEM software packages, such as *lavaan* (Rosseel, 2012) and Mplus (Muthén & Muthén, 1998–2017).

One commonly used scaled test statistic is Satorra and Bentler's (1994) mean-adjusted test statistic which is calculated as

$$T_M = cT, \tag{1}$$

where $T = NF_{LS}$ is the conventional test statistic given a certain least squares estimator, $c = \frac{df}{tr(\mathbf{U\Gamma})}$ is a scaling factor, df is the model degrees of freedom, $\mathbf{\Gamma}$ is the asymptotic covariance matrix, $\mathbf{U} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\hat{\mathbf{\Delta}}\left(\hat{\mathbf{\Delta}}'\mathbf{W}^{-\mathbf{1}}\hat{\mathbf{\Delta}}\right)^{-1}\hat{\mathbf{\Delta}}'\mathbf{W}^{-1}$ is the residual weight matrix, $\mathbf{W}$ is the weight matrix used in the estimation, and $\hat{\mathbf{\Delta}}$ is the matrix of model derivatives. As discussed in Satorra and Bentler (1994), the mean of the asymptotic distribution of T is consistently estimated by $tr(\mathbf{U\Gamma})$. By employing this mean in the scaling factor, T is rescaled to T_M such that the expected value of T_M matches the theoretical mean of the target chi-square distribution, which is equal to df. In other words, this rescaling ensures that the mean of T_M is the same as the mean of the chi-square distribution it aims to approximate. However, T_M does not adjust the distribution of T to match higher moments, such as variance, skewness, and kurtosis.

Another commonly used scaled test statistic is Asparouhov and Muthén's (2010a) mean- and variance-adjusted test statistic which is calculated as

$$T_{MV} = aT + b, \tag{2}$$

where $a = \sqrt{\frac{df}{tr\left((\mathbf{U\Gamma})^2\right)}}$ and $b = df - atr(\mathbf{U\Gamma})$ are scaling and shifting factors, respectively, $\mathbf{U}$ and $\mathbf{\Gamma}$ are the same as in Equation (1). $tr(\mathbf{U\Gamma})$ and $2tr\left((\mathbf{U\Gamma})^2\right)$ consistently estimate the mean and the variance of the asymptotic distribution of T, respectively (Satorra & Bentler, 1994). a and b are chosen so that the expected value of T_{MV} is *df* and its variance is 2*df* (Asparouhov & Muthén, 2010a). This combination of rescaling and shifting ensures that T_{MV} closely reflects both the theoretical mean (*df*) and the theoretical variance (2*df*) of the target chi-square distribution.

Scaled test statistics can be obtained from both DWLS and ULS. Simulation studies found that the relative performance between the DWLS- and ULS-based scaled test statistics (e.g., DWLS-based T_{MV} versus ULS-based T_{MV}) was largely similar. However, ULS-based test statistics could perform slightly better under challenging conditions, such as when sample sizes were small ($N \leq 200$) or with severely asymmetric thresholds (Forero et al., 2009; C.-H. Li, 2016; Savalei & Rhemtulla, 2013; Shi et al., 2018; Yang-Wallentin et al., 2010). When comparing T_M and T_{MV}, previous studies found that T_{MV} generally produced Type I error rates close to the nominal level, particularly with symmetric thresholds and larger sample sizes, whereas T_M consistently yielded higher Type I error rates than T_{MV} (Asparouhov & Muthén, 2010a; DiStefano & Morgan, 2014). However, the difference between T_M and T_{MV} can be substantial when the model contains a large number of items. For example, one study showed that, when fitting a model with 60 items, T_{MV} was extremely conservative, whereas T_M was somewhat inflated. Nonetheless, with moderate or large sample sizes (N = 500 or 1,000) and high factor loadings (.80), the ULS-based T_M performed adequately (Shi et al., 2018). Given these findings, we focus on ULS-based T_M and T_{MV} in our study.[1]

In addition to the scaled test statistics, another way to evaluate the global model fit following DWLS/ULS is to employ Browne's (1984) residual-based test statistic:

$$T_B = N(\mathbf{r} - \hat{\boldsymbol{\rho}})^{'} \mathbf{U}_{\boldsymbol{\Gamma}} (\mathbf{r} - \hat{\boldsymbol{\rho}}), \tag{3}$$

where $\mathbf{r} - \hat{\boldsymbol{\rho}}$ is the model residuals, $\mathbf{U}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Delta}}\left(\hat{\boldsymbol{\Delta}}^{'}\boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Delta}}\right)^{-1}\hat{\boldsymbol{\Delta}}^{'}\boldsymbol{\Gamma}^{-1}$ is the residual weight matrix, $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix, and $\hat{\boldsymbol{\Delta}}$ is the matrix of model derivatives. T_B is based on the distribution of residuals, which has an asymptotic chi-square distribution even when DWLS/ULS is used. Unfortunately, T_B tends to be inflated in finite-sample sizes, primarily because its calculation involves the inversion of the entire asymptotic covariance matrix, which can be inaccurate when the sample size is not sufficiently large. To avoid the inflated value of T_B, Yuan and Bentler (1998) proposed a corrected version of the residual-based test statistic[2]:

$$T_{YB} = T_B / [1 + N T_B / (N - 1)^2]. \tag{4}$$

This correction shrinks T_B as N decreases. Previous simulations showed that the performance of T_{YB} was generally good, but it could be lower than the nominal value with small sample sizes (Yuan & Bentler, 1998). Therefore, T_{YB} is often preferred over T_B, unless a very large sample size is available, where T_B and T_{YB} are equivalent. In our study, we include both T_B and T_{YB} because both have been examined in previous studies using the MI2S approach (e.g., Chung & Cai, 2019).

All the chi-square test statistics (i.e., T_M, T_{MV}, T_B, and T_{YB}) reviewed above test the hypothesis that the population and model-implied population moment structure are exactly equal. In practice, a chi-square test statistic is always significant for models with large sample sizes. As such, additional measures of fit are often provided (Jackson et al., 2009). In this study, we also include the RMSEA because it is not only widely reported (e.g., 73% of empirical studies; Zyphur et al., 2023) but also has a variant version specifically designed for SEM with ordinal variables (Lai, 2020). Xia and Yang (2019) demonstrated that employing the traditional RMSEA, developed for continuous data, in a least-squares estimation with ordinal variables tended to result in a lower RMSEA value, suggesting a better fit between the model and data than there actually was.

The RMSEA quantifies the amount of discrepancy between the population and the hypothesized model as the average amount of discrepancy per constrained parameter. The RMSEA is defined as follows:

$$RMSEA = \sqrt{\frac{F}{df}}, \tag{5}$$

where F is the fit function and df is the model degrees of freedom. A bias-corrected sample estimate for F under ULS is $\hat{F}_{BC} = \hat{\epsilon}'\hat{\epsilon} - \frac{1}{2n}tr(\mathbf{Q}\boldsymbol{\Gamma})$, where $\hat{\epsilon}$ is a vector of residuals, $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix, and $\mathbf{Q}$ is a matrix of model second derivatives. More technical details can be found in Lai (2020).

In summary, DWLS/ULS is frequently employed when conducting SEM with ordinal data. However, the conventional chi-square test statistic does not follow a chi-square distribution under DWLS/ULS. As a result, the scaled test statistics and the residual-based test statistics are preferred. In addition, approximate fit indices such as the RMSEA are also recommended to determine to what extent the hypothesized model fits the data. Yet, the presence of missing data, particularly in the context of ordinal data, adds additional complexity to this process. Next, we discuss how these fit statistics are derived within the MI2S approach.

Two-Stage Multiple Imputation

Chung and Cai (2019) proposed the MI2S approach as an innovative inferential procedure for SEM with ordinal variables based on multiple imputation. This approach differs from the standard multiple imputation procedure in the sense that it does not require the pooling of m sets of fit statistics, such as m chi-square test statistics, from repeated fitting of the hypothesized model.

The MI2S approach consists of two stages. In the first stage, m imputed data sets are created from a single data set with missing data. Then, m sets of $\mathbf{r}$ (a vector of the nonredundant elements of the sample polychoric correlation matrix and thresholds) and $\boldsymbol{\Gamma}$ (the asymptotic covariance matrix) are calculated from m imputed data sets. These m sets of $\mathbf{r}$ and $\boldsymbol{\Gamma}$ are then combined using Rubin's (1987) rules. Specifically, a vector of pooled polychoric correlations and thresholds is calculated as follows:

$$\bar{\mathbf{r}} = \frac{1}{m}\sum\nolimits_{i=1}^{m} \mathbf{r}_i. \tag{6}$$

As can be seen, $\bar{\mathbf{r}}$ is just the average of the m vectors of $\mathbf{r}$. Next, the total asymptotic covariance matrix, which contains the estimates of the asymptotic covariance matrix that have been corrected for the additional uncertainty due to missing data, is calculated as follows:

$$\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_W + \boldsymbol{\Gamma}_B + \frac{\boldsymbol{\Gamma}_B}{m}, \tag{7}$$

where $\boldsymbol{\Gamma}_W = \frac{1}{m}\sum\nolimits_{i=1}^{m} \boldsymbol{\Gamma}_i$ is the within-imputation asymptotic covariance matrix, which is simply the average over m sets of the asymptotic covariance matrix, and $\boldsymbol{\Gamma}_B = \frac{1}{m-1}\sum\nolimits_{i=1}^{m}(\mathbf{r}_i - \bar{\mathbf{r}})(\mathbf{r}_i - \bar{\mathbf{r}})'$ is the between-imputation asymptotic covariance matrix, which represents the additional variability due to the missing data.

In the second stage, $\bar{\mathbf{r}}$ and $\tilde{\boldsymbol{\Gamma}}$, which are consistent estimates of $\mathbf{r}$ and $\boldsymbol{\Gamma}$, serve as the input for the least squares fit function. That is, the hypothesized model is fitted to the pooled polychoric correlations and thresholds using DWLS/ULS. As the hypothesized model is fitted only once, a single set of results is readily available. This set includes the MI2S versions of T_M, T_{MV}, T_B, and T_{YB}, denoted as $\tilde{T}_M$, $\tilde{T}_{MV}$, $\tilde{T}_B$, and $\tilde{T}_{YB}$, respectively. The calculations of the MI2S-based test statistics mirror that of their complete data counterparts given in Equations 1 to 4. Similarly, approximate fit indices can be obtained in a similar fashion.

Previous studies have examined the performance of $\tilde{T}_B$ and $\tilde{T}_{YB}$. Chung and Cai (2019) found that, with a relatively small model size (a three-factor model with nine items) and a sufficient number of imputed data sets (m = 60 for high missing data rates), Type I error rates of $\tilde{T}_B$ and $\tilde{T}_{YB}$ were close to the nominal level when the sample sizes were large ($N \geq$ 1,000), regardless of the number of categories, missing data rate (up to 40% on a third of the analysis variables), and missing data mechanisms (missing completely at random [MCAR] or MAR). With smaller sample sizes ($N \leq 500$), $\tilde{T}_B$ was often too high, while $\tilde{T}_{YB}$ generally performed well (Chung & Cai, 2019). However, with a larger model size (15 items), Liu et al. (2021) found that $\tilde{T}_B$ was highly inflated even when the sample size was large (N = 1,000), while the performance of $\tilde{T}_{YB}$ did not always improve with an increased sample size. In addition, both $\tilde{T}_B$ and $\tilde{T}_{YB}$ were clearly more inflated with m = 100 than m = 300 or higher in many conditions, suggesting that a large number of imputed data sets is required for residual-based test statistics (Liu et al., 2021).

The likely cause of the suboptimal performance of $\tilde{T}_B$ and $\tilde{T}_{YB}$ in some contexts can be partially attributed to the known instability of $\tilde{\boldsymbol{\Gamma}}$ (Enders, 2022). Moreover, even without multiple imputation, these residual-based test statistics involve the inversion of $\boldsymbol{\Gamma}$ (see Equation 3), which may encounter some numeric challenges similar to those when utilizing the WLS estimator (Wirth & Edwards, 2007). Particularly, for small sample sizes, $\boldsymbol{\Gamma}$ exhibits considerable sampling variation, and its inversion could become infeasible (Browne, 1984). In this study, we focus on the performance of $\tilde{T}_M$ and $\tilde{T}_{MV}$. These scaled test statistics could offer a viable alternative as they do not require the inversion of $\tilde{\mathbf{\Gamma}}$. When calculating $\tilde{T}_M$ and $\tilde{T}_{MV}$, the information in $\tilde{\mathbf{\Gamma}}$ is condensed into one or two values (as shown in Equations 1 and 2). These values are then used to rescale the MI2S-based conventional test statistic.

In summary, several scaled test statistics and residual-based test statistics can be used as tests of model fit when the MI2S approach is applied. However, when a large number of imputed data sets is not feasible, the residual-based statistics may not be trustworthy. Although the scaled test statistics are widely reported and show promise, their performance within the MI2S framework has yet to be examined. To fill this gap, the primary objective of this study is to evaluate the performance of $\tilde{T}_M$ and $\tilde{T}_{MV}$. In addition, since previous studies have not evaluated the performance of any MI2S-based approximate fit indices, this study also investigates the performance of the MI2S-based RMSEA.

Simulation Study

To examine the performance of MI2S-based fit statistics, we conducted a simulation study using a 2 (number of response categories: $C = 2, 5$) × 2 (threshold distribution: symmetric thresholds, asymmetric thresholds) × 3 (sample size: $N = 250, 500, 1{,}000$) × 2 (missing data mechanism: MAR1, MAR2) × 2 (missing data rate: 20%, 40%) × 4 (number of imputed data sets: $m = 20, 50, 100, 300$) × 4 (analysis model: one correct model and three incorrect models) full factorial design.

For each combination of the first four factors, 1,000 complete data sets were generated.[3] Then, two incomplete data sets with different missing data rates were generated from each complete data set. For each incomplete data set, we generated a total of $m = 300$ imputed data sets. The imputed data sets were then used in four different sets of analyses that were based on $m = 20, 50, 100$, and all 300 imputed data sets. Each set of analyses consisted of four analysis models. Complete data were also analyzed.

We used Mplus 8.6 for multiple imputation and the *lavaan* package in R for data analysis. Computer scripts used for our simulation are available on the Open Science Framework (OSF) at https://osf.io/64yjv/?view_only=1edc2972045243f09a0a323c1774c335.

Data Generation

The data-generating model was a three-factor confirmatory factor analysis (CFA) model with six items per factor (X_1–X_6, M_1–M_6, and Y_1–Y_6). All factor loadings and error variances were set to .80 and .36, respectively (i.e., the variances of all latent continuous variables underlying the ordinal observed items were 1). The correlations between the three factors were set to .40 (medium-high effect size). These values are comparable with Chung and Cai (2019), except that the number of items was doubled so that the impact of the number of imputed data sets and other factors can be clearly observed. A review of 194 studies using CFA found that the median numbers of observed items and latent factors were 17 and three, respectively (Jackson et al., 2009). Thus, the model we used (three factors with 18 items) was common and could be considered a medium size.

To simulate the data, we first generated the underlying continuous response variables from a multivariate normal distribution. Then, continuous variables were categorized by thresholds (described below), resulting in observed ordinal items.

MAR Data Generation. The missing data were generated using two MAR mechanisms: MAR1 and MAR2.[4] Supplemental Table S1 in the online supplemental materials summarizes key similarities and differences between the two MAR mechanisms.

MAR1. Missing values were imposed on every item in the last factor (Y_1–Y_6), that is, 33.3% of items were incomplete. A case with incomplete data had no data on all six Y items. The missingness in Y_1–Y_6 was determined by the sum of the first 12 items (X and M items). Specifically, we sorted the data set in ascending order by the sum scores and deleted the first 20% or 40% of observations on Y_1–Y_6, meaning that missing data were perfectly determined by the sum scores, with low scores corresponding to missing data. This MAR data generation is similar to several studies investigating SEM with incomplete ordinal data (e.g., Shi et al., 2020), resulting in a monotone missing data pattern. This pattern is common in longitudinal studies (e.g., drop-out scenarios).

MAR2. Similar to MAR1, six items contain missing values, but these six items were not in the same factor. In MAR2, missing values were imposed on M_1–M_3 and Y_1–Y_3. An auxiliary variable, drawn from a standard normal distribution (M = 0, SD = 1) and correlated with M_1–M_3 and Y_1–Y_3 before categorization at r = .40, was the cause of missingness. The missingness was determined according to a logistic regression model with McKelvey and Zavoina's (1975) pseudo-R^2 value of .40, with parameters $\beta_0 = -1.91$ and $\beta_1 = 1.50$ for a 20% missing data rate, and $\beta_0 = -0.57$ and $\beta_1 = 1.50$ for a 40% missing data rate. Given these parameters, higher values of the auxiliary variable were associated with higher probability of having missing data. Unlike MAR1, the cause of missingness was not perfectly related to the occurrence of missing data. This resulted in a general missing data pattern where missing values occurred in a seemingly random manner. This MAR data generation resembles what has been used in previous simulation studies to examine the performance of test statistics in SEM with incomplete ordinal data (e.g., Liu & Sriutaisuk, 2020).

Multiple Imputation. We used multiple imputation based on an unrestricted variance-covariance model implemented in Mplus (Asparouhov & Muthén, 2022). All items were included in the imputation model. Specifically, for MAR1, the model included X_1–X_6, M_1–M_6, and Y_1–Y_6. For MAR2, the model additionally included the auxiliary variable along with X, M and Y items. By including the cause of missingness (X and M items for MAR1 and the auxiliary variable for MAR2) in the imputation model, multiple imputation is unbiased under MAR (Enders, 2022). To evaluate the convergence of MCMC estimation, we utilized the highest potential scale reduction (PSR; Gelman et al., 2013). The burn-in iterations ranged from 2,000 to 38,000 across missing data generation conditions (see Supplemental Table S2).

Number of Response Categories and Threshold Distribution

The same set of thresholds was used to discretize all underlying continuous variables, resulting in either binary (C = 2) or polytomous (C = 5) items with symmetric or (severely) asymmetric thresholds. The thresholds were chosen so that the percentages of cases in each category of each item before imposing missing data would be 50% and 50% (symmetric) or 85% and 15% (asymmetric) when C = 2; and 7%, 24%, 38%, 24%, and 7% (symmetric) or 52%, 15%, 13%, 11%, and 9% (asymmetric) when C = 5 (see Supplemental Table S3 for threshold values). These threshold values are the same as those used in previous studies (e.g., Rhemtulla et al., 2012). We chose two and five categories because both are popular (e.g., responses coded as 0 or 1, 5-point Likert-type scales). Moreover, the effect of discretization is greatest with few categories, and items with more than five categories can sometimes be treated as continuous (Rhemtulla et al., 2012). It is noteworthy that MAR1 mitigated the extremity of the asymmetry, whereas MAR2 exacerbated it. Supplemental Figures S1 and S2 graphically show the distributions of an item with and without missing data by the number of response categories and threshold distribution under MAR1 and MAR2, respectively.

Sample Size

Three sample sizes were investigated: N = 250, 500, and 1,000, representing relatively small, medium, and large samples. We did not consider smaller sample sizes despite their high prevalence because they often yielded high convergence failures and biased parameter estimates, particularly with binary data and asymmetric thresholds, even with complete data (e.g., Forero & Maydeu-Olivares, 2009; Forero et al., 2009; Shi et al., 2018). With missing data, one study had to discard results in N = 150 conditions due to severe convergence problems (Jia & Wu, 2019).

Missing Data Rate

Similar to Chung and Cai (2019), we examined two missing data rates: low (20%) and high (40%).

Number of Imputed Data Sets

Simulation studies on missing data in SEM with ordinal variables often set m between 20 and 100 imputed data sets (e.g., Chung & Cai, 2019; Jia & Wu, 2019; Shi et al., 2020). We generated m = 300 imputed data sets for each incomplete data set. To gain a better understanding of the effect of the number of imputed data sets on test statistics, we examined results based on 20, 50, 100, and all 300 imputed data sets, representing a small, moderate, large, and extremely large number of imputed data sets.

Analysis Model

We examined one correctly specified model and three incorrectly specified models. The correct model was the same as the data-generating model; that is, a three-factor CFA (X, M, and Y) with six items per factor ($df = 132$). The first incorrect model (ICM1) had the same number of factors and items, but a full latent mediation model, where the direct effect was fixed at 0, was specified ($df = 133$). We chose this model to examine the power to detect a relatively small model misspecification with the population RMSEA of 0.065. The second incorrect model (ICM2) was a three-factor CFA model with 18 items, but X_6 was erroneously loaded on the Y factor, rather than the X factor ($df = 132$). Finally, the third incorrect model (ICM3) was a two-factor CFA model, with one factor measured by X_1 to X_6 and the other by M_1–M_6 and Y_1–Y_6 ($df = 134$). ICM2 and ICM3 had relatively large and very large model misspecifications, with the population RMSEA values of 0.095 and 0.134, respectively.

Model Fit Statistic

We obtained $\tilde{T}_M$, $\tilde{T}_{MV}$, and $\tilde{T}_B$, as well as their complete data counterparts (T_M, T_{MV}, and T_B) from *lavaan*. $\tilde{T}_{YB}$ and T_{YB} were manually computed. We used Lai's (2020) RMSEA instead of the traditional RMSEA. The complete data RMSEA and MI2S-based RMSEA were calculated using a custom R script, provided by Lai (2020). As Lai's (2020) RMSEA was developed for ULS, all analysis models were fitted using the ULS estimator.

Simulation Results

We summarize the results with respect to five outcomes: (a) convergence failures and improper solutions; (b) relative bias in parameter estimates; (c) Type I error rates; (d) statistical power; and (e) means and coverage rates for the RMSEA. Although our focus is on test statistics, we briefly present relative bias in parameter estimates for the correct model to show that MI2S yields reasonable estimates under MAR. Additional results, such as the distributions of the test statistics in terms of means and variances, as well as those results omitted in the main text, can be found in the online supplemental materials.

Convergence Failures and Improper Solutions

There was no sign of a severe convergence problem. All MCMC chains for multiple imputation converged, with the highest PSR of 1.05, which was below the widely used cutoff of PSR < 1.10 (Gelman et al., 2013). However, around 2% of imputed data sets under MAR1 (i.e., monotone missing data pattern) and less than 1% under MAR2 (i.e., general missing data pattern) with $C = 5$, $N = 250$, symmetric thresholds, and 40% missing data were not included in our analyses. The reason to eliminate these imputed data sets was that they contained item(s) with fewer observed response categories (four rather than five categories) than specified in the data generation model, resulting in a different model *df* and different theoretical distributions of the test statistics compared with other replications in the same condition.

For analyses of imputed data, at least 97% of the replications converged to admissible solutions across most conditions (see Table 1). However, in conditions with $C = 2$, $N = 250$, and asymmetric thresholds, the rate of proper solutions dropped to 95% under MAR1 with 40% missing data, 94% under MAR2 with 20% missing data, and 85% under MAR2 with 40% missing data. Conditions with fewer response categories, asymmetric thresholds, smaller sample sizes, higher rates of missing data, and MAR2 mechanism generally posed more challenges for convergence. All improper solutions were attributed to a non-positive definite covariance matrix (e.g., negative estimated residual variances) and were thus excluded.

Relative Bias in Parameter Estimates

The relative bias for a specific parameter estimate was calculated as $\mathrm{RB} = \frac{\hat{\theta}-\theta}{\theta}$, where $\hat{\theta}$ is the parameter estimate from a given replication, and θ is the population value. The absolute value of the mean RB less than 10% was considered acceptable (Flora & Curran, 2004). For each combination of the number of response categories, threshold distribution, sample size, and missing data rate, we examined the mean RB across replications for the correctly specified model using $m = 300$ imputed data sets. Given the minimal differences in RB between the two MAR mechanisms, the results were reported collectively. Table 1 shows the highest mean RB for factor loadings, covariances, and thresholds.

First, factor loadings were acceptable across conditions with $m = 300$, with the highest mean RB being -4.0%. Second, factor covariances were also acceptable, showing the highest mean RB of 9.8%. However, threshold values were only partially acceptable, with the highest mean RB of -36%. The excessively high mean RB values originated from conditions with $C = 5$ and asymmetric thresholds. In these conditions, one population threshold value was near zero (i.e., 0.05; see Supplemental Table S3), meaning that even a slight deviation could result in a substantial relative bias. Upon excluding this near-zero population threshold value, all threshold values were acceptable, with the highest mean RB of -5.1%.[5]

Empirical Type I Error Rates for the Correct Model

We defined the empirical Type I error rate of a test statistic in a certain condition as the proportion of non-excluded replications in that condition that produced a

Table 1. Highest Mean Relative Biases at m = 300.

				#Rep		Highest mean relative bias			
C	TD	*N*	Miss	MAR1	MAR2	Loading	Covariance	Threshold	Threshold[a]
2	sym	250	20	994	1000	–.007	.029	—	—
			40	980	990	–.011	.053	—	—
		500	20	1000	1000	–.004	.012	—	—
			40	999	1000	.007	.033	—	—
		1000	20	1000	1000	–.003	.006	—	—
			40	1000	1000	.005	.008	—	—
	asym	250	20	972	936	–.017	–.035	.016	.016
			40	954	852	–.040	.098	.038	.038
		500	20	1000	1000	.008	.013	–.010	–.010
			40	999	979	.009	.046	–.019	–.019
		1000	20	1000	1000	.014	.010	–.008	–.008
			40	1000	998	.019	.027	–.015	–.015
5	sym	250	20	1000	1000	–.006	.009	–.020	–.020
			40	977	996	–.008	–.017	–.038	–.038
		500	20	1000	1000	–.003	.006	.013	.013
			40	1000	1000	–.006	.006	–.021	–.021
		1000	20	1000	1000	–.002	–.005	.009	.009
			40	1000	1000	–.004	.003	–.014	–.014
	asym	250	20	1000	1000	–.008	.013	**.170**	–.018
			40	999	1000	–.014	–.041	–**.360**	–.051
		500	20	1000	1000	.007	.010	**.133**	–.017
			40	1000	1000	.013	–.023	–**.323**	–.045
		1000	20	1000	1000	.006	.005	**.100**	–.011
			40	1000	1000	.013	–.008	**.143**	–.023

Note. Relative biases between −.100 and .100 are bolded. Relative bias is not available when the population threshold equals zero (conditions with *C* = 2 and symmetric thresholds). *C* = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; *N* = sample size; Miss = missing data rate; #Rep = number of useable replications (up to 1,000).

[a]Highest relative bias among thresholds, excluding the near-zero (i.e., 0.05) population threshold value.

significant ($p < .05$) test statistic when the model was correctly specified. The values within the range of [.025, .075] were considered acceptable (Bradley, 1978).

$\tilde{T}_M$ and $\tilde{T}_B$. Figures 1 and 2 show the Type I error rates of $\tilde{T}_M$ and $\tilde{T}_B$ under MAR1 and MAR2, respectively. As illustrated in the figures, the Type I error rates for $\tilde{T}_M$ and $\tilde{T}_B$, along with their complete data counterparts, were inflated across all conditions, regardless of the number of response categories, threshold distribution, sample size, missing data mechanism, missing data rate, and number of imputed data sets. For $\tilde{T}_B$, there was a pronounced trend of inflated Type I error rates when m was insufficient, particularly with a high proportion of missing data, resulting in Type I error rates that considerably exceeded the nominal level of .05. This inflation was reduced as m increased, showing marked improvement when m reached 300. Similarly, the impact of N was noticeable, with larger sample sizes resulting in lower rejection rates

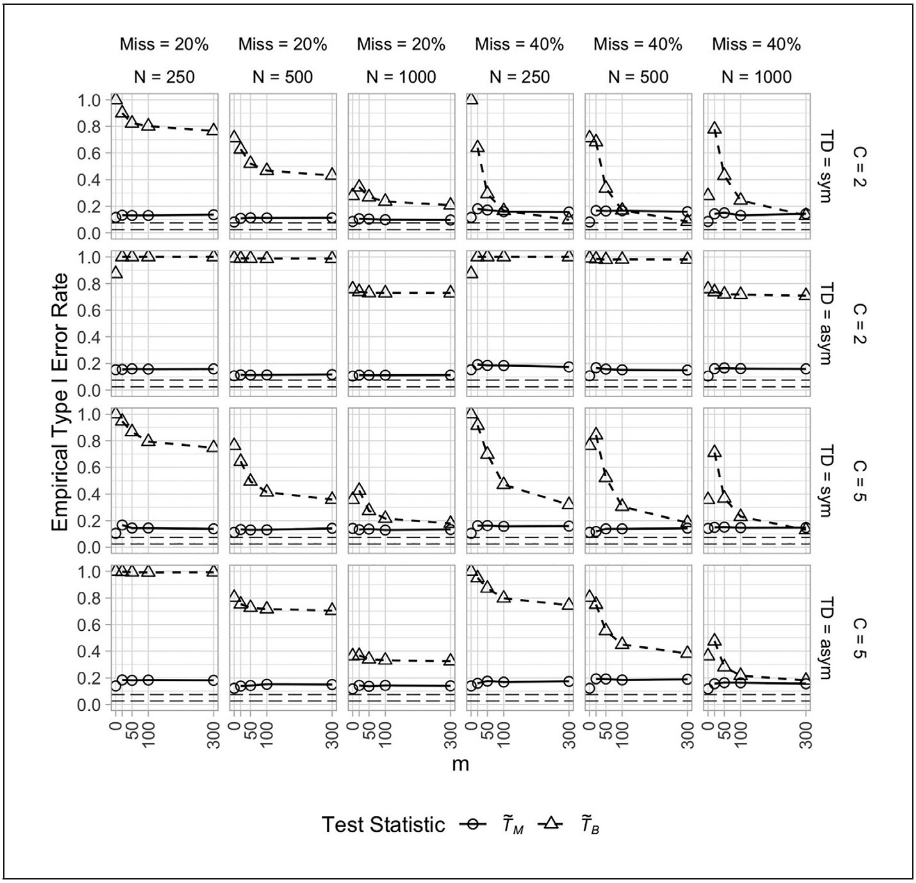

Figure 1. Empirical Type I Error Rates of $\tilde{T}_M$ and $\tilde{T}_B$ Under MAR1.

Note. The results from corresponding complete data analyses are displayed at m = 0. The acceptable range for the Type I error rate [.025, .075] is between the long-dashed lines. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; m = number of imputed data sets.

for $\tilde{T}_B$, approaching the expected nominal level. In contrast, $\tilde{T}_M$ exhibited a more stable pattern across m and other factors, yielding lower and more consistent rejection rates. Nevertheless, both $\tilde{T}_M$ and $\tilde{T}_B$ showed inflated Type I error rates.

***$\tilde{T}_{MV}$* and *$\tilde{T}_{YB}$*.** Figures 3 and 4 show the Type I error rates of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ under MAR1 and MAR2, respectively. Overall, out of 192 Type I error rates for $\tilde{T}_{MV}$ presented in the two figures, 90% were within the acceptable range. As shown in Figure 3, the Type I error rates for $\tilde{T}_{MV}$ were well-calibrated to those obtained from complete data, staying reasonably close to the nominal level, across all conditions under the MAR1 mechanism, with only one exception (C = 5, N = 500, symmetric

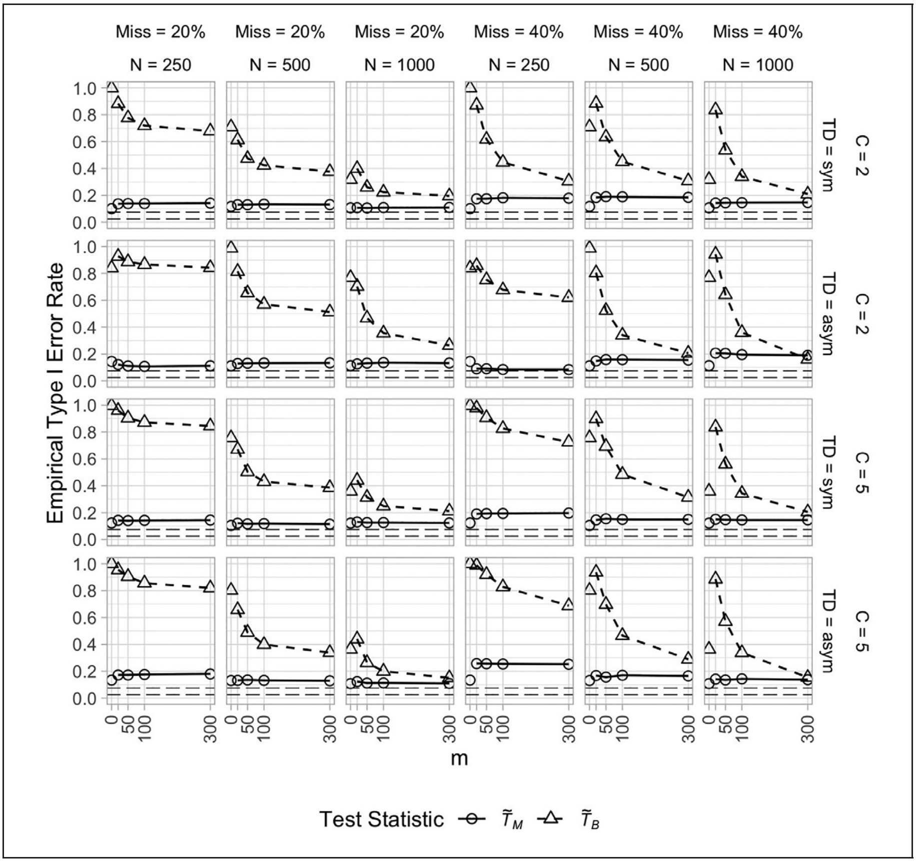

Figure 2. Empirical Type I Error Rates of $\tilde{T}_M$ and $\tilde{T}_B$ Under MAR2.

Note. The results from corresponding complete data analyses are displayed at m = 0. The acceptable range for the Type I error rate [.025, .075] is between the long-dashed lines. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; m = number of imputed data sets.

thresholds, 40% missing data, and m = 20) that yielded a slightly deflated Type I error rate. Similarly, as depicted in Figure 4, the Type I error rates for $\tilde{T}_{MV}$ under MAR2 closely matched those from complete data and were near the nominal level when the missing data rate was 20%. However, some conditions with 40% missing data under MAR2 exhibited inflated Type I error rates, particularly most conditions with N = 250, as well as those with C = 2 and symmetric thresholds at N = 500. In addition, the C = 2 conditions with N = 1,000 and asymmetric thresholds also yielded inflated Type I error rates with 40% missing data under MAR2. These results generally suggested that $\tilde{T}_{MV}$ performed well in a wide range of conditions, but could be suboptimal even when the number of imputed data sets appeared sufficient,

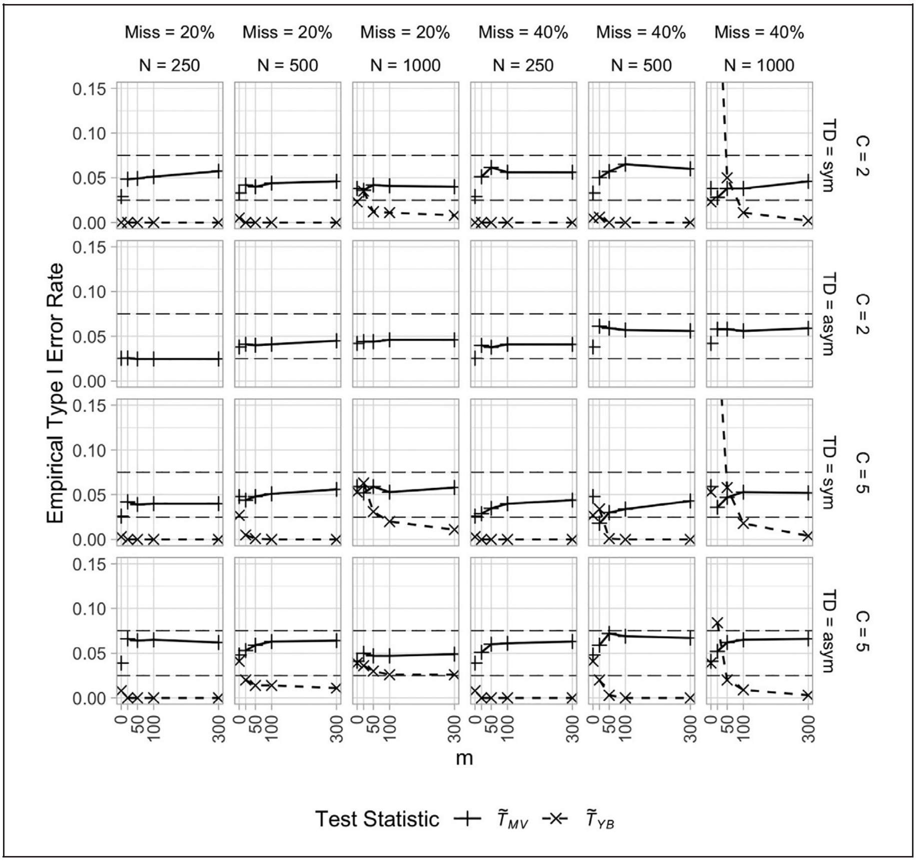

Figure 3. Empirical Type I Error Rates of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ Under MAR1.

Note. The results from corresponding complete data analyses are displayed at m = 0. The y-axis is truncated to provide better visualization of the differences between the test statistics. As a result, highly inflated test statistics are omitted, such as many of those in the C = 2 and asymmetric threshold conditions. The acceptable range for the Type I error rate [.025, .075] is between the long-dashed lines. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; m = number of imputed data sets.

especially at high missing data rates in certain combinations of sample size, threshold distribution, and missing data mechanism.

Regarding $\tilde{T}_{YB}$, only 12.5% of its Type I error rates fell within the acceptable range, with the majority (67%) yielding deflated Type I error rates. The rejection rates of $\tilde{T}_{YB}$ varied across and significantly influenced by the number of imputed data sets; they tended to become more conservative as m increased, especially in conditions with larger sample sizes, higher rates of missing data, and under the MAR2 mechanism. Moreover, due to the small sample adjustment, when $N = 250$, $\tilde{T}_{YB}$ consistently led to Type I error deflation, except in some conditions with C = 2 and

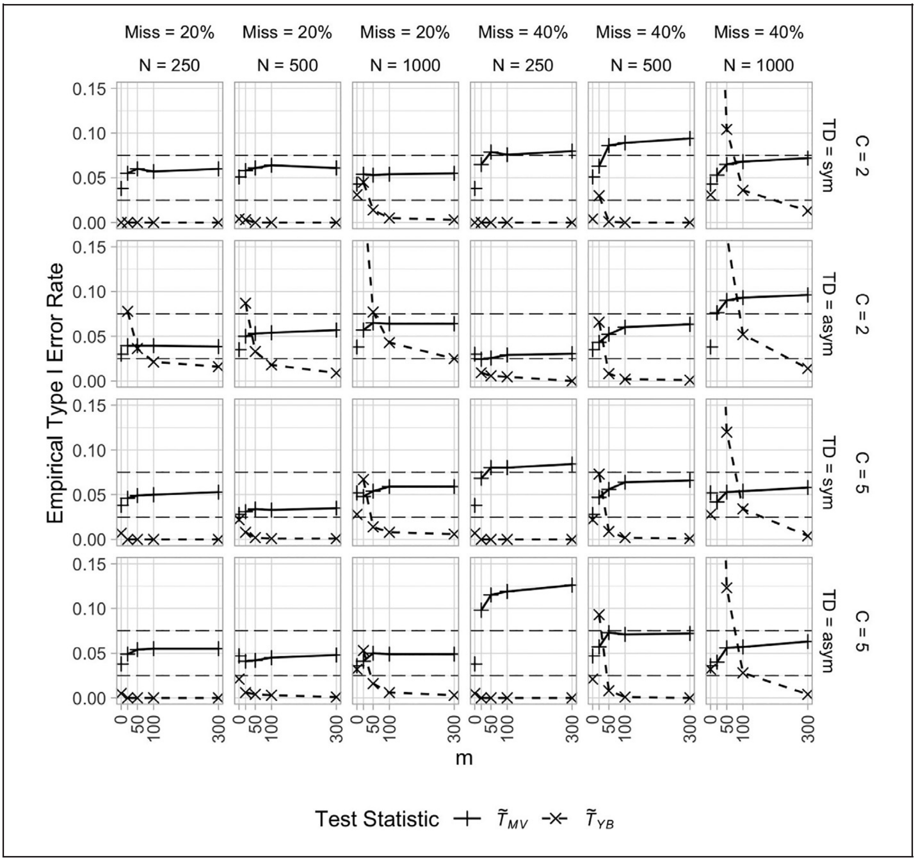

Figure 4. Empirical Type I Error Rates of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ Under MAR2.

Note. The results from corresponding complete data analyses are displayed at m = 0. The y-axis is truncated to provide better visualization of the differences between the test statistics. As a result, highly inflated test statistics are omitted, such as many of those in the C = 2 and asymmetric threshold conditions. The acceptable range for the Type I error rate [.025, .075] is between the long-dashed lines. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; m = number of imputed data sets.

asymmetric thresholds. Notably, its complete data counterpart, T_{YB}, also tended to perform poorly in most conditions, especially with smaller sample sizes. When the data were binary and asymmetric, the Type I error rates of T_{YB} were substantially inflated. Our results generally showed that $\tilde{T}_{YB}$, as well as its complete data counterpart, did not perform well in terms of Type I error control in most of the conditions examined.

In summary, our results revealed distinct patterns in the control of Type I error rates by the MI2S-based test statistics. $\tilde{T}_{MV}$ yielded decent performance, maintaining acceptable error rates in most conditions. Conversely, $\tilde{T}_{M}$ demonstrated a tendency to slightly inflate Type I error rates across all conditions. $\tilde{T}_B$ and $\tilde{T}_{YB}$ revealed more pronounced limitations: $\tilde{T}_B$ was prone to severe error rate inflation while $\tilde{T}_{YB}$ often resulted in overly conservative Type I error rates, especially in cases of smaller sample sizes. Furthermore, the performance of $\tilde{T}_B$ and $\tilde{T}_{YB}$ varied across m, suggesting that a large number of m is required.

Statistical Power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ for the Incorrect Models

The statistical power of a test statistic in a condition was defined as the proportion of non-excluded replications in that condition that produced a significant test statistic ($p < .05$) when fitting a misspecified model. Since the Type I error rates of $\tilde{T}_M$ and $\tilde{T}_B$ were inflated, resulting in artificially high power across all conditions, we chose to focus only on the power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$. For the same reason, we omitted the power of a certain test statistic when its corresponding Type I error rate in the m = 300 condition or in the complete data analysis was found to be inflated. For example, the Type I error rates of T_{YB} from complete data analyses were highly inflated when C = 2 and thresholds were asymmetric. Consequently, we should not use either T_{YB} or $\tilde{T}_{YB}$ in these conditions. Although T_{MV} from complete data analyses yielded acceptable Type I error rates across all conditions, $\tilde{T}_{MV}$ showed inflated Type I error rates in some conditions under MAR2. As such, the power of the $\tilde{T}_{MV}$ in those conditions was omitted to avoid misinterpretation that might arise from the artificially high power due to inflated Type I error rates.

Incorrect Model 1. Figures 5 and 6 summarize the statistical power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ for ICM1 under MAR1 and MAR2, respectively. As shown in the figures, the power of $\tilde{T}_{MV}$ remained relatively stable across m. In contrast, the power of $\tilde{T}_{YB}$ tended to decrease as m increased, especially in conditions with larger sample sizes and a high missing data rate. Our results showed that $\tilde{T}_{YB}$ consistently exhibited lower power compared with $\tilde{T}_{MV}$. Specifically, with N = 250, $\tilde{T}_{MV}$ had power above .20, whereas the power of $\tilde{T}_{YB}$ was virtually nil. With N = 500, the power of $\tilde{T}_{MV}$ exceeded .50, while $\tilde{T}_{YB}$ still rarely rejected ICM1, especially with larger m. Finally, with N = 1,000, $\tilde{T}_{MV}$ often had essentially 100% power, whereas $\tilde{T}_{YB}$ only reached power in the range of around .20 to .85 when m = 300, which is on par with or even lower than the power of $\tilde{T}_{MV}$ when N = 250. Notably, in conditions where $\tilde{T}_{MV}$ yielded inflated Type I error rates (e.g., MAR2, 40% missing data rate, N = 250, and C = 5; see Figure 4), the power of $\tilde{T}_{YB}$ in these conditions was non-existent when m = 300.

Incorrect Models 2 and 3. The varied pattern across m for ICM2 and ICM3 was the same as for ICM1. As such, we focused only on m = 300 because, theoretically, this should produce the most reliable conclusions. The results for N = 1,000 were not presented, as all MI2S-based test statistics could consistently reject ICM2 and ICM3 with this large sample size. As shown in Table 2, the power of $\tilde{T}_{MV}$ for both ICM2 and ICM3 was virtually 100% across conditions, except for those

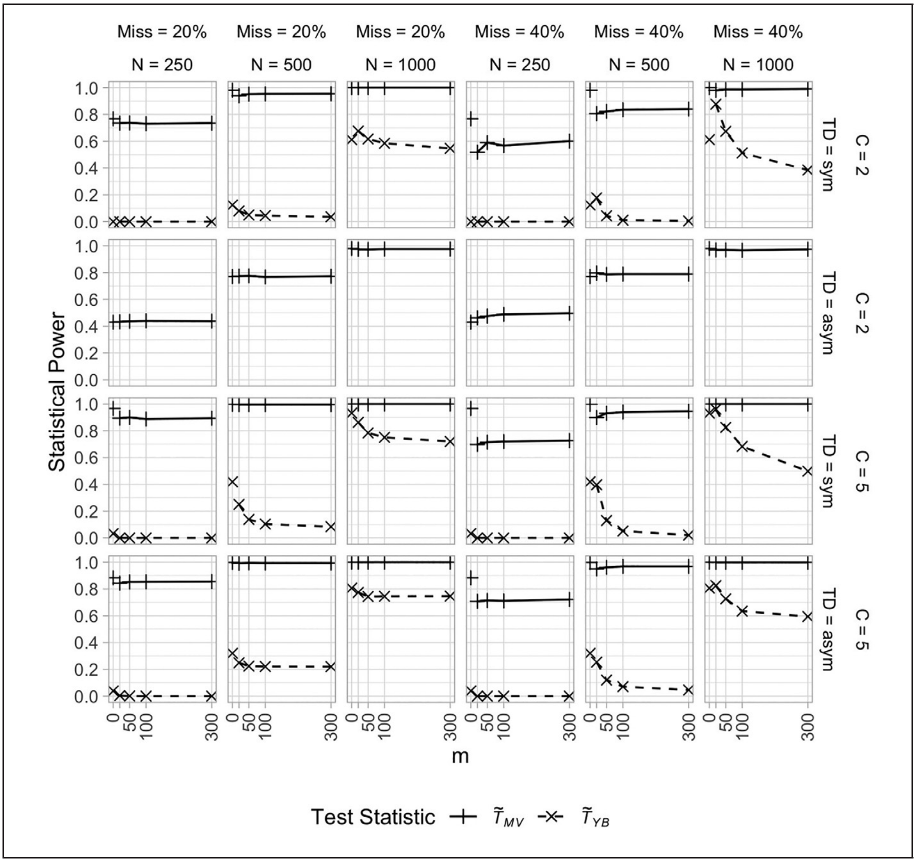

Figure 5. Empirical Power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ for Incorrect Model 1 Under MAR1.
Note. Power is omitted when the corresponding Type I error rate of *m* = 300 or complete data analysis is inflated. The results from corresponding complete data analyses are displayed at *m* = 0. *C* = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; *N* = sample size; Miss = missing data rate; *m* = number of imputed data sets.

with $N = 250$, $C = 2$, and asymmetric thresholds. As expected, the power of $\tilde{T}_{MV}$ was higher for conditions with more missing data and for ICM3, where the misspecification was more severe.

Comparing $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$, the results echoed those of ICM1. Specifically, $\tilde{T}_{YB}$ had lower power compared with $\tilde{T}_{MV}$ across all conditions, except in a few conditions, where both yielded a power of 100%. Focusing on the conditions, where $\tilde{T}_{MV}$ yielded inflated Type I error rates, the power of $\tilde{T}_{YB}$ was non-existent, consistent with the results of ICM1. This suggested that $\tilde{T}_{YB}$ might not be a viable alternative to $\tilde{T}_{MV}$, even when $\tilde{T}_{MV}$ did not perform well.

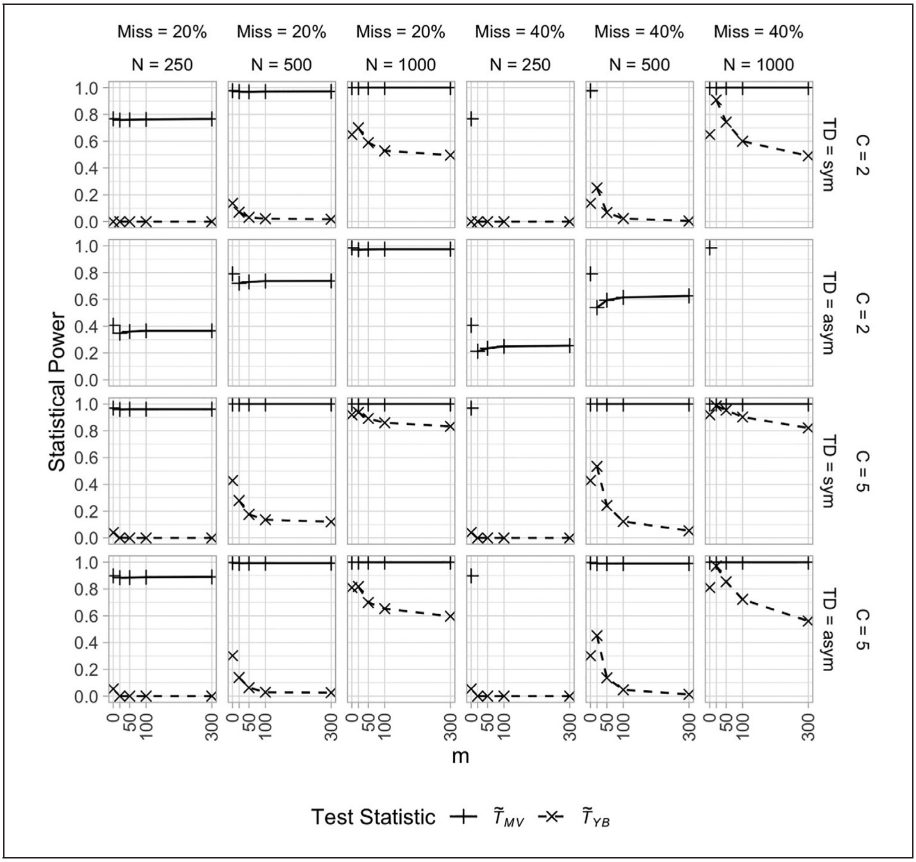

Figure 6. Empirical Power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ for Incorrect Model 1 Under MAR2.
Note. Power is omitted when the corresponding Type I error rate of m = 300 or complete data analysis is inflated. The results from corresponding complete data analyses are displayed at m = 0. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; m = number of imputed data sets.

In summary, the results showed that the statistical power of MI2S-based test statistics $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ varied across different conditions. $\tilde{T}_{MV}$ demonstrated consistent and relatively high power in detecting misspecified models. In contrast, $\tilde{T}_{YB}$ exhibited noticeably lower power, especially with smaller N and larger m.

MI2S-Based RMSEA for the Incorrect Models

To examine the performance of the MI2S-based RMSEA in each condition, we compared its mean values across replications with the population RMSEA value, as well as those from complete data analyses. Table 3 shows the mean MI2S-based RMSEA

Table 2. Empirical Power of $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ for Incorrect Models 2 and 3 With N $\leq$ 500 and m = 300.

				ICM2				ICM3			
				$\tilde{T}_{MV}$		$\tilde{T}_{YB}$		$\tilde{T}_{MV}$		$\tilde{T}_{YB}$	
C	TD	*N*	Miss	MAR1	MAR2	MAR1	MAR2	MAR1	MAR2	MAR1	MAR2
2	sym	250	0	1	1	.073	.072	1	1	.015	.009
			20	.999	1	.001	.004	1	1	.001	0
			40	.986	—	.005	0	1	—	.003	0
		500	0	1	1	.998	.995	1	1	.998	.999
			20	1	1	.779	.976	1	1	.803	.991
			40	1	—	.122	.948	—	—	.084	.964
	asym	250	0	.935	.940	—	—	.998	.998	—	—
			20	.910	.880	—	—	.997	.990	—	—
			40	.873	.658	—	—	.991	.943	—	—
		500	0	1	1	—	—	1	1	—	—
			20	.999	1	—	—	1	1	—	—
			40	.999	.997	—	—	1	1	—	—
5	sym	250	0	1	1	.822	.828	1	1	.744	.770
			20	1	1	.006	.168	1	1	0	.031
			40	1	—	.004	.066	1	—	.002	.004
		500	0	1	1	1	1	1	1	1	1
			20	1	1	.999	1	1	1	.998	1
			40	1	1	.648	1	1	1	.588	1
	asym	250	0	1	1	.567	.563	1	1	.425	.465
			20	1	1	.083	.024	1	1	.012	.001
			40	1	—	.015	.008	1	—	.002	.001
		500	0	1	1	1	1	1	1	1	1
			20	1	1	1	1	1	1	.998	1
			40	1	1	.822	.999	1	1	.843	.998

Note. Power results are omitted when the corresponding Type I error rate at m = 300 or in the complete data analysis is inflated. C = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; N = sample size; Miss = missing data rate; ICM = incorrect model.

at $m = 300$, alongside its complete data counterpart, for ICM1, ICM2, and ICM3. It is important to note that the effect of m on the MI2S-based RMSEA was trivial, thus we only included $m = 300$ in the table.

For ICM1, the mean MI2S-based RMSEA values ranged from 0.054 to 0.069, which were reasonably close to the population value of 0.065. These corresponded to the results obtained from complete data analyses ranging from 0.061 to 0.064, except in some extreme conditions (e.g., $N = 250$% and 40% missing data rate). For ICM2, the mean MI2S-based RMSEA values ranged from 0.090 to 0.097, aligning well with the population value of 0.095 and the results from complete data analyses ranging from 0.094 to 0.096. For ICM3, the mean MI2S-based RMSEA values ranged from 0.130 to 0.138, close to the population value of 0.134 and the results from complete data analyses, ranging from 0.133 to 0.137. The ranges of the mean MI2S-based RMSEA values for ICM1 (0.054 to 0.069), ICM2 (0.090 to 0.097), and ICM3 (0.130 to 0.138) did not change when including or excluding conditions with $m = 20$, 50, and 100 imputed data sets.

In addition, we assessed the empirical coverage of 90% confidence intervals (CIs) for the MI2S-based RMSEA. Coverage represents the proportion of time that a CI contains the true parameter value. We chose a 90% CI because it is commonly reported in the literature. Given that the effect of the number of imputed data sets was marginal, we present coverage rates only when $m = 300$. As shown in Table 4, the coverage rates were generally acceptable, closely approaching the nominal level of 0.90. However, in a few conditions under MAR1 and ICM1, the coverage rates appeared concerning, dropping below 0.85, particularly when $N \leq 500$, $C = 5$, and 40% missing data rate. Nevertheless, all coverage rates remained above 0.80.

Overall, these results suggested that the MI2S-based RMSEA generally performed well for gauging the model-data fit, but it tended to be less optimal with smaller sample sizes, higher rates of missing values, and in cases of smaller model misfit.

In summary, our simulation results showed that $\tilde{T}_{MV}$ outperformed $\tilde{T}_M$, $\tilde{T}_B$, and $\tilde{T}_{YB}$, and the MI2S-based RMSEA appeared to provide useful information in determining the fit of SEM with multiply imputed ordinal data.

Empirical Example

In this section, we demonstrate the practical application of the MI2S approach for ordinal data using a real-world data set. We consider a scenario where our interest lies in testing the configural invariance of positive affect across three-time points (see Liu et al., 2017, for details on measurement invariance). The R scripts and Mplus files used for this demonstration are available from the OSF.

Data Sets

We used publicly available data sets from the MIDUS (Midlife in the United States) longitudinal study (www.midus.wisc.edu). The first wave began in 1995 (MIDUS 1)

Table 3. Mean MI2S-Based RMSEA for Incorrect Models 1, 2, and 3 at m = 300.

				ICM1		ICM2		ICM3	
C	TD	*N*	Miss	MAR1	MAR2	MAR1	MAR2	MAR1	MAR2
2	sym	250	0	.061	.061	.094	.095	.134	.134
			20	.062	.062	.093	.096	.132	.134
			40	.063	.063	.092	.097	.131	.134
		500	0	.063	.063	.094	.094	.133	.133
			20	.063	.064	.094	.094	.133	.134
			40	.063	.064	.093	.095	.132	.134
		1000	0	.064	.064	.095	.094	.134	.134
			20	.063	.064	.095	.095	.135	.134
			40	.063	.064	.094	.095	.134	.134
	asym	250	0	.063	.062	.096	.096	.135	.137
			20	.065	.061	.095	.096	.133	.136
			40	.069	**.054**	.092	.090	.130	.131
		500	0	.061	.062	.095	.095	.134	.135
			20	.062	.062	.095	.096	.134	.136
			40	.065	.062	.093	.096	.132	.138
		1000	0	.064	.063	.094	.095	.134	.134
			20	.064	.064	.094	.095	.134	.135
			40	.065	.065	.094	.097	.133	.138
5	sym	250	0	.062	.063	.094	.095	.133	.135
			20	.061	.064	.095	.095	.134	.135
			40	**.058**	.064	.096	.095	.135	.135
		500	0	.063	.064	.095	.095	.134	.134
			20	.063	.064	.095	.095	.134	.134
			40	.062	.064	.095	.095	.134	.134
		1000	0	.064	.064	.095	.095	.134	.134
			20	.063	.064	.095	.095	.134	.134
			40	.063	.064	.095	.095	.134	.134
	asym	250	0	.062	.062	.095	.094	.133	.134
			20	.061	.063	.095	.095	.134	.134
			40	**.058**	.065	.097	.096	.136	.134
		500	0	.063	.063	.094	.095	.134	.134
			20	.063	.063	.095	.095	.134	.134
			40	.062	.063	.096	.096	.136	.135
		1000	0	.064	.064	.094	.095	.134	.134
			20	.064	.064	.095	.095	.135	.134
			40	.064	.064	.095	.095	.135	.135
				(.065)	(.065)	(.095)	(.095)	(.134)	(.134)

Note. The population RMSEA for each incorrect model is in parentheses. Relative biases (in absolute values) above .100 are bolded. *C* = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; *N* = sample size; Miss = missing data rate; ICM = incorrect model.

Table 4. Empirical Coverage Rates of the MI2S-Based RMSEA for Incorrect Models 1, 2, and 3 at 90% CIs and m = 300.

				ICM1		ICM2		ICM3	
C	TD	*N*	Miss	MAR1	MAR2	MAR1	MAR2	MAR1	MAR2
2	sym	250	0	.883	.872	.920	.927	.898	.909
			20	.889	.887	.908	.922	.878	.917
			40	.893	.912	.886	.934	.858	.904
		500	0	.879	.869	.910	.909	.906	.905
			20	.879	.872	.908	.913	.905	.899
			40	.864	.889	.895	.918	.882	.912
		1,000	0	.890	.894	.923	.912	.917	.893
			20	.886	.896	.920	.912	.921	.891
			40	.879	.897	.924	.917	.908	.902
	asym	250	0	.933	.919	.943	.940	.918	.905
			20	.938	.924	.929	.953	.897	.904
			40	.932	.932	.916	.950	.853	.877
		500	0	.879	.900	.925	.912	.900	.902
			20	.888	.913	.922	.919	.896	.919
			40	.903	.930	.912	.940	.873	.917
		1,000	0	.887	.884	.926	.933	.906	.901
			20	.882	.904	.927	.935	.904	.898
			40	.883	.917	.896	.947	.876	.907
5	sym	250	0	.866	.877	.911	.884	.898	.899
			20	**.849**	.878	.899	.895	.909	.886
			40	**.824**	.879	.901	.893	.908	.898
		500	0	.875	.888	.882	.892	.884	.899
			20	.861	.890	.898	.890	.895	.897
			40	**.846**	.891	.894	.906	.904	.896
		1,000	0	.891	.881	.913	.888	.888	.895
			20	.887	.884	.917	.892	.900	.894
			40	.875	.879	.908	.883	.886	.890
	asym	250	0	.851	.858	.898	.906	.890	.895
			20	**.826**	.872	.879	.912	.885	.901
			40	**.817**	.882	.889	.937	.897	.896
		500	0	.862	.877	.894	.903	.888	.896
			20	.872	.882	.897	.907	.873	.900
			40	**.837**	.894	.907	.898	.884	.899
		1,000	0	.909	.868	.907	.888	.897	.886
			20	.892	.862	.894	.893	.888	.892
			40	.885	.881	.906	.893	.895	.904

Note. The nominal level of the coverage rate is .900. Values below .850 are bolded. *C* = number of response categories; TD = threshold distribution; sym = symmetric; asym = asymmetric; *N* = sample size; Miss = missing data rate; ICM = incorrect model.

and two follow-up waves began in 2004 (MIDUS 2) and 2013 (MIDUS 3). More details can be found in the MIDUS documentation (www.icpsr.umich.edu/web/ICPSR/series/203). The data used for this demonstration contained 895 adults who graduated from college at MIDUS 1 ($M_{\text{age}} = 45.7$, $SD = 12.4$). Among them, 601 (67%) and 468 (52%) returned at MIDUS 2 and 3, respectively. Within the first, second, and third wave, 1 (0.1%), 13 (2.1%), and 3 (0.6%) participated adults had partial missing data on the positive affect items, respectively. That is, missing values were mostly due to dropouts. Positive affect was measured by six items. Participants responded on a 6-point scale from 1 (*all of the time*) to 5 (*none of the time*), ''During the past 30 days, how much of the time did you feel [cheerful/in good spirits/extremely happy/calm and peaceful/satisfied/full of life]?'' (Mroczek & Kolarz, 1998).

Multiple Imputation, Analysis, and Results

Similar to our simulation study, we utilized multiple imputation using an unrestricted variance-covariance model for ordinal data in Mplus 8.6. Given that missing data were mainly due to loss to follow-up, we incorporated six auxiliary variables to make the MAR assumption more plausible. The auxiliary variables—age, gender, marital status, parental status, physical health, and mental health—were measured at MIDUS 1. Song et al. (2021) showed that these variables were predictive of retention and attrition in the MIDUS longitudinal study. These auxiliary variables, particularly mental health, were also correlated with positive affect. Based on convergence diagnostics (i.e., PSR values and trace plots), we anticipated convergence to occur after approximately 9,200 iterations. To ensure convergence, we doubled the number of iterations. Therefore, we specified 18,400 burn-in iterations for multiple imputation and generated 500 imputed data sets.

For each imputed data set, the polychoric correlations, thresholds, and asymptotic covariances were estimated. These 500 sets of summary statistics were then combined into a single set using the MI2S approach. This process was automated through the *lavaan.mi* package in R (Jorgensen, 2023). Following this, the pooled summary statistics were utilized as input data for estimating the three-factor longitudinal CFA model via the *lavaan* package. Several MI2S-based statistics, including $\tilde{T}_M$, $\tilde{T}_{MV}$, and $\tilde{T}_B$, were readily obtained. $\tilde{T}_{YB}$ was calculated using Equation 4. Finally, the MI2S-based RMSEA for ordinal data was computed with a custom R script provided by Lai (2020).

The results showed that, although the chi-square test statistics were significant, approximate fit indices did not suggest serious misfit, $\tilde{T}_M = 1152.0$, $\tilde{T}_{MV} = 778.3$, $\tilde{T}_B = 515.1$, $\tilde{T}_{YB} = 326.7$, $df = 132$, ps $< .001$, MI2S-based RMSEA = 0.067, 90% CI [0.059, 0.073]. Thus, we concluded that the fit of the longitudinal configural invariance model was acceptable, indicating that the same construct was being measured across time.

Discussion

Chung and Cai's (2019) MI2S approach holds promise for evaluating the model fit of SEMs with multiply imputed ordinal data. Under MI2S, the hypothesized model is fitted once, and a single set of imputation-based fit statistics is readily available from SEM software packages. While previous studies evaluated $\tilde{T}_B$ and $\tilde{T}_{YB}$, this study additionally examined $\tilde{T}_M$ and $\tilde{T}_{MV}$ across a wide range of conditions.

The present study extends previous research in several ways. First, we found that $\tilde{T}_{MV}$ generally exhibited good performance, consistent with the performance of its complete-data counterpart (e.g., Savalei & Rhemtulla, 2013). $\tilde{T}_{MV}$ consistently maintained acceptable Type I error control, except for a few conditions with a high missing data rate under MAR2, where $\tilde{T}_{MV}$ could produce slightly inflated Type I error rates with smaller sample sizes. Although MAR1 and MAR2 differed in several ways, we speculate that the poorer performance under MAR2 is partly due to the strength of the association between the cause of missingness and the binary missing data indicators.

To support this speculation, we conducted an additional simulation where everything remained the same as MAR2 except we set parameters β_0 to -2.1 and β_1 to 8, leading to 40% missing data and a pseudo-R^2 value of .95 (instead of .40). These adjustments made the auxiliary variable almost perfectly predicted missingness, thus making the missing data pattern closely resemble the monotone pattern of MAR1. This simulation showed that $\tilde{T}_{MV}$ yielded non-inflated Type I error rates across conditions, except for a slight inflation when $N = 250$ and $C = 5$ with asymmetric thresholds (see Supplemental Figure S3). These results suggested that when the cause of missingness was more strongly related to the occurrence of missing data, the Type I error rates tended to be closer to the nominal level.

Apart from the impact of the strength of the association, the poorer performance in asymmetric threshold conditions under MAR2 could be explained by the fact that MAR2 made the severe asymmetry even more extreme, whereas MAR1 mitigated it (see Supplemental Figures S1 and S2). Similar to the findings of Jia and Wu (2019) with MAR-tail (more extreme) and MAR-head (less extreme), this also explains why convergence issues were likely to occur in conditions with asymmetric thresholds under the more challenging MAR2 mechanism.

The second novel contribution of our study is that we found $\tilde{T}_{MV}$ generally outperformed $\tilde{T}_M$, $\tilde{T}_B$, and $\tilde{T}_{YB}$. Both $\tilde{T}_M$ and $\tilde{T}_B$ showed inflated Type I error rates across all conditions. The inflation in Type I error rates for $\tilde{T}_B$ aligns with our expectations and is consistent with previous findings (Liu et al., 2021). Although $\tilde{T}_M$ is a novel aspect of our study, investigations of its complete data equivalent, T_M, have indicated that it can be slightly inflated in certain situations, such as with medium-sized models containing around 15-20 ordinal items (DiStefano & Morgan, 2014; Liu & Sriutaisuk, 2020; Savalei & Rhemtulla, 2013). Unlike other test statistics examined, $\tilde{T}_{YB}$ often produced conservative Type I error rates which translated into a substantial loss of power to detect misspecifications. However, when data were binary and asymmetrical, $\tilde{T}_{YB}$ tended to produce substantially inflated Type I error rates. In light of these findings, we generally recommend $\tilde{T}_{MV}$ over $\tilde{T}_{M}$, $\tilde{T}_{B}$, and $\tilde{T}_{YB}$.

In contrast to our study, Chung and Cai (2019) found that $\tilde{T}_{YB}$ generally exhibited good performance. This inconsistency is likely because of the difference in the simulation design, such as the number of indicators and the missing data generation process. We also could not fully investigate very large samples (e.g., N = 5,000) due to constraints in computing resources. We believed that $\tilde{T}_{YB}$ should perform well given a large enough sample size and number of imputed data sets. To provide some evidence, we conducted an additional small-scale simulation with N = 5,000 and symmetric thresholds under the MAR1 mechanism (everything else remained the same as the main simulation). The findings for the correct model were as expected. That is, the empirical means and variances of $\tilde{T}_{YB}$ were close to the relevant results obtained from complete data analyses, and the empirical Type I error rates of $\tilde{T}_{YB}$ were all within the acceptable range even with 40% missing data (see Supplemental Figure S4).

Third, this study provides evidence indicating that the MI2S approach can perform well in challenging conditions with high rates of missing data, even without an extremely large number of imputed data sets. We found that $\tilde{T}_{MV}$ produced more stable solutions compared with $\tilde{T}_{YB}$. Specifically, with up to 20% missing data, $\tilde{T}_{MV}$ and $\tilde{T}_{YB}$ required at least m = 20 and 100, respectively. With a higher missing data rate (40%), $\tilde{T}_{MV}$ required m = 100 whereas $\tilde{T}_{YB}$ did not perform well even with m = 300. The instability of $\tilde{T}_{YB}$ is probably caused by the well-documented instability of the total asymptotic covariance matrix (Enders, 2022).[6] Although $\tilde{T}_{MV}$ also relies on the total asymptotic covariance matrix, the calculation of $\tilde{T}_{MV}$ is much simpler and does not involve inverting the entire asymptotic covariance matrix (see Equation 2).

Fourth, this study extends the literature by showing that the MI2S-based RMSEA can be used as a good measure of the model–data fit. It generally exhibited good performance, except in some conditions, where the missing data were high, coupled with a small sample size and a relatively small model misspecification. Our results are consistent with a recent simulation using the standard multiple imputation approach, which showed that Lai's (2020) RMSEA performed reasonably well, especially as sample size increased (from N = 200 to 1,000) and missingness decreased (from 50% to 15%; Shi et al., 2023).

How does MI2S perform relative to the standard multiple imputation approach in terms of model test statistics? To answer this question, we conducted additional analyses at m = 300, comparing $\tilde{T}_{MV}$ with the pooled chi-square test statistics from the standard multiple imputation approach, sometimes referred to as the D_2 test statistics (Liu & Sriutaisuk, 2020; see also Jia, 2023).[7] Compared with $\tilde{T}_{MV}$, D_2 produced lower rejection rates, resulting in Type I error rates that were zero or below .025, and consistently lower power than $\tilde{T}_{MV}$. Moreover, the performance of D_2 was considerably influenced by the missing data rate, such that the statistical power was greatly reduced when 40% of data were missing (see Supplemental Figures S5–S6). Overall, we found initial evidence favoring $\tilde{T}_{MV}$ over D_2 in most conditions. Nevertheless, in the few challenging conditions where $\tilde{T}_{MV}$ yielded inflated Type I error rates, D_2 could be a viable alternative. Future research is needed to thoroughly compare $\tilde{T}_{MV}$ with D_2 (see Liu et al., 2021 for the comparison of $\tilde{T}_{YB}$ and D_2).

There are several limitations and directions for future research. First, while we carefully designed our simulation, there are additional factors that may warrant a more thorough investigation that we did not investigate to limit the scope of this study (e.g., the reliability of items, proportion of incomplete items, and model size; Liu et al., 2021). Second, while our multiple imputation and analyses generally resulted in converged solutions and unbiased parameter estimates, we only examined the unrestricted variance-covariance model implemented in Mplus under MAR. The robustness of MI2S-based fit statistics across different multiple imputation approaches (e.g., fully conditional specification imputation; Van Buuren, 2007), missing data mechanisms, and software programs would be an avenue for future research. Third, studies recently proposed ways to test measurement invariance with multiply imputed data (e.g., Chen, Wu, Garnier-Villarreal, et al., 2020) and search for potential sources of misfit (e.g., Chen, Wu, Brandt, et al., 2020; Mansolf et al., 2020). An interesting avenue of future research is to explore the possibility of using MI2S for measurement invariance testing and model modification. Finally, recent studies have shown that Meng and Rubin's (1992) method for pooling likelihood-based statistics performs well for continuous normal and nonnormal data (Enders & Mansolf, 2018; Jia, 2023). Future research might compare the MI2S approach to the method proposed by Meng and Rubin (1992) in the context of ordinal data. It would also be beneficial to compare the standard multiple imputation-based approximate fit indices with the MI2S-based counterparts.

Conclusion

This study provides significant understanding regarding the performance of the MI2S approach in the context of SEM with ordinal variables and missing data. We found that $\tilde{T}_{MV}$ outperformed previously examined MI2S-based test statistics in most conditions. Importantly, it provided more stable solutions, particularly when working with a practical model size and a reasonable number of imputed data sets. While $\tilde{T}_{MV}$ has shown considerable promise, it ideally should be used in conjunction with approximate fit indices for a more comprehensive evaluation of model fit. Our study also supports the use of the MI2S-based RMSEA as a feasible measure for assessing the extent of misfit between the hypothesized model and the data.

Authors' Note

Seungwon Chung is an employee of the US Food and Drug Administration (FDA). The views expressed in this document are those of the author and should not be construed to represent official FDA views or policies. The information and analyses included in this document are provided for academic research purposes only and should not be considered FDA recommended approaches.

Yu Liu is no longer affiliated with the University of Houston.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has been granted funds from Faculty of Psychology, Chulalongkorn University. This project was completed in part with the supporting computing infrastructure provided by the Research Computing Data Core at the University of Houston, as well as by the National Science and Technology Development Agency, National e-Science Infrastructure Consortium, Chulalongkorn University and the Chulalongkorn Academic Advancement into Its 2nd Century Project (Thailand). URL: www.e-science.in.th

ORCID iD

Suppanut Sriutaisuk https://orcid.org/0000-0001-9412-5880

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Another reason we used ULS was because Lai's (2020) RMSEA is only appropriate under ULS. However, we also examined DWLS in several conditions. We omitted these results because they were essentially the same as those obtained with ULS and did not impact our conclusions.
2. A typographical error exists in Chung and Cai's (2019) Equation 24, where the term (N-1) is mistakenly not squared.
3. The missing data mechanism was a between-subjects factor where complete data sets for MAR1 and MAR2 were generated seperately. Our supplementary simulation showed that treating this factor as either a within-subjects or a between-subjects factor did not impact the conclusions.
4. We also examined the missing completely at random (MCAR) mechanism in several conditions. We omitted the results because they were essentially the same as MAR and did not impact our conclusions.
5. To evaluate the practical significance of the biases in thresholds, we examined the ''standardized'' bias (Collins et al., 2001). The results showed that the biases in thresholds, as well as factor loadings and covariances, were not practically significant. The largest standardized bias was 38.5% (in absolute value), which is less than the cutoff of 40% proposed by Collins et al. (2001, p. 340).
6. A more stable total asymptotic covariance matrix may be obtained using a simplified estimate of the total asymptotic covariance matrix proposed by K. H. Li, Raghunathan, et al. (1991). We did not examine this relatively common approach as $\tilde{T}_{MV}$ exhibited good performance without using it.

7. There are variants of the D_2 test statistics. The original D_2 is the result of pooling T_{MV} across imputations (Liu & Sriutaisuk, 2020). D_{2ASN} is the result of pooling T across imputations and then applying the scaling-and-shifting transformation to the pooled T (Jia, 2023). Our conclusions from comparing $\tilde{T}_{MV}$ with D_2 hold for both variants. D_2 and D_{2ASN} can be obtained using the *semTools* package (Jorgensen et al., 2022) in R.

References

Asparouhov, T., & Muthén, B. O. (2010a). *Simple second order chi-square correction*. https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Asparouhov, T., & Muthén, B. O. (2010b). *Weighted least squares estimation with missing data*. https://www.statmodel.com/download/GstrucMissingRevision.pdf

Asparouhov, T., & Muthén, B. O. (2022). *Multiple imputation with Mplus: Technical implementation*. https://www.statmodel.com/download/Imputations7.pdf

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

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). Sage.

Chen, P.-Y., Wu, W., Brandt, H., & Jia, F. (2020). Addressing missing data in specification search in measurement invariance testing with Likert-type scale variables: A comparison of two approaches. *Behavior Research Methods*, *52*, 2567–2587.

Chen, P.-Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2020). Testing measurement invariance with ordinal missing data: A comparison of estimators and missing data techniques. *Multivariate Behavioral Research*, *55*(1), 87–101.

Chung, S., & Cai, L. (2019). Alternative multiple imputation inference for categorical structural equation modeling. *Multivariate Behavioral Research*, *54*(3), 323–337.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438.

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford Press.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000563

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*(1), 76–93.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275–299.

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*, *16*(4), 625–641.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23.

Jia, F. (2023). Pooling test statistics across multiply imputed datasets for nonnormal items. *Behavior Research Methods*, *56*, 1229–1243. https://doi.org/10.3758/s13428-023-02088-3

Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior Research Methods*, *51*(5), 2337–2355.

Jorgensen, T. D. (2023). *lavaan.mi* (Version 0.1-0.0006) [Computer software]. GitHub. https://github.com/TDJorgensen/lavaan.mi

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (Version 0.5-6) [Computer software]. CRAN. https://cran.r-project.org/web/packages/semTools/index.html

Lai, K. (2020). Correct point estimator and confidence interval for RMSEA given categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 678–695.

Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics*, *37*(6), 675–702.

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*, 1065–1073.

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506.

Liu, Y., & Sriutaisuk, S. (2020). Evaluation of model fit in structural equation models with ordinal missing data: An examination of the D2 method. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 561–583.

Liu, Y., Sriutaisuk, S., & Chung, S. (2021). Evaluation of model fit in structural equation models with ordinal missing data: A comparison of the D2 and MI2S methods. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 740–762.

Mansolf, M., Jorgensen, T. D., & Enders, C. K. (2020). A multiple imputation score test for model modification in structural equation models. *Psychological Methods*, *25*(4), 393–411.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120.

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

Mroczek, D. K., & Kolarz, C. M. (1998). The effect of age on positive and negative affect: A developmental perspective on happiness. *Journal of Personality and Social Psychology*, *75*(5), 1333–1349.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.).

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*(1), 149–160.

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223.

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

Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 924–945.

Shi, D., Lee, T., Fairchild, A. J., & Maydeu-Olivares, A. (2020). Fitting ordinal factor analysis models with missing data: A comparison between pairwise deletion and multiple imputation. *Educational and Psychological Measurement*, *80*(1), 41–66.

Shi, D., Zhang, B., Liu, R., & Jiang, Z. (2023). Evaluating close fit in ordinal factor analysis models with multiply imputed data. *Educational and Psychological Measurement*, *84*, 171–189. https://doi.org/10.1177/001316442311588

Song, J., Radler, B. T., Lachman, M. E., Mailick, M. R., Si, Y., & Ryff, C. D. (2021). Who returns? Understanding varieties of longitudinal participation in MIDUS. *Journal of Aging and Health*, *33*(10), 896–907.

Teman, E. D. (2012). *The performance of multiple imputation and full information maximum likelihood for missing ordinal data in structural equation models* (Publication No. 3555133) [Doctoral dissertation, University of Northern Colorado]. ProQuest Dissertations and Theses Global.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392–423.

Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309.

Zyphur, M. J., Bonner, C. V., & Tay, L. (2023). Structural equation modeling in organizational research: The state of our science and some proposals for its future. *Annual Review of Organizational Psychology and Organizational Behavior*, *10*, 495–517.