Behavior Research Methods (2019) 51:2337–2355
https://doi.org/10.3758/s13428-018-1187-4

Evaluating methods for handling missing ordinal data in structural equation modeling

Fan Jia[1] · Wei Wu[2]

Published online: 25 January 2019

Abstract
Missing ordinal data are common in studies using structural equation modeling (SEM). Although several methods for dealing with missing ordinal data have been available, these methods often have not been systematically evaluated in SEM. In this study, we used Monte Carlo simulation to evaluate and compare five existing methods, including one direct robust estimation method and four multiple imputation methods, to deal with missing ordinal data. On the basis of the simulation results, we provide practical guidance to researchers in terms of the best way to deal with missing ordinal data in SEM.

Keywords Missing ordinal data · Structural equation modeling · Robust estimation · Multiple imputation

Ordinal data, such as those measured using Likert scales, are very common in the social and behavioral sciences. An ordinal variable contains a few response points, which are ordered, but the distances among the values are not meaningful. The past research on missing data has primarily focused on continuous missing data. Little guidance has been provided to researchers in terms of how to appropriately deal with missing ordinal data for their studies.

Such guidance is especially needed, given that multiple methods to deal with ordinal missing data have been made available, due to recent advances in missing data analysis and software developments (Enders, 2001b, 2010; Graham, 2009; Rubin, 1976, 1996; Schafer & Graham, 2002). Many of these methods have not been thoroughly studied and compared. Thus, it is not clear which method(s) should be adopted in a study that involves ordinal missing data. Of course, the appropriate method(s) may also vary depending on the kind of analysis adopted in the study. This article evaluates the performance of available methods for missing ordinal data, focusing on one of the most popular data analytical frameworks, structural equation modeling (SEM). These methods can be grouped into two categories: robust estimation methods and multiple imputation (MI) methods. Robust estimation methods deal with missing ordinal data without filling in the missing values. MI methods, on the other hand, replace missing ordinal data with multiple sets of plausible values.

It is important to point out that the performance of a missing data method is highly related to the mechanism through which data are missing. Rubin (1976) classified such processes into three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR refers to the case in which the probability of missing data for a variable is unrelated to the underlying values of the missing data or to any observed variables. MAR refers to the case in which the probability of missing data for a variable is related only to other observed variables. And MNAR refers to the case in which the probability of missing data for a variable is determined by the underlying values of the missing data themselves. Both MCAR and MAR are deemed ignorable, in the sense that the missing data generation process does not need to be explicitly modeled. We only consider methods developed for ignorable missingness in this study.

The rest of the article is organized as follows. We first describe the six methods that have been used or that theoretically can be used to deal with missing ordinal data in SEM. We then present a simulation study conducted to evaluate the performance of five of these methods under a variety of conditions for fitting a typical structural equation model. We conclude the article by discussing the results and limitations of the simulation study and providing recommendations on the use of these methods in practice.

✉ Fan Jia
fanjia@ku.edu

[1] University of Kansas, Lawrence, KS, USA

[2] Indiana University–Purdue University Indianapolis, Indianapolis, IN, USA

Robust estimation methods for missing ordinal data

Robust full information maximum likelihood (RFIML) RFIML treats missing ordinal data as if they were continuous. To understand how RFIML works, we start with the maximum-likelihood estimation method (ML) for continuous complete data. ML assumes that the data are continuous and multivariate normally distributed. Under certain assumptions, such as normality, independent observations, large sample size, and a correctly specified model (Bollen, 1989; Finney & DiStefano, 2006; Savalei & Falk, 2014), the parameter estimates produced by ML have desirable asymptotic properties, such as unbiasedness (i.e., they are close enough to the true population values), consistency (they converge to the true values as the sample size goes large), and efficiency (the sampling distribution of the estimates has minimum variance). If any of the assumptions are violated, these properties might not hold. In reality, the assumption of normality is always violated when data are not continuous. According to Bollen, when the observed indicators in SEM models are ordinal, there are at least two important consequences if the ordinal data are treated as normal: (1) the linear measurement model does not hold for ordinal indicators, and (2) the fundamental hypothesis of SEM does not hold—in other words, the population covariance matrix is not equal to the model-implied covariance matrix.

The effect of discontinuity on the performance of ML is dependent on at least two factors: (1) the distribution of categorical variables and (2) the number of categories. Researchers generally believe that when ordinal data are not severely skewed or kurtotic and have at least five categories, treating them as continuous does not result in severe bias in the parameter estimates, standard errors, or fit indices (Finney & DiStefano, 2006). In other situations, with a small number of categories or/and higher levels of skewness and kurtosis, bias in the parameter estimates and standard errors could be more pronounced, and the fit indices could be misleading (Dolan, 1994; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Muthén & Kaplan, 1985).

One solution to these problems is to use ML accompanied with robust correction for nonnormality. Satorra and Bentler (1994) developed a correction method by rescaling the standard errors and the test statistic from ML. This method is well-known as *Satorra–Bentler scaling* or *robust ML* (RML). However, RML is only applicable when the data are complete. When data are incomplete but normally distributed, SEM estimates can be obtained by iteratively maximizing the sum of N case-wise log-likelihood functions (Enders, 2001b). This method is referred to as *full information maximum likelihood* (FIML), which is one of the popular missing data techniques. To deal with missing nonnormal data with FIML, Yuan and Bentler (2000) developed three correction methods, of which the "direct" method is the most commonly used, known as *robust FIML* (RFIML). For continuous data, past research has shown that RFIML generally performed well under MCAR and MAR, except for one situation in which MAR missingness occurred mainly in the heavy tail of the distribution and the proportion of missing data was large (30%; Enders, 2001a; Savalei & Falk, 2014). Although RFIML was developed for missing continuous data, it has been widely used to deal with missing ordinal data in practice. However, research on the performance of RFIML for ordinal incomplete data is lacking.

Diagonally weighted least squares estimation methods (cat-DWLS) Another solution is to use an estimation method developed specifically for ordinal data. For example, weighted least squares (WLS) estimators are often used for ordinal data. WLS estimators account for ordinal data by fitting a SEM model to the polychoric correlation matrix. There are different versions of WLS estimators for ordinal data. Among these, cat-DWLS is the most popular. Briefly speaking, cat-DWLS uses the diagonal elements of the asymptotic polychoric correlation matrix as a correction factor for the covariance matrix of parameters in calculating the standard errors. Since cat-DWLS uses summary statistics in the estimation process, it cannot deal with missing data by itself, but needs to be combined with a missing data technique (Asparouhov & Muthén, 2010).

A common missing data technique used along with cat-DWLS is pairwise deletion. Specifically, pairwise deletion is used to calculate the polychoric correlations, which are then used in the cat-DWLS fit function. Asparouhov and Muthén (2010), however, showed that pairwise deletion could produce biased parameter estimates and unacceptable confidence interval coverage rates when the data were not MCAR. In addition, once the polychoric correlations are calculated on the basis of pairwise deletion, cat-DWLS treats them as if they were from complete data when estimating the model parameters. Thus, the uncertainty due to missing data is not taken into account, leading to inflated Type I error rates. Given the obvious disadvantages of pairwise deletion, we do not examine it in the present study. In comparison, multiple imputation (MI) is a better strategy to use when combined with cat-DWLS (Asparouhov & Muthén, 2010).

MI methods for missing ordinal data

MI is a widely used modern missing data technique designed for ignorable missingness (Rubin, 1987; Schafer & Graham, 2002). A standard MI procedure involves three phases: (1) the *imputation* phase—generate multiple sets of complete data with missing values filled in using a specific imputation model; (2) the *analysis* phase—fit the hypothesized model to each of the imputed data sets; and (3) the *pooling* phase—pool the results (e.g., the parameter estimates, standard errors, and fit indices) across the imputed data sets to produce a final set of results. We refer to the model used to impute/predict missing data as an *imputation model*. An imputation model can be either parametric or nonparametric. In this article, we consider MI with three parametric imputation models and one nonparametric imputation model. Depending on which imputation model is used, there are different MI methods. The parametric imputation methods include imputing the data based on multivariate normal distributions (MI-MVN), ordinal logistic regression models (MI-LOGIT), or latent variable models (MI-LV). The nonparametric imputation method is MI using random forests (MI-RF). These methods are described in detail below.

MI-MVN MI-MVN treats ordinal data as if they were continuous and generates imputed data sets based on a multivariate normal distribution. Although MI-MVN is not designed for ordinal missing data, it has been used in practice to deal with ordinal missing data, because of its wide availability. The imputed values from this method are continuous. For ordinal missing data, past research has recommended keeping the fractional part of the continuous imputed values rather than rounding them, unless a follow-up analysis requires use of a categorical metric for the imputed values (Enders, 2010; Graham, 2009; Honaker, King, & Blackwell, 2011; Schafer & Graham, 2002). Wu, Jia, and Enders (2015) found that MI-MVN generally performed well when imputing Likert-type ordinal missing values that were then aggregated to scale scores for regression analysis, unless the sample size was small and the distributions of the ordinal variables were severely asymmetrical. Limited research has examined the performance of MI-MVN in the context of SEM with missing ordinal data.

MI-LV The imputation model used in MI-MVN is not designed specifically for ordinal data. Given this reality, it is natural to think, why not use a statistical model designed for ordinal data instead? One popular model for predicting ordinal data is the so-called latent variable model, which is basically a formulation of the cumulative/ordinal probit model (Cowles, 1996). The latent variable model assumes that a continuous latent variable underlies each observed ordinal variable (Asparouhov & Muthén, 2010). The latent variables are typically assumed to follow a multivariate normal distribution. When this model is used for imputation, the missing values are first imputed at the continuous latent-variable level and then discretized on the basis of estimated thresholds.

MI-LV has received increasing attention in recent years. Asparouhov and Muthén (2010) compared using the MI-LV method followed by cat-DWLS with using direct cat-DWLS along with pairwise deletion for estimating a growth model of five binary variables observed at five time points. They found that MI-LV outperformed direct cat-DWLS by providing more accurate parameter estimates and higher confidence interval coverage under MAR. Wu et al. (2015) also found that MI-LV performed well in a context in which the ordinal variables were to be aggregated to scale scores for regression analysis, regardless of the missing data proportions, sample sizes, numbers of categories of the ordinal data, and the degree of unbalance of the categories. However, the performance of MI-LV has not been systematically examined in SEM.

MI-LOGIT Another popular model for predicting ordinal data is ordinal logistic regression. This imputation model is used with the chained equations algorithm (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren & Groothuis-Oudshoorn, 2011). Unlike MI-MVN or MI-LV, the chained equations algorithm does not impute on the basis of a joint distribution. Rather, it imputes missing data on a variable-by-variable basis. Prediction of the missing data for each variable is conditional on the current values of the other variables at a specific iteration. The imputation model for each missing data variable can be specified individually. Van Buuren et al. (2006) found that MI-LOGIT was superior to listwise deletion when estimating odds ratios. Van Buuren (2007) also recommended using MI-LOGIT rather than MI-MVN for ordinal logistic regression analysis. Wu et al. (2015), however, found that MI-LOGIT led to substantial bias in estimating reliability coefficients, mean scale scores, and regression coefficients for predicting one scale score from another when the items that formed the scale were ordinal, especially with small sample sizes, unbalanced categories, and more than five categories. Thus, an imputation model designed specifically for ordinal data might not necessarily have satisfying empirical performance in predicting missing ordinal data.

MI-RF All of the imputation models introduced above are parametric. Random forests (RF), on the other hand, is a nonparametric method that can be used to predict both continuous and categorical data, including ordinal data using the chained equations algorithm. Briefly speaking, RF is a recursive partitioning method that predicts a variable with missing values by successively splitting the data set based on one predictor at a time, so that the subsets become more homogeneous with each split (Breiman, 2001). One advantage of RF is that it does not rely on distributional assumptions, and thus has the potential to accommodate nonnormality and nonlinear relationships among the variables that cannot be easily parameterized (Doove, van Buuren, & Dusseldorp, 2014; Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014). When used for MI, the data are bootstrapped first. Each bootstrapped sample is then split into several subsets. The values in each subset are called "donors." Missing data are then imputed by random draws from the donors. Doove et al. (2014) compared MI-RF with MI-LOGIT for recovering interaction effects in logistic regression analyses. They found that MI-RF produced more accurate estimates of the interaction effects and was more efficient (yielded smaller standard errors) than MI-LOGIT. However, MI-RF and MI-LOGIT have not been compared in the context of SEM.

In sum, in this study we considered five methods that have the potential to be used for missing ordinal data in SEM. These methods include RFIML and different forms of MI based on the multivariate normal distribution (MI-MVN), based on the latent variable model (MI-LV), using ordinal logistic regression (MI-LOGIT), and using random forests (MI-RF). Among these methods, RFIML and MI-MVN treat missing ordinal data as if they were continuous. All of the methods are parametric, except for MI-RF. A brief summary of the characteristics of the five methods can be found in Table 1.

Simulation study

In this section, we use Monte Carlo simulation to evaluate the performance of the five methods. We attempt to address the following three questions.

> Question 1: Are the continuous-data methods RFIML and MI-MVN applicable to ordinal data? Under what situations and to what extent are the two methods robust to discontinuity?
> Question 2: How is the performance of each of the methods influenced by number of categories, asymmetry of thresholds, sample size, missing data proportion, and missing data mechanism?
> Question 3: Which of the five methods performs best under the examined conditions?

Data generation model

Following Ferrari and Barbiero (2012), we generated continuous data first and then discretized the continuous data into ordinal data. The continuous data were generated on the basis of the SEM used in Enders (2001a; see Fig. 1), which represents a model that is often seen in the SEM literature (e.g., Bollen, 1989; Palomo, Dunson, & Bollen, 2011). The model has three latent variables: η_1, η_2, and η_3, where η_3 was predicted by η_1 and η_2, and η_2 was predicted by η_1. As is shown in Fig. 1, the values of the structural paths among the three variables were .4 ($\eta_1 \rightarrow \eta_2$), .286 ($\eta_2 \rightarrow \eta_3$), and .286 ($\eta_1 \rightarrow \eta_3$), respectively. The variance of η_1 was fixed at 1 for identification purposes. The residual variances of η_2 and η_3 were set to .840 and .771, respectively, so that their total variances would both equal 1. Each latent variable was indicated by three continuous variables/items with all loadings set to .70. The residual variances on the indicators were all set to .51 so that the indicators would form a standardized metric. These continuous indicators were then discretized on the basis of thresholds in order to create ordinal indicators. For ordinal indicators with G categories, there were G minus 1 thresholds. To examine how the methods performed under different conditions, we manipulated the following factors.

Design factors

Number of categories Both dichotomous and polytomous ordinal data were considered in the study. The numbers of ordinal categories were set at two, three, and five.

Asymmetry of thresholds In practice, the distributions of ordinal indicators can be either symmetric or asymmetric. Following Rhemtulla, Brosseau-Liard, and Savalei (2012), we varied the asymmetry of the thresholds at three levels (symmetry, moderate asymmetry, and severe asymmetry) in order to introduce different levels of asymmetry in the item distributions (see Table 2). For the sake of simplicity, all variables shared the same set of thresholds and number of categories in each condition.

Table 1 Summary of methods for missing ordinal data

Method label	Method description	Software
RFIML	FIML with the rescaling strategy proposed by Yuan and Bentler (2000).	R (lavaan with the "MLR" estimator)
MI-MVN	Multivariate normal imputation using the EMB algorithm (Honaker et al., 2011). Robust ML (Satorra & Bentler, 1994) was used in the analysis phase.	R (Amelia, and lavaan with the "MLM" estimator)
MI-LV	Latent variable imputation (Asparouhov & Muthén, 2010). Cat-DWLS (Muthén & Muthén, 2012) was used in the analysis phase.	MPlus with the "WLSMV" estimator
MI-LOGIT	Multiple imputation by chained equations with a logistic regression model for dichotomous variables, and with an ordinal logistic regression model for polytomous variables (van Buuren, 2012). Cat-DWLS (Muthén & Muthén, 2012) was used in the analysis phase	R (mice, and lavaan with the "WLSMV" estimator)
MI-RF	Multiple imputation by chained equation with random forests, which involves a random selection of a smaller group of predictors at each split (Doove et al., 2014). Cat-DWLS (Muthén & Muthén, 2012) was used in the analysis phase.	R (mice, and lavaan with the "WLSMV" estimator)

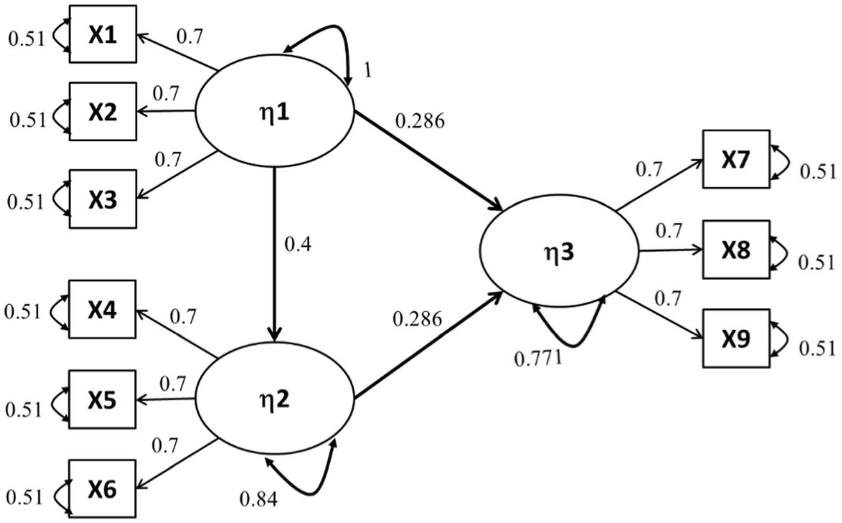

Fig. 1 The structural equation model for data generation

Sample size Two levels of sample size (N) were examined: 300 and 600,[1] which represent typical sample sizes in studies that have used SEM.

Missing data proportion Missing data proportions (mp) were manipulated at two levels: low (15%) and high (30%).

Missing data mechanism Missing data were imposed on two of the indicators for each latent variable. Specifically, missing values occurred for X_1, X_2, X_4, X_5, X_7, and X_8. The missing data were generated using three mechanisms: MCAR, MAR-head, and MAR-tail. MCAR data were generated by randomly selecting a desired proportion of values to be missing for the variables. When the ordinal indicators were asymmetric, we considered two versions of MAR: missingness occurring more frequently on the head of the distribution (MAR-head) or on the tail of the distribution (MAR-tail). To generate the MAR data, the rank order of the values on each of the fully observed variables (i.e., X_3, X_6, and X_9) was used to determine the probability of having a missing observation on the other two indicators for the same latent variable. For example, the missingness for X_1 and X_2 was determined by X_3. MAR-head data were generated on the basis of ascending rank order. Using X_1 as an example, the probability of having missing data for X_1 was computed as 1 - (the ascending order of the values on X_3/N). Because all variables were positively correlated, the probability of having missing observations for X_1 increased as X_3 increased. In addition, because all of the indicators were positively skewed, MAR-head led to more missing data on the head of the X_1 distribution.

In contrast, MAR-tail data were generated in such a way so that the probability of having missing data for X_1 decreased as X_3 increased. Consequently, more data were missing on the tail of the X_1 distribution; the distribution of the observed data became more skewed; and the density of the higher levels in the ordinal variable (e.g., four or five in a five-category variable) drastically decreased. In more extreme cases, some of the categories (e.g., five) might have zero observations. Therefore, we believe that MAR-tail was a more challenging situation than MAR-head, and it was necessary to differentiate the two situations. Figure 2 demonstrates the distribution of one three-category indicator from one replication with different degrees of asymmetry and missing data mechanisms.

In sum, there were 108 fully crossed conditions (3 × 3 × 2 × 2 × 3 = 108). One thousand replicated samples were created for each condition. The analysis model was the same as the data generation model. For the imputation methods, 50 imputed data sets were obtained for each replication. Following the guideline of White, Royston, and Wood (2011), 50 imputations should be sufficient for the amount of missing data simulated.

Computational characteristics

This simulation study was carried out using various packages in R 3.2 (R Core Team, 2015) and Mplus 7.2 (Muthén & Muthén, 2012). Data were generated using functions provided in the R package GenOrd (Ferrari & Barbiero, 2012). RFIML was implemented in lavaan (Rosseel, 2012). MI-MVN was implemented using Amelia (Honaker et al., 2011). MI-LOGIT and MI-RF were implemented through functions in the package mice (van Buuren & Groothuis-Oudshoorn, 2011), with ten burn-in iterations (van Buuren et al., 2006;

[1] Originally, we also considered a smaller sample size of 150. However, severe convergence problems were found for this sample size, and no estimates could be obtained in most conditions with the examined levels of missing data proportion and asymmetry of the distribution. Therefore, we do not report this sample size in the present article.

Table 2 Distributions of ordinal data used in the simulation

Threshold Condition	Number of Categories	Thresholds as *z* Scores				Percentages of Cases in Each Category				
Symmetric	2	0.00				50	50			
	3	– 0.83	0.83			20	59	20		
	5	– 1.50	– 0.50	0.50	1.50	7	24	38	24	7
Moderately Asymmetrical	2	0.36				64	36			
	3	– 0.50	0.76			31	47	22		
	5	– 0.70	0.39	1.16	2.05	24	41	22	10	2
Severely Asymmetrical	2	1.04				85	15			
	3	0.58	1.13			72	15	13		
	5	0.05	0.44	0.84	1.34	52	15	13	11	9

This is a subset of conditions from Rhemtulla, Brosseau-Liard, and Savalei (2012)

White et al., 2011). For MI-RF, ten bootstrap samples were generated from each original sample, based on the suggestion of Shah, Bartlett, Carpenter, Nicholas, and Hemingway (2014), and the default minimum number of donors (i.e., 1) was used to create classification trees (Liaw & Wiener, 2002). MI-LV was implemented in Mplus.

We used cat-DWLS to analyze the imputed data sets from all of the imputation methods, given that past research on complete ordinal data had shown that this technique generally outperformed RML (Li, 2016). The only exception was MI-MVN, for which RML was used in the follow-up analysis, because MI-MVN produced continuous imputed data, and cat-DWLS could not be applied. For the imputation methods, a replication was determined to be converged if the model converged for all 50 imputed datasets.

Evaluation criteria

The performance of the five methods was evaluated for four outcomes: proportion of convergence failures, relative bias in the parameter estimates (Est bias), relative bias in the standard errors (*SE* bias), and confidence interval coverage rate (CIC).

Proportion of convergence failures We defined convergence failures as replications that failed to converge to proper solutions. These included replications that failed to produce any solutions and replications that produced improper solutions (i.e., extreme parameter or standard error estimates). Improper solutions included (1) standard error estimates greater than 10, (2) parameter estimates ten *SD*s above or below the mean parameter estimate for the design cell, and (3) standard error estimates ten *SD*s above or below the mean standard error for the design cell. The convergence failures were removed before computing the other three outcomes. We calculated the proportion of the convergence failures in each condition.

Relative bias in parameter estimates (Est bias) The relative bias for a specific parameter estimate was calculated as the percentage of raw bias relative to the true population value:

$$\mathrm{Est\,Bias} = \frac{\left(\bar{\hat{\theta}}_{est} - \theta_0\right)}{\theta_0} \times 100\%$$

where the numerator represents the raw bias, which is the difference between the average parameter estimate across replications within a design cell ($\bar{\hat{\theta}}$) and the population value (θ_0). According to Hoogland and Boomsma (1998), an Est bias less than 5% is considered acceptable. However, Muthén, Kaplan, and Hollis (1987) argued that "a bias of less than 10%–15% may not be serious in most SEM contexts." We used 10% as the cutoff in the present study.

Relative bias in standard error estimates (SE bias) Bias in standard error estimates is the degree to which a standard error accurately reflects the sampling standard deviation of the corresponding parameter estimate, which can be calculated using the following formula:

$$\mathrm{SE\,Bias} = \frac{\overline{\mathrm{SE}} - ESE}{ESE} \times 100\%,$$

where $\overline{\mathrm{SE}}$ is the average standard error across replications in a design cell, and ESE is the empirical standard error (i.e., the standard deviation of the parameter estimates across converged replications). An SE bias is considered acceptable if its absolute value is less than 10% (Hoogland & Boomsma, 1998).

Confidence interval coverage (CIC) Confidence interval coverage was estimated as the percentage of replications in a design cell for which the 95% confidence intervals covered the population value. Ideally, the CIC values should be equal to 95%. Following Collins, Schafer, and Kam (2001), a coverage value below 90% was considered problematic.

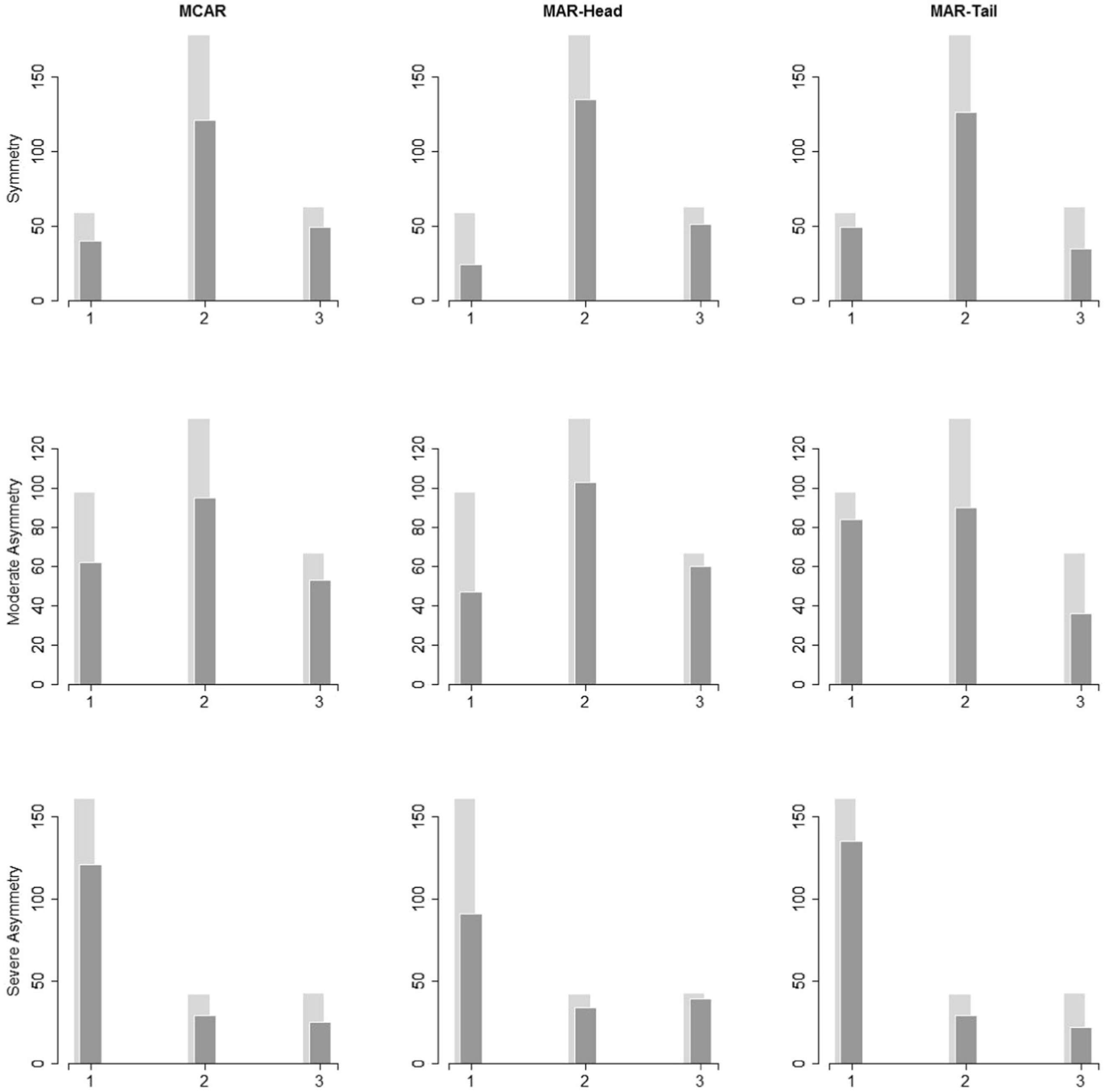

Fig. 2 Distributions of X_1 (three categories) from one replication with $N = 300$, both before (light gray) and after (dark gray) imposing 30% missing data

Results

Convergence failures

Symmetrical thresholds The convergence failures were low (below 15%) for all methods, regardless of *N*, *mp*, and missing data mechanism.

Moderately asymmetrical thresholds Some of the methods show significant convergence problems when the item distributions were moderately asymmetrical, particularly for dichotomous data and MAR-tail conditions. Figure 3 displays the proportions of convergence failures in conditions with moderately asymmetrical thresholds. All methods converged well under MCAR or MAR-head, except that MI-LOGIT failed in more than 60% of replications for five-category data with N = 300 and mp = 30%. More convergence failures occurred under MAR-tail. Specifically, for dichotomous data, all methods had convergence failures to some degree, with MI-RF and MI-MVN being the most problematic (more than 98% failures) when $N = 300$ and $mp = 30\%$. The convergence rates improved in general as *N* increased to 600. However,

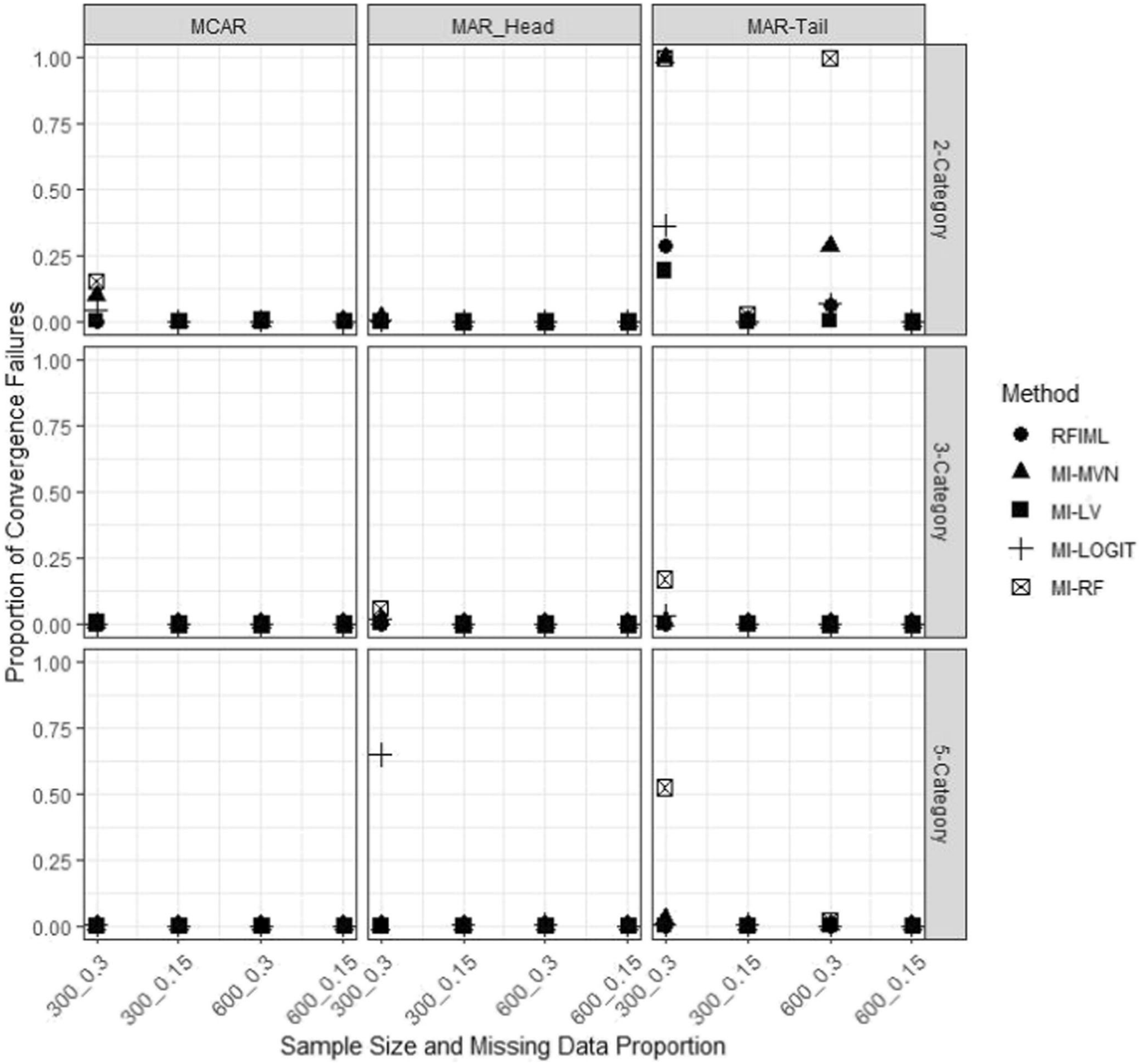

Fig. 3 Proportions of convergence failures with moderately asymmetrical distributions. RFIML = robust full-information maximum liklihood; MI-MVN = multiple imputation based on multivariate normal distributions; MI-LV = latent variable imputation; MI-LOGIT = multiple imputation using logistic regression or ordinal logistic regression; MI-RF = multiple imputation using random forests

with $N = 600$, MI-RF still had substantial convergence problems when mp = 30%. For polytomous data, all methods converged well, except that MI-RF produced substantial convergence failures for both three-catgory data and five-category data when $N = 300$ and mp = 30%. Overall, MI-LV seemed the best performer, followed by RFIML, and MI-RF was the worst performer, followed by MI-MVN and MI-LOGIT in terms of convergence.

Severely asymmetrical thresholds Convergence failures occurred more frequently for almost all methods when the item distributions were severely asymmetrical (see Fig. 4). Again, the convergence problem was more severe if N was small and mp was large. Also, MAR-tail continued to be the most challenging situation. Both RFIML and MI-LV seemed to have the least convergence problems under MCAR or MAR-head. Other methods also converged well under MCAR or MAR-head, except that they encountered substantial convergence problems for dichotomous data. Under MAR-tail, all methods had substantial convergence failures with dichotomous data, but MI-LV started to outperform the other methods quickly as sample size increased. For polytomous data, only MI-LV worked well for all conditions, except that it had substantial convergence problems (91% failures) for three-category data when $N = 300$ and $mp = 30\%$. The other methods all had substantial convergence failures under either $N = 300$, $mp = 30\%$, or both. Overall, in terms of convergence, MI-LV appeared to be the best performer, followed by RFIML, and MI-RF was the worst performer, followed by MI-MVN.

Note that if the proportion of convergence failures was higher than 75% for a method in a condition (Figs. 3 and 4), the method was considered failed for that condition, and the other three outcomes (i.e., Est bias, *SE* bias, and CIC) are not reported.

Est biases, SE biases, and CICs for factor loadings

We report Est biases, SE biases, and CICs for factor loadings first, followed by those for structural path coefficients. For ease of presentation, we report the averaged results across all factor loadings or all structural paths. The results for the conditions with 15% missing data followed patterns very similar to those for the conditions with 30% missing data. Thus, we report only the results for 30% missing data. (The results for 15% missing data can be requested from the authors.) In addition, the results were similar between symmetric thresholds

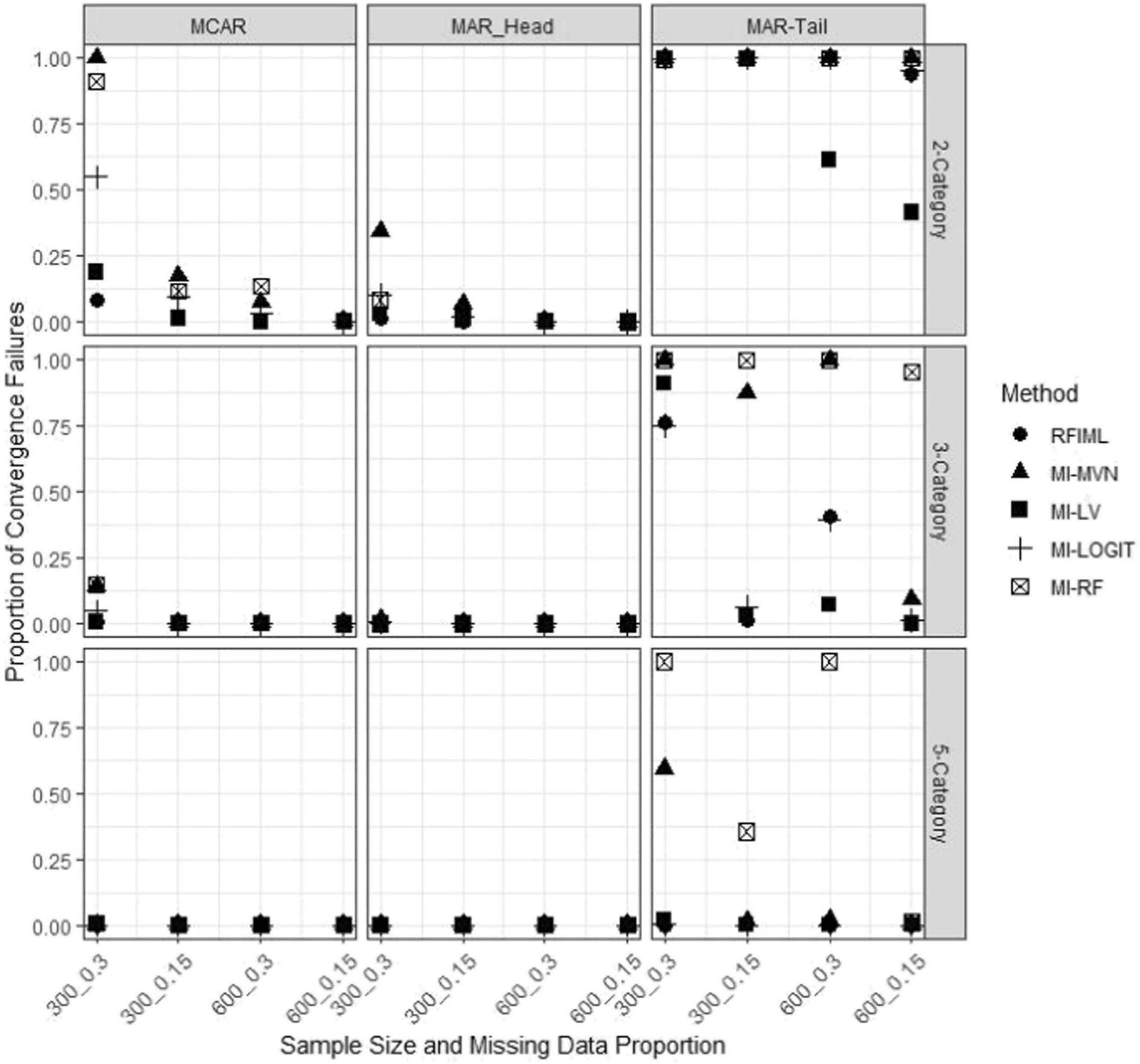

Fig. 4 Proportions of convergence failures with severely asymmetrical distributions. RFIML = robust full-information maximum liklihood; MI-MVN = multiple imputation based on multivariate normal distributions; MI-LV = latent variable imputation; MI-LOGIT = multiple imputation using logistic regression or ordinal logistic regression; MI-RF = multiple imputation using random forests

and moderately asymmetrical thresholds. Thus, the results for the two threshold conditions are described together below.

Symmetric or moderately asymmetrical thresholds For factor loadings, the results for Est bias, SE bias, and CIC under the two threshold conditions are summarized in Tables 3 and 4. The results show the same pattern across the two tables for all methods. Under MCAR, all methods performed well except for MI-RF, which produced unacceptable bias (> 10%) in both the parameter estimates and standard error estimates. Under MAR-head or MAR-tail, all methods performed well except for MI-RF and MI-LOGIT. Although MI-RF and MI-LOGIT produced acceptable parameter estimates, the standard error estimates from both methods were substantially biased, especially when the ordinal data had more than two categories. MI-LOGIT tended to produce too narrow CIs, whereas the CIs from MI-RF were too wide. Note that although MI-MVN and MI-LV performed well in general, there was one condition (moderately asymmetrical thresholds with N = 300, dichotomous indicators, and 30% MAR-tail missing data) under which MI-MVN failed to converge and MI-LV yielded unacceptable SE bias (15.2%; see Table 4).

Severely asymmetrical thresholds The results for factor loadings under severely asymmetrical thresholds are summarized in Table 5. Under MCAR, MI-LOGIT produced acceptable results under all conditions. RFIML, MI-MVN, and MI-LV also performed well, except for the dichotomous data with N = 300. MI-RF yielded problematic replications for dichotomous data and N = 300, and also failed all other conditions by producing biased results in terms of all three outcomes.

Under MAR-head, all methods performed well except for MI-RF. Under MAR-tail, however, all methods failed for dichotomous data, by either having convergence problems or resulting in biased results. For ordinal data with three categories, only MI-LV performed well, and only with a relatively large sample size of N = 600. For ordinal data with five categories, only MI-LV and RFIML performed well with both sample sizes. MI-MVN show acceptable performance only with N = 600.

Est biases, SE biases, and CICs for path coefficients

Symmetric or moderately asymmetrical thresholds The results for the path coefficients under symmetric thresholds and moderately asymmetrical thresholds are presented in Tables 6 and 7, respectively. Under MCAR, all methods performed well in general except for MI-RF, which produced substantial biases in the parameter or standard error estimates, as well as too-wide CIs. There was also one condition (N = 300 with dichotomous indicators) in which MI-LV produced slightly above 10% biases in the path coefficient estimates.

Under MAR-head and MAR-tail, RFIML was the best performer when the thresholds were not severely asymmetric. MI-MVN performed well in general, but it had convergence problems for dichotomous data with N = 300. MI-LV also had difficulties with this condition. It produced substantial biases in parameter estimates or standard error estimates, especially when the thresholds were moderately asymmetric. MI-RF and MI-LOGIT led to substantially biased parameter estimates or standard error estimates for many of the conditions, with MI-RF being more problematic. Again, MI-LOGIT produced too-narrow CIs, whereas MI-RF produced too-wide CIs.

Severely asymmetrical thresholds The results for path coefficients with severely asymmetrical thresholds are presented in Table 8. Under MCAR, only RFIML performed well for all outcomes. MI-MVN produced acceptable results for all conditions, except that it had convergence problems with N = 300 and dichotomous indicators. MI-LV and MI-LOGIT was acceptable only when the ordinal data had five categories. MI-RF produced substantial *SE* biases, or both Est biases and *SE* biases for all conditions.

Under MAR-head, the only method that worked under all conditions was RFIML. MI-MVN produced acceptable results for all outcomes for only polytomous indicators. MI-LV and MI-LOGIT produced substantial bias in the path coefficient estimates for all conditions except for conditions with five-category indicators. MI-RF produced unbiased parameter estimates across all conditions but overestimated the standard errors, leading to wide CIs.

Under MAR-tail, all methods failed for dichotomous data, regardless of sample size. For three-category data, no methods worked well with N = 300, and only RFIML produced acceptable results with N = 600. For ordinal data with five categories, RFIML, MI-MVN, and MI-LV were acceptable. MI-LOGIT failed to produce accurate standard errors and CIs under any conditions for polytomous data. MI-RF also encountered severe convergence problems for polytomous data.

Conclusion and discussion

In this article, we evaluated five available methods to deal with missing ordinal data in SEM across a broad range of conditions, using Monte Carlo simulation. In this section, we summarize the major findings for each of the research questions raised above.

> *Question 1:* Are the continuous-data methods RFIML and MI-MVN applicable to ordinal data? Under what situations and to what extent are the two methods robust to discontinuity?

Table 3 Results for loadings with symmetric thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
				N = 300, *ncat*=2					
RFIML	3.0	3.2	95	3.1	2.1	95	3.2	2.0	95
MI-MVN	3.4	2.9	94	4.4	1.8	95	4.5	2.0	95
MI-LV	1.9	3.9	94	2.8	2.6	95	3.0	1.1	95
MI-LOGIT	2.3	7.4	94	2.9	9.7	94	3.0	9.4	94
MI-RF	**13.0**	**26.7**	99	8.9	**22.9**	98	8.9	**22.6**	98
				N = 600, *ncat*=2					
RFIML	1.5	2.6	95	1.3	1.9	95	1.0	1.4	95
MI-MVN	1.7	4.0	95	1.8	1.6	95	1.5	2.9	95
MI-LV	0.9	3.8	95	1.0	1.2	96	0.9	1.5	96
MI-LOGIT	1.2	6.7	94	1.1	7.8	93	1.0	8.8	93
MI-RF	**10.7**	**21.9**	99	5.2	**19.6**	99	5.1	**18.8**	98
				N = 300, *ncat*=3					
RFIML	2.1	3.1	95	1.8	1.4	95	2.2	1.1	95
MI-MVN	2.3	4.9	94	3.3	2.0	95	3.7	2.5	95
MI-LV	1.3	4.7	95	2.4	2.0	96	2.4	2.3	95
MI-LOGIT	2.5	6.1	95	5.3	**27.1**	90	5.6	**25.2**	90
MI-RF	**16.6**	**28.9**	99	8.0	**35.7**	99	9.0	**36.5**	99
				N = 600, *ncat*=3					
RFIML	0.7	2.5	96	0.5	2.0	95	0.8	2.0	95
MI-MVN	0.8	1.2	95	1.2	1.8	95	1.5	1.4	95
MI-LV	0.3	1.3	95	0.8	2.8	95	1.0	1.6	96
MI-LOGIT	0.9	3.9	95	2.3	**27.7**	**87**	2.5	**28.5**	**86**
MI-RF	**14.7**	**32.1**	97	5.6	**35.7**	99	6.9	**37.1**	99
				N = 300, *ncat*=5					
RFIML	1.5	1.8	95	1.3	2.0	95	1.9	0.9	95
MI-MVN	1.7	2.8	95	2.5	2.0	95	3.2	2.2	95
MI-LV	1.0	3.4	95	1.7	3.0	96	2.1	3.0	96
MI-LOGIT	3.2	2.4	96	3.9	**16.4**	92	3.5	**31.2**	**85**
MI-RF	**15.4**	**31.8**	99	7.9	**31.9**	99	8.7	**34.9**	99
				N = 600, *ncat*=5					
RFIML	0.5	1.4	95	0.1	2.1	95	0.9	1.8	95
MI-MVN	0.6	1.9	95	0.6	1.7	95	1.5	1.9	95
MI-LV	0.3	1.9	95	0.5	2.5	95	1.2	1.8	95
MI-LOGIT	1.4	4.0	95	2.1	**29.7**	**85**	3.0	**43.9**	**77**
MI-RF	**13.5**	**30.6**	93	5.1	**32.1**	99	6.2	**34.7**	99

Values highlighted are the smallest Est or SE bias in each design cell. Values in bold are unacceptable (i.e., Est bias ≥ 10%, SE bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if the proportion of convergence failures ≥ 75%

The results show that the continuous-data methods RFIML and MI-MVN in general worked quite well for ordinal data in estimating factor loadings and structural path coefficients. In fact, in most conditions they outperformed the methods designed specifically for ordinal data, except for MI-LV, probably because of their simplicity. RFIML was reliable under

Table 4 Results for loadings with moderately asymmetrical thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
N = 300, ncat=2									
RFIML	3.7	2.5	95	1.4	3.9	95	7.9	4.9	95
MI-MVN	4.1	2.5	94	2.1	4.4	94	--	--	--
MI-LV	2.5	2.7	95	2.3	3.4	95	5.9	**15.2**	95
MI-LOGIT	3.2	9.4	94	2.9	7.4	94	4.6	**17.7**	91
MI-RF	**16.1**	**26.2**	99	9.5	**17.1**	98	--	--	--
N = 600, ncat=2									
RFIML	1.4	2.0	95	-0.3	2.5	94	5.0	2.5	95
MI-MVN	1.7	3.6	94	-0.1	2.6	94	7.4	3.4	96
MI-LV	0.8	1.5	94	1.0	1.7	95	3.0	3.3	96
MI-LOGIT	1.1	5.8	94	1.5	4.4	95	1.9	**19.3**	90
MI-RF	**13.4**	**23.3**	99	8.0	**19.1**	98	**--**	**--**	--
N = 300, ncat=3									
RFIML	1.8	3.1	95	0.8	3.8	95	2.4	1.4	95
MI-MVN	2.0	2.8	95	2.0	4.2	95	3.9	1.9	96
MI-LV	1.3	2.3	95	2.2	2.0	96	2.6	1.6	96
MI-LOGIT	1.9	5.1	95	3.2	**19.5**	92	4.8	**26.3**	90
MI-RF	**14.1**	**29.4**	99	9.6	**29.9**	99	**10.2**	**32.9**	99
N = 600, ncat=3									
RFIML	0.5	1.1	95	-0.5	3.7	95	1.3	2.7	96
MI-MVN	0.7	0.8	95	0.1	4.9	95	2.0	2.8	96
MI-LV	0.4	1.7	95	0.9	2.3	96	1.4	2.3	96
MI-LOGIT	0.7	5.3	94	1.6	**19.3**	90	2.6	**25.2**	**88**
MI-RF	**12.0**	**26.4**	97	6.6	**27.3**	99	7.8	**30.6**	99
N = 300, ncat=5									
RFIML	1.6	1.6	95	-0.1	1.7	94	4.1	0.9	95
MI-MVN	1.8	5.4	94	0.6	3.8	93	6.3	1.7	96
MI-LV	1.1	3.4	95	1.6	2.8	95	2.5	2.8	96
MI-LOGIT	3.3	3.3	96	2.9	**19.4**	91	3.0	**23.5**	91
MI-RF	**16.7**	**32.7**	99	8.5	**32.2**	99	**10.1**	**46.5**	99
N = 600, ncat=5									
RFIML	0.4	1.1	95	-1.3	2.5	93	2.5	2.2	96
MI-MVN	0.5	2.8	94	-1.0	2.3	93	3.5	2.8	96
MI-LV	0.2	1.2	95	0.6	2.5	95	1.0	1.7	96
MI-LOGIT	1.5	1.8	95	1.4	**21.3**	89	0.4	**26.6**	**86**
MI-RF	14.5	**35.1**	93	6.1	**34.6**	99	7.7	**40.7**	99

The highlighted values are the smallest Est or SE biases in each design cell. The second smallest biases might also be highlighted if the method with the smallest bias yielded unacceptable result on the other outcomes. Unacceptable values are in bold (i.e., Est bias ≥ 10%, SE bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if proportion of convergence failures ≥ 75%

various conditions examined in this study, except that it failed to converge for dichotomous data when the distributions were severely asymmetrical and were likely truncated by missing data (i.e., MAR-tail). MI-MVN also had convergence

Table 5 Results for loadings with severely asymmetrical thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
N = 300, ncat=2									
RFIML	**11.3**	3.3	94	5.7	4.7	93	--	--	--
MI-MVN	--	--	--	5.9	3.9	91	--	--	--
MI-LV	4.8	**13.5**	94	4.4	2.4	94	--	--	--
MI-LOGIT	5.9	3.7	94	6.2	4.8	94	--	--	--
MI-RF	--	--	--	**15.7**	**14.5**	98	--	--	--
N = 600, ncat=2									
RFIML	4.7	1.4	94	1.0	1.1	94	--	--	--
MI-MVN	5.0	9.8	92	1.1	8.2	92	--	--	--
MI-LV	1.7	2.9	94	2.2	2.6	95	**13.6**	**56.2**	97
MI-LOGIT	3.4	8.2	94	3.5	4.8	95	--	--	--
MI-RF	**21.7**	**28.5**	99	**13.0**	**14.5**	99	--	--	--
N = 300, ncat=3									
RFIML	4.1	2.9	95	0.9	2.2	94	--	--	--
MI-MVN	4.5	8.2	93	1.2	7.0	92	--	--	--
MI-LV	2.5	2.3	95	2.9	4.3	95	--	--	--
MI-LOGIT	3.5	6.1	95	3.2	5.7	95	--	--	--
MI-RF	**17.5**	**25.3**	99	9.8	**14.4**	98	--	--	--
N = 600, ncat=3									
RFIML	1.5	1.4	95	-1.3	0.6	93	**10.8**	**10.5**	94
MI-MVN	1.8	9.6	93	-1.2	4.5	92	--	--	--
MI-LV	0.5	1.7	95	1.2	2.1	95	7.5	3.2	96
MI-LOGIT	1.3	6.7	94	1.9	4.2	95	6.2	**43.1**	**79**
MI-RF	**14.4**	**22.7**	98	8.3	**15.1**	97	--	--	--
N = 300, ncat=5									
RFIML	2.1	2.2	95	-0.4	2.4	95	7.7	2.7	95
MI-MVN	2.4	6.4	94	0.0	3.9	94	**10.3**	**13.1**	96
MI-LV	1.4	2.3	95	2.2	2.2	96	4.1	4.6	96
MI-LOGIT	2.3	7.2	94	1.6	6.3	94	5.0	**32.7**	**87**
MI-RF	**15.0**	**30.0**	99	7.7	**25.8**	99	--	--	--
N = 600, ncat=5									
RFIML	1.0	1.1	95	-1.5	2.9	93	5.5	1.2	96
MI-MVN	1.1	5.5	94	-1.4	3.2	93	7.9	1.7	96
MI-LV	0.4	1.6	95	1.1	3.3	95	1.7	2.6	95
MI-LOGIT	1.0	6.1	94	1.4	5.7	94	4.0	**42.2**	**77**
MI-RF	**13.0**	**28.0**	97	6.5	**25.0**	98	--	--	--

The highlighted values are the smallest Est or SE biases in each design cell. The second smallest biases might also be highlighted if the method with the smallest bias yielded unacceptable result on the other outcomes. Unacceptable values are in bold (i.e., Est bias ≥ 10%, SE bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if the proportion of convergence failures ≥ 75%

problems when the ordinal data had ≤ 3 categories and when the distributions were asymmetrical.

As compared to RFIML, MI-MVN required a larger sample size or a smaller proportion of missing data in some of the

Table 6 Results for structural paths with symmetric thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
N = 300, ncat=2									
RFIML	2.8	4.3	95	4.4	3.3	95	4.3	3.7	95
MI-MVN	3.3	3.6	95	5.4	2.8	94	5.1	2.8	95
MI-LV	**10.3**	2.8	96	**11.8**	2.3	96	**11.5**	2.9	96
MI-LOGIT	9.2	6.9	95	**10.5**	6.3	94	**10.1**	9.1	94
MI-RF	**11.8**	**16.7**	99	**20.1**	**20.3**	99	**19.7**	**19.5**	99
N = 600, ncat=2									
RFIML	1.5	3.5	95	1.9	2.0	96	2.5	1.8	95
MI-MVN	1.8	5.5	94	2.5	3.4	95	3.0	3.8	94
MI-LV	6.7	4.9	95	7.1	2.8	96	7.5	3.7	95
MI-LOGIT	6.5	7.4	94	6.6	7.7	94	6.9	9.3	93
MI-RF	7.5	**13.5**	98	**14.4**	**12.0**	98	**15.2**	**10.8**	98
N = 300, ncat=3									
RFIML	1.1	3.1	95	2.5	2.3	95	3.5	1.5	96
MI-MVN	1.4	2.2	94	3.1	1.8	95	4.3	1.4	95
MI-LV	4.0	2.4	95	4.6	3.1	95	5.6	0.6	96
MI-LOGIT	3.9	4.4	95	5.4	**12.3**	92	4.6	**11.3**	93
MI-RF	5.0	**21.2**	99	**20.0**	**28.6**	99	**22.3**	**31.2**	100
N = 600, ncat=3									
RFIML	1.6	1.8	95	2.1	1.5	95	2.5	1.1	96
MI-MVN	1.8	2.9	94	2.6	2.1	95	3.0	1.2	95
MI-LV	2.7	2.9	94	2.9	3.0	95	3.4	1.6	95
MI-LOGIT	2.7	5.8	94	3.1	**16.7**	91	3.6	**15.1**	92
MI-RF	3.3	**20.4**	98	**16.7**	**20.9**	99	**17.4**	**23.3**	99
N = 300, ncat=5									
RFIML	0.9	2.7	95	1.9	1.5	95	2.4	3.9	95
MI-MVN	1.0	1.7	95	2.7	0.8	95	3.1	3.5	94
MI-LV	2.8	2.8	95	3.3	1.1	95	3.8	3.5	95
MI-LOGIT	2.8	3.2	95	6.2	6.6	94	6.9	**13.1**	92
MI-RF	4.7	**18.9**	98	**19.4**	**22.0**	99	**23.0**	**22.0**	99
N = 600, ncat=5									
RFIML	1.1	1.6	95	1.3	1.0	96	1.6	3.5	95
MI-MVN	1.3	2.6	94	1.6	1.4	95	2.0	3.2	95
MI-LV	1.9	3.1	95	1.8	1.7	95	2.0	2.4	95
MI-LOGIT	1.9	4.9	94	3.2	**15.0**	91	4.7	**23.9**	**89**
MI-RF	3.4	**16.7**	98	**16.3**	**16.8**	99	**19.0**	**19.3**	98

The highlighted values are the smallest Est or SE biases in each design cell. The second smallest biases might also be highlighted if the method with the smallest bias yielded unacceptable result on the other outcomes. Unacceptable values are in bold (i.e., Est bias ≥ 10%, SE bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if proportion of convergence failures ≥ 75%

most difficult situations to converge to admissible solutions. This is probably due to the fact that RFIML is a one-step approach that handles missing data directly in the estimation process, whereas MI-MVN separates the missing data

Table 7 Results for structural paths with moderately asymmetrical thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
N = 300, ncat=2									
RFIML	2.6	2.7	95	2.6	1.1	96	5.3	1.6	94
MI-MVN	3.3	3.3	94	4.0	5.1	95	--	--	--
MI-LV	**11.7**	1.2	96	**10.1**	5.0	96	**16.4**	**30.4**	97
MI-LOGIT	**10.4**	6.6	94	9.3	7.4	95	9.3	**12.2**	93
MI-RF	**11.9**	**22.3**	98	6.2	**11.4**	98	--	--	--
N = 600, ncat=2									
RFIML	2.0	2.7	95	2.5	2.4	95	4.1	2.1	95
MI-MVN	2.2	4.5	94	3.0	4.3	94	3.1	3.1	94
MI-LV	8.5	4.3	95	7.7	3.7	95	**11.6**	3.4	97
MI-LOGIT	8.1	6.5	94	7.0	6.4	94	9.9	**13.5**	92
MI-RF	8.6	**15.8**	98	3.2	**11.6**	98	**--**	**--**	**--**
N = 300, ncat=3									
RFIML	0.7	2.3	95	1.8	2.4	95	3.6	4.4	96
MI-MVN	0.9	3.4	94	2.5	0.9	95	4.4	4.4	96
MI-LV	4.4	3.0	95	4.5	1.0	95	6.6	3.7	96
MI-LOGIT	4.1	5.9	94	5.4	9.9	93	3.3	**11.5**	92
MI-RF	5.9	**18.0**	98	**21.2**	**28.1**	99	**25.6**	**32.4**	99
N = 600, ncat=3									
RFIML	1.2	3.4	95	1.6	2.9	96	1.9	2.5	95
MI-MVN	1.3	4.6	95	2.0	3.2	96	2.4	3.1	94
MI-LV	3.3	4.8	95	2.9	3.1	96	3.6	3.4	95
MI-LOGIT	3.1	6.3	94	3.3	**10.0**	93	2.7	**14.2**	91
MI-RF	4.2	**15.1**	98	**16.9**	**18.7**	99	**18.7**	**19.3**	99
N = 300, ncat=5									
RFIML	0.7	1.4	95	2.8	1.4	94	2.0	2.9	94
MI-MVN	0.9	3.3	94	3.1	2.2	94	2.6	3.0	95
MI-LV	3.6	3.3	95	5.2	2.9	95	3.5	2.6	95
MI-LOGIT	4.6	3.6	95	**15.3**	8.2	95	-1.1	**11.2**	91
MI-RF	5.7	**19.6**	98	**14.0**	**19.2**	99	**35.0**	**50.4**	**100**
N = 600, ncat=5									
RFIML	1.1	2.3	95	2.6	1.7	95	0.4	3.1	95
MI-MVN	1.2	3.0	95	2.5	3.1	95	1.2	3.3	95
MI-LV	2.9	3.3	95	3.7	2.8	95	1.4	4.6	94
MI-LOGIT	3.1	4.0	94	7.7	**13.4**	92	-1.4	**14.3**	91
MI-RF	4.3	**20.4**	98	**12.1**	**17.2**	98	**30.5**	**23.2**	99

The highlighted values are the smallest Est or SE biases in each design cell. The second smallest biases might also be highlighted if the method with the smallest bias yielded unacceptable result on the other outcomes. Unacceptable values are in bold (i.e., Est bias ≥ 10%, SE bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if proportion of convergence failures ≥ 75%

handling and model estimation processes. Thus, RFIML is more efficient than MI-MVN. In addition, although RFIML does not account for the ordinal nature of the data, it accounts for the nonnormality due to ordinal data. In comparison, MI-

Table 8 Results for structural paths with severely asymmetrical thresholds (*mp* = 30%)

	MCAR			MAR-Head			MAR-Tail		
	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC	Est Bias	SE Bias	CIC
				N = 300, ncat=2					
RFIML	7.0	2.5	93	7.6	4.3	94	--	--	--
MI-MVN	--	--	--	**11.3**	5.7	93	--	--	--
MI-LV	**30.3**	**19.7**	97	**25.7**	4.3	97	--	--	--
MI-LOGIT	**24.5**	0.3	96	**24.6**	2.4	96	--	--	--
MI-RF	--	--	--	9.0	**13.5**	98	--	--	--
				N = 600, ncat=2					
RFIML	4.1	1.8	94	5.6	0.7	95	--	--	--
MI-MVN	4.2	8.5	92	7.2	9.9	93	--	--	--
MI-LV	**21.5**	2.2	96	**19.6**	4.1	95	**40.5**	**137.6**	99
MI-LOGIT	**19.4**	6.2	94	**18.0**	5.3	95	--	--	--
MI-RF	**18.0**	**20.9**	99	5.0	**11.1**	98	--	--	--
				N = 300, ncat=3					
RFIML	1.6	2.5	94	3.4	3.3	95	--	--	--
MI-MVN	2.0	7.1	93	5.0	6.7	94	--	--	--
MI-LV	**15.0**	3.2	96	**13.7**	3.1	95	--	--	--
MI-LOGIT	**13.6**	6.5	95	**15.7**	4.8	95	--	--	--
MI-RF	**14.5**	**16.7**	98	3.8	**10.7**	97	--	--	--
				N = 600, ncat=3					
RFIML	1.4	1.5	95	3.6	0.3	95	4.9	3.6	91
MI-MVN	1.5	6.7	93	4.3	6.9	94	--	--	--
MI-LV	**12.1**	3.0	95	**11.3**	3.4	95	**17.9**	8.1	97
MI-LOGIT	**11.4**	6.0	94	**11.3**	4.9	95	1.6	**21.4**	**88**
MI-RF	**11.3**	**16.0**	98	2.1	**11.1**	97	--	--	--
				N = 300, ncat=5					
RFIML	0.7	2.4	95	2.5	4.1	96	2.6	3.8	94
MI-MVN	1.0	4.8	94	3.1	2.9	95	3.1	6.4	94
MI-LV	8.2	2.4	95	7.8	2.5	95	9.1	2.5	96
MI-LOGIT	8.5	6.3	94	**13.1**	4.2	95	-5.3	**15.2**	90
MI-RF	9.3	**18.2**	99	5.1	**16.3**	98	--	--	--
				N = 600, ncat=5					
RFIML	1.1	2.1	95	2.8	1.9	96	0.6	3.2	94
MI-MVN	1.2	6.2	94	2.9	3.5	95	1.4	4.4	94
MI-LV	7.0	3.6	95	6.4	3.7	95	6.5	3.5	95
MI-LOGIT	6.7	6.8	94	7.7	5.7	94	-3.4	**21.2**	**88**
MI-RF	7.8	**15.9**	98	3.3	**12.2**	98	--	--	--

The highlighted values are the smallest Est or SE biases in each design cell. The second smallest biases might also be highlighted if the method with the smallest Bias yielded unacceptable result on the other outcomes. Unacceptable values are in bold (i.e., Est bias ≥ 10%, *SE* bias ≥ 10%, or CIC < 90% or = 100%). Values are not reported if the proportion of convergence failures ≥ 75%

MVN only partially account for nonnormality (in the analysis stage, by using RML, but not in the imputation stage).

Question 2: How is the performance of each of the methods influenced by number of categories, asymmetry of thresholds, sample size, missing data proportion, and missing data mechanism*?*

RFIML was least impacted by the examined design factors. However, it may fail to converge or may generate large biases under the most difficult conditions (i.e., the distributions were severely asymmetrical, the missing data were MAR-tail, and the number of categories was two or three). RFIML could also produce substantial biases in the loading estimates for dichotomous data, if the distributions were severely asymmetrical. Otherwise, the performance of RFIML was stable and reliable in general.

The other methods were influenced by all design factors to some degree. MI-MVN and MI-LV generally performed well, except that they could fail when the distributions were asymmetrical, the missing data were MAR-tail, and the number of categories was small (i.e., two or three). MI-MVN was more sensitive to asymmetrical distributions and MAR-tail than was MI-LV, given that MI-MVN treated missing data as continuous and did not account for nonnormality when imputing the missing data.

As compared to the other methods, MI-LOGIT was impacted by the factors in different ways. In general, it worked adequately under MCAR or for dichotomous data, except for data with severely asymmetrical distributions. The performance of MI-RF was not satisfactory across all conditions for the outcomes considered in the study. One possible explanation for this poor performance of MI-RF is that, as a nonparametric method, MI-RF probably required a larger sample size (e.g., $N = 1{,}000$ in Doove et al., 2014) than those examined in the study. Specifically, MI-RF predicts missing data from donors that share similar properties with the incomplete cases. With a small sample size, it could have been difficult for MI-RF to find donors that were sufficiently homogeneous to those incomplete cases, resulting in either convergence problems or biased estimates. Future research will be warranted to study the conditions under which MI-RF may deliver satisfactory performance.

In terms of the general impact of a single factor, we found that the missing data mechanism was most influential on the performance of the examined methods. In other words, the methods performed quite differently under MAR-tail than under MCAR or MAR-head. In our study, MAR-tail was created in such a way that the values on the right tail of the distribution were most likely to be truncated. Thus, when the distribution was right-skewed, the least frequently observed values became sparser or completely missing, creating a challenging situation for the missing data techniques we examined.

Question 3: Which of the five methods performs best under the examined conditions?

For dichotomous data, RFIML appeared to be the best method among the five. For indicators with three categories, RFIML was also the best performer, closely followed by MI-MVN and MI-LV. RFIML was slightly better than MI-MVN by being easier to converge and slightly more robust to asymmetrical distributions and to MAR-tail. MI-LV combined with cat-DWLS also worked well for three-category data, except that it could result in substantially biased estimates for structural path coefficients when the item distributions were severely asymmetrical. For indicators with five categories, both RFIML and MI-LV combined with cat-DWLS seemed the best methods. They converged properly and produced accurate loading and structural path estimates under all conditions. MI-MVN also had comparable performance, except that it might produce biased loading estimates in the most challenging situation (small sample size, MAR-tail, and a distribution that was severely asymmetrical). The other two methods, MI-LOGIT and MI-RF, could produce biased results or fail to converge to a proper solution for various types of data examined in this study, with MI-RF being the worst.

Limitations and future directions

The findings and conclusions from this article are limited to the scope of the study. Several of these limitations are worth mentioning. First, we only examined one type of SEM model and focused on factor loadings and three latent paths. More studies will be needed to examine the performance of these methods for other types of parameters and other types of SEMs, such as growth curve models and mixture models.

Second, when generating the data, we assumed that missing data occurred on two indicators of all three latent variables. In such a model, with three latent variables playing different roles, it would be interesting to examine to what extent the location of the missing data might have affected the performance of the five methods. To shed some light on the impact of this factor, we conducted a small-scale simulation study based on a representative set of conditions (i.e., the distributions were moderately asymmetrical, the number of categories was three, and the missing data proportion was 30%). The same data generation and analysis model was used, but missing data occurred on two indicators of only one latent variable at a time. The results showed that only MI-RF was sensitive to the location of the missing data. Specifically, when the missing data only occurred among the indictors for η_1, MI-RF yielded extremely large positive biases in the path coefficient estimates. The biases decreased when the missing data occurred only among the indictors for η_2. When the missing data occurred only among the indictors for η_3, MI-RF yielded negative but acceptable Est biases for the structural path coefficients. Given that the other methods did not seem to be affected by the location of the missing data, varying the location of missing data would not alter our conclusions. More research, however, should be conducted to examine the reason why MI-RF performed differently with different missing data locations.

Third, the examined methods were implemented on the basis of the commonly used settings of the software packages. For the methods that had frequent convergence problems, modifying the settings may improve convergence. For example, one might try different optimization techniques, if they are available, or a larger number of iterations. For MI-RF, in particular, one might try different numbers of bootstrapped samples or different minimum numbers of donors. One might also use a less conservative rule (e.g., 50% imputations converged) to determine convergence for the imputation methods. Future research could be conducted to examine the existing strategies or to develop new strategies to improve the convergence of the missing data methods.

Fourth, we only considered the situation in which all the indicators were ordinal and had the same thresholds. In practice, continuous and ordinal missing data may coexist, and the thresholds or number of categories might also vary across ordinal items. It would be interesting to investigate whether the same conclusions would hold when different types of incomplete variables needed to be analyzed simultaneously.

Finally, we focused on the accuracy and precision of parameter estimates when evaluating the performance of the methods. In SEM, researchers often care about model fit information such as the chi-square test statistic and practical fit indices. Thus, it would be interesting to examine which method(s) lead to accurate statistical inference in terms of model fit (e.g., the chi-squared test statistic). We did record the Type I error rates of the chi-squared test statistic associated with RFIML (the detailed results can be requested from the authors). We found that the Type I error rates from RFIML were in general close to 5%, except they were inflated substantially (above 10%) with severely asymmetrical thresholds, especially when the ordinal data had only two or three categories. Again, these findings are limited to the conditions manipulated in the simulation. For example, the Type I error rates might be inflated when the threshold values, missing data mechanisms, or missing data proportions are substantially different across items (Savalei & Falk, 2014). It is important to note that valid model fit test statistics cannot be obtained from the multiple imputation methods investigated in this study, because there is no good way so far to pool the rescaled test statistics across imputations for cat-DWLS or RML (Enders & Mansolf, 2018). For this reason, we did not report fit indices in the Results section. Developing an appropriate method to pool the rescaled test statistics over imputations for RML or cat-DWLS would be a fruitful avenue for future research.

To conclude, among the five methods examined in the present study, RFIML performed the best for the outcomes considered in the study, and was most stable for a wide range of conditions. Aside from RFIML, MI-MVN and MI-LV also performed well most of the time, unless the conditions were extremely difficult (e.g., small number of categories, small sample size, severely asymmetrical distributions, and a large proportion of data missing on the tail of the distribution). Thus, MI-MVN and MI-LV can serve as good alternatives to RFIML, especially when MI is chosen to deal with missing data or RFIML fails to converge. Interested researchers might also try more than one method to see whether the different methods converge to similar results. Although MI-LOGIT and MI-RF are theoretically appealing, their empirical performance was not satisfying for the model and conditions examined in the present study. Thus researchers should be cautious about using them to deal with missing ordinal data in SEM.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Asparouhov, T., & Muthén, B. (2010). *Multiple imputation with Mplus* (Technical report). www.statmodel.com.

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

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*, 330–351.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*, 101–111. https://doi.org/10.1007/BF00162520

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, *72*, 92–104.

Enders, C. K. (2001a). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370. https://doi.org/10.1037/1082-989X.6.4.352

Enders, C. K. (2001b). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128–141.

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

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*, 76–93. https://doi.org/10.1037/met0000102

Ferrari, P. A., & Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research*, *47*, 566–589.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling*, *4*, 108–120.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*, 1–47.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling An overview and a meta-analysis. *Sociological Methods and Research*, *26*, 329–367.

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, 369–387. https://doi.org/10.1037/met0000093

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

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

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Muthén, L. K., & Muthén, B. (2012). *Mplus user's guide* (Version 7). Los Angeles, CA: Muthén & Muthén.

Palomo, J., Dunson, D. B., & Bollen, K. (2011). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 163–188). Amsterdam, The Netherlands: Elsevier.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation Statistical Computing, Vienna, Austria. Retrieved from www.R-project.org

Rhemtulla, M., Brosseau-Liard, P. E., & 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*, 354–373.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

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

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 473–489. https://doi.org/10.1080/01621459.1996.10476908

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 variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling*, *21*, 280–302. https://doi.org/10.1080/10705511.2014.882692

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

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, *179*, 764–774. https://doi.org/10.1093/aje/kwt312

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

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman and Hall/CRC Press.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377–399.

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. https://doi.org/10.1080/00273171.2015.1022644

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.