



A comparison of multiple imputation strategies to deal with missing nonnormal data in structural equation modeling

Fan Jia¹ · Wei Wu²

Accepted: 18 July 2022 / Published online: 29 August 2022
© The Psychonomic Society, Inc. 2022

Abstract

Missing data and nonnormality are two common factors that can affect analysis results from structural equation modeling (SEM). The current study aims to address a challenging situation in which the two factors coexist (i.e., missing nonnormal data). Using Monte Carlo simulation, we evaluated the performance of four multiple imputation (MI) strategies with respect to parameter and standard error estimation. These strategies include MI with normality-based model (MI-NORM), predictive mean matching (MI-PMM), classification and regression trees (MI-CART), and random forest (MI-RF). We also compared these MI strategies with robust full information maximum likelihood (RFIML), a popular (non-imputation) method to deal with missing nonnormal data in SEM. The results suggest that MI-NORM had similar performance to RFIML. MI-PMM outperformed the other methods when data were not missing on the heavy tail of a skewed distribution. Although MI-CART and MI-RF do not require any distribution assumption, they did not perform well compared with the others. Based on the results, practical guidance is provided.

Keywords Missing data · Nonnormality · Multiple imputation · Full information maximum likelihood · Predictive mean matching · Classification and regression trees · Random forest

Structural equation modeling (SEM) is a flexible and powerful analytical framework for testing complex multivariate relationships at the observed and/or latent variable levels (Bollen, 1989). It offers various estimation methods (e.g., maximum likelihood [ML] or weighted least squares estimation methods) and can handle different types of data (e.g., continuous or categorical). The current article aims to address the coexistence of two common factors that could affect the performance of SEM with continuous data and ML estimation. These two factors are nonnormality and missing data. When mishandled, either factor alone could cause biased results. Specifically, nonnormality alone could lead to biased standard error estimates (Browne, 1984; Chou et al., 1991; Fan & Wang, 1998; Finch et al., 1997; Olsson et al., 2000). Missing data alone could cause bias not only

in standard error estimates but also in parameter estimates (Enders, 2001a, 2001b).

Extensive research has addressed nonnormality with complete data (see Browne, 1984; Satorra & Bentler, 1994; Yuan & Hayashi, 2006). However, much less research has tackled the issue of nonnormality when data are incomplete, especially in the SEM framework. Missing data, ubiquitous in social and behavioral research, could add an extra layer of complexity on the top of nonnormality. There are different missing data techniques, such as full information maximum likelihood (FIML) and multiple imputation (MI; Enders, 2001a, 2010; Graham, 2009; Rubin, 1976, 1996; Schafer & Graham, 2002). For MI in particular, missing data can be imputed in different ways (e.g., through a parametric or nonparametric model). Consequently, researchers are faced with multiple options, and it is not clear which option(s) would work best.

The purpose of the current study is thus to provide a systematic comparison of several available MI strategies to deal with missing nonnormal data in the context of SEM using Monte Carlo simulation. Although some of the strategies have been examined in the past in univariate or regression analyses (see more details below), the findings

✉ Fan Jia
fjia3@ucmerced.edu

¹ Psychological Sciences, University of California, Merced,
5200 N. Lake Road, Merced, CA 95343, USA

² Indiana University-Purdue University Indianapolis,
Indianapolis, IN, USA

are not automatically generalizable to SEM, as SEM allows researchers to model relationships at the latent variable level. Thus, we believe that our study can provide new insights into the missing data analysis literature.

The rest of the article is organized as follows. We first provide the background information for the study, including missing data mechanisms, a general description and comparison of FIML and MI in dealing with normal missing data, and how the two approaches can be extended to accommodate missing nonnormal data. Because there are many ways to deal with missing nonnormal data when MI is used, we select a few available MI methods to study and provide the technical details for them. We then present the simulation study conducted to evaluate the performance of the selected MI methods compared with FIML in terms of parameter estimation. Finally, we discuss the results and limitations of the simulation study and provide practical recommendations to researchers on the use of these methods.

Background

Missing data mechanisms

Missing data mechanisms characterize the processes by which data become missing. Rubin (1976) developed a classification scheme with three missing data mechanisms. Suppose the probability of having missing data on a variable Y is not related to the missing values of Y itself after controlling for the other variables in the analysis. Then, the data on Y are said to be missing at random (MAR). Otherwise, the data are said to be missing not at random (MNAR). A special case of MAR is missing completely at random (MCAR), in which the probability of missing data on Y is unrelated to Y 's values or any other observed variables in the data set.

A general description of FIML and MI

Evidence shows that when missingness occurs in normal data, MAR, including MCAR, could be appropriately handled by modern missing data techniques such as full information maximum likelihood (FIML) and multiple imputation (MI; Enders, 2001a, 2010; Graham, 2009; Rubin, 1976, 1996; Schafer & Graham, 2002). FIML is a one-step approach that handles missing data simultaneously in the model estimation process. Specifically, FIML produced parameter estimates by iteratively maximizing the sum of N case-wise log-likelihood functions tailored to individual patterns of missing data (Enders, 2001a). MI, in comparison, typically involves three steps. It first generates multiply imputed data with missing values

filled in (imputation phase), then fits the hypothesized model to each of the imputed data sets (analysis phase), and pools the results across imputed data sets to produce the final results (Rubin, 1987; pooling phase). Under certain assumptions, such as MAR, multivariate normality, and a plausible imputation model, both FIML and MI were found to produce unbiased parameter and standard error estimates (Collins et al., 2001; Enders & Bandalos, 2001; Rubin, 1987; Savalei & Rhemtulla, 2012; Schafer & Graham, 2002).

Both methods have been widely used in practice. Although MI is more cumbersome to implement and can be less efficient than FIML (Yuan et al., 2012), there are unique benefits of using MI. First, MI is flexible, with a variety of imputation algorithms and imputation models available. Thus, it could potentially provide better treatments for nonnormal and nonlinear relationships among variables (Asparouhov & Muthén, 2010; White et al., 2011). Second, MI can incorporate many more auxiliary variables than FIML. Savalei and Bentler (2009) showed that incorporating many auxiliary variables into the FIML analysis could yield odd structures for certain covariance matrices, causing convergence problems. In contrast, MI incorporates auxiliary variables in the imputation phase only, so it is less likely to cause problems in the analysis phase. In addition, MI creates complete data sets; thus, statistical methods that work only with complete data can be applied. For example, FIML cannot be used to deal with item-level missing data when the items are to be parceled and will not be directly included in the analysis model (Little et al., 2013). Although a two-stage ML could be used in this situation (TSML; Savalei & Rhemtulla, 2017), it requires sophisticated matrix algebra and has not been automated in standard software packages. Thus, most researchers do not have access to this approach. On the other hand, MI is widely available and can easily handle such situations by generating complete item scores before parceling (Enders & Mansolf, 2018; Gottschall et al., 2012).

Extending FIML and MI to account for nonnormality

Past research had found that nonnormality could lead to underestimated standard errors when FIML was used (Enders, 2001b). Methods to correct the bias have thus been developed. The most popular correction method is known as robust FIML (a.k.a., RFIML; Savalei & Falk, 2014), which uses a sandwich-like covariance matrix based on the results from FIML (Yuan & Bentler, 2000; Yuan & Hayashi, 2006). Research has found that RFIML performed well under MCAR or MAR, except when MAR data occurred mainly on the heavy tail of a distribution, and the proportion of

missing data was large (e.g., 30%, Enders, 2001b; Savalei & Falk, 2014).

There is no consensus on how to best account for nonnormality in the imputation stage of MI. Our literature review suggests that four types of MI strategies have been examined for missing nonnormal data: (1) MI based on the assumption of multivariate normality, (2) normalizing the data using a transformation method first and then using the first approach to impute, (3) MI based on generalized parametric families that account for some specific nonnormal distributions, and (4) MI based on semi-parametric or nonparametric models that do not have distributional assumptions. The four types of strategies are explained below.

The first strategy ignores the nonnormality of the data (MI-NORM). Past research found that the robustness of MI-NORM to the violation of normality varied across different parameter estimates. Demirtas et al. (2008) examined MI-NORM by generating continuous data from a broad range of distributions, including normal, t , Laplace, and Beta distributions. They found that MI-NORM accurately estimated means and regression coefficients with these nonnormal distributions. However, the variance parameters could be biased, particularly when the sample size was small (e.g., $N = 40$). Other parameters that rely more on the tails of a distribution, such as extreme quantiles, were sensitive to nonnormality (Demirtas et al., 2008; Schafer, 1997). Yuan et al. (2012) also concluded that nonnormal data could severely impact the estimates of variance-covariance parameters.

The second strategy uses transformation to normalize nonnormal data before imputation (MI-TRANS). Transformation is a traditional approach for dealing with nonnormal data. Many transformation functions, such as log, exponential, square root, Box-Cox, and non-parametric, have been used in the past to reduce the skewness of a distribution (Allison, 2000; Honaker et al., 2011; Lee & Carlin, 2017; Schafer & Graham, 2002; von Hippel, 2005). After transformation, MI-NORM can then be used to fill in the missing values in the transformed metric. These imputed values may be converted back to their original scales before the target analysis. Although the transformation method seems straightforward, it is usually not recommended, as it is often challenging to determine the best transformation function. If a wrong/suboptimal transformation method is used, it could hurt the imputation by distorting the relationships between the variable and the others, resulting in biased imputed values and follow-up analyses (von Hippel, 2013).

The third strategy is to impute continuous missing values based on generalized parametric families for continuous data. A parametric family is a family of distribution functions whose forms depend on a set of parameters. Rather than assuming a normal distribution, a generalized parametric family allows for various data distributions, making the imputation more flexible. In the last two decades, several

generalized parametric families have been considered in MI, such as Tukey's gh distribution (Demirtas & Hedeker, 2008; He & Raghunathan, 2009), t , lognormal, Beta, and Weibull distributions (Demirtas & Hedeker, 2008), Fleishman's power polynomials (Demirtas & Hedeker, 2008), and the generalized lambda distribution (Demirtas, 2009). Focusing on univariate distributions and MCAR data, these approaches outperformed the first strategy in estimating quantiles of continuous data; however, they did not show advantages in estimating the means of continuous variables. Like the transformation approach, it is also challenging to determine which distribution will best fit the data. Consequently, additional bias could be introduced if a wrong distribution is used.

The fourth strategy is to impute missing nonnormal data based on semi-parametric or nonparametric methods. In the MI literature, semi-parametric methods such as local residual draws (LRD) and predictive mean matching (PMM), and nonparametric methods such as classification and regression trees (CART) and random forest (RF), have been evaluated for imputing missing nonnormal data. We refer to MI with these four methods as MI-LRD, MI-PMM, MI-CART, and MI-RF, respectively. He and Raghunathan (2009) found that MI-LRD and MI-PMM performed well for estimating marginal means, proportions, and regression coefficients when the error distribution was uniform or moderately skewed. However, both seemed to have difficulty handling extreme values and performed poorly under severe nonnormality. Lee and Carlin (2017) showed that MI-PMM could produce acceptable results for estimating means and regression coefficients with $N = 1000$. MI-CART and MI-RF were found to outperform parametric and semi-parametric imputation methods, such as MI with logistic regression and MI-PMM, in dealing with nonlinear relationships of categorical variables (Doove et al., 2014; Shah et al., 2014). Hayes and McArdle (2017) examined the performance of MI-CART and MI-RF in estimating an interaction effect under various missing data generating mechanisms and distribution conditions (both normal and nonnormal). Their results show that MI-CART and MI-RF were superior to MI-NORM under certain conditions, particularly when the sample size was large ($N = 500$ – 1000). However, MI-CART and MI-RF performed poorly with small sample sizes and when the missingness had nonlinear relationships with other variables, regardless of the degree of nonnormality.

The current study is designed to systematically evaluate different MI methods to deal with missing nonnormal data in SEM in comparison to RFIML. Our goal is to investigate to what extent the methods can recover parameters in SEM under nonnormality. We hope that this investigation can provide valuable insights for future research.

To keep the scope of our study manageable, we did not include all available methods but those we believed

promising or likely to be used by practical researchers. We include MI-NORM to further examine its robustness to nonmorality in the context of SEM. We expect MI-NORM to have some robustness to nonnormality when it is not extreme. We also include MI-PMM, MI-CART, and MI-RF because they rely less on distributional assumptions and have shown some good performance in regression analyses. We omitted MI-LRD because it performs similarly to MI-PMM and has limited software implementation (He & Raghunathan, 2009; Morris et al., 2014). We did not consider the transformation or the generalized parametric family strategies because of the limitations mentioned above. We included RFIML as a comparison approach. To help researchers better understand how the selected MI methods work, we provide the technical details below, including how the data are imputed in the imputation phase, the estimation method used in the analysis phase, and how the parameter estimates are pooled across imputations.

Selected MI methods for missing nonnormal data

Normal-theory-based imputation (MI-NORM)

MI-NORM ignores nonnormality. It is typically implemented using either of two algorithms: joint modeling (JM; Schafer, 2010) and expectation-maximization with bootstrapping (EMB; Honaker et al., 2011). Both algorithms fill in missing values on multiple incomplete variables simultaneously and are theoretically equivalent. In the current study, we used the EMB algorithm for convenience. Briefly speaking, EMB generates a large number of bootstrapped samples first (Efron, 1979) and then uses the EM algorithm to obtain the maximum likelihood estimates of the mean and covariance matrix for the variables included in an imputation model for each bootstrapped sample. The EM estimates are then treated as a random draw of the imputation parameters and used to impute the missing data.

MI with semi-parametric or nonparametric models

MI with semi-parametric or nonparametric models is implemented using a so-called fully conditional specification (FCS) algorithm, also known as MI by chained equations (MICE; van Buuren et al., 2006; van Buuren & Groothuis-Oudshoorn, 2011). Unlike JM or EMB, FCS imputes missing data on a variable-by-variable basis without relying on a multivariate normal distribution.

To illustrate a typical FCS process, let y_1, y_2, \dots, y_p be the p variables that need to be imputed and $\theta_1, \theta_2, \dots, \theta_p$ be the parameters that describe the distributions of the p variables

(Mistler & Enders, 2017; van Buuren, 2018). Then the FCS at the t^{th} iteration can be described as follows.

$$\begin{aligned}
 &\theta_1^{(t)} \text{ from } P\left(\theta_1 | y_1^{obs}, y_2^{(t-1)}, \dots, y_p^{(t-1)}\right) \\
 &y_1^{miss(t)} \text{ from } P\left(y_1^{miss} | y_1^{obs}, y_2^{(t-1)}, \dots, y_p^{(t-1)}, \theta_1^{(t)}\right) \\
 &\theta_2^{(t)} \text{ from } P\left(\theta_2 | y_2^{obs}, y_1^{(t)}, y_3^{(t-1)}, \dots, y_p^{(t-1)}\right) \\
 &y_2^{miss(t)} \text{ from } P\left(y_2^{miss} | y_2^{obs}, y_1^{(t)}, y_3^{(t-1)}, \dots, y_p^{(t-1)}, \theta_2^{(t)}\right) \\
 &\dots \\
 &\theta_p^{(t)} \text{ from } P\left(\theta_p | y_p^{obs}, y_1^{(t)}, y_2^{(t)}, \dots, y_{p-1}^{(t)}\right) \\
 &y_p^{miss(t)} \text{ from } P\left(y_p^{miss} | y_p^{obs}, y_1^{(t)}, y_2^{(t)}, \dots, y_{p-1}^{(t)}, \theta_p^{(t)}\right)
 \end{aligned} \tag{1}$$

Because FCS imputes on a variable-by-variable basis, it has the flexibility to tailor the imputation model according to the nature of each incomplete variable. In other words, the imputation models can vary across variables. It can also accommodate a wide variety of imputation models, such as the semi-parametric and nonparametric methods considered in the current study.

MI with predictive mean matching (MI-PMM)

The idea of MI-PMM is to impute each missing value by randomly drawing a value from its nearest observed neighbors (also called candidate donors) in terms of the predicted value of the same variable (Little, 1988). Different versions of PMM have been developed by varying one or some of the computational details. First, there are different ways to estimate the parameters in a predictive model (van Buuren, 2018). Using linear regression as an example, the parameters could be (i) least square parameters, (ii) random parameter values drawn from their posterior distributions (Bayesian approach), or (iii) least square parameters computed from a bootstrap sample taken from the observed data. The first method ignores the sampling variability of the parameters and tends to produce biased results, especially when there are only a small number of predictors (Heitjan & Little, 1991; van Buuren, 2018). This problem can be alleviated by using the Bayesian approach or bootstrapping (Koller-Meinfelder, 2010).

Second, there are different matching methods. For example, matching can be done based on the distance between the predicted values or random draws from the posterior distribution of the observed or the posterior distribution of the missing data (e.g., type 1 matching or type 2 matching; van Buuren, 2018). Lee and Carlin (2017) found that type 1 matching outperformed type 2 matching for PMM in estimating marginal means and regression coefficients under various types of nonnormality. Finally, the number of candidate donors (denoted as d ; see

Andridge & Little, 2010, for more details) can vary within a reasonable range. A general rule is that d should not be too small (results in little variability across imputed data sets) or too large (increases the chance of poor matches). Common values for d are 3, 5, and 10 (van Buuren, 2018; Morris et al., 2014). More research, however, is needed to establish a guideline for specifying d .

MI-PMM preserves the original distributions; thus, it has the potential to deal with missing nonnormal data. PMM is semi-parametric because it does not require a parametric model to define the distribution of missing data; however, a parametric predictive model, usually a linear regression model, is still needed to determine the candidate donor pool (Heitjan & Little, 1991; Schenker & Taylor, 1996). Although MI-PMM has been found to work well with nonnormal data in various scenarios (e.g., Di Zio & Guarnera, 2009; Kleinke, 2017; Morris et al., 2014), it has not been examined in the context of SEM.

MI with classification and regression trees (MI-CART)

CART is a recursive partitioning method. CART stands for classification trees or regression trees, depending on whether the response variable is categorical or continuous (Breiman et al., 1984). Unlike traditional regression and classification methods, CART predicts a variable by successively splitting a data set based on one other variable at a time. The resulting subsets of data become more homogeneous with each split (Breiman, 2001). Because this splitting procedure and resulting subsets can be represented as a tree structure, these subsets are also referred to as leaves of the tree (James et al., 2013). For the observations within the same leaves, the mean of the response values is then used as the predicted value. When CART is used for imputation, the missing values are imputed based on these predicted values.

MI with random forest (MI-RF)

RF is an extension of CART. A single classification tree from CART is often prone to sample noise, limiting its generalizability (Doove et al., 2014; Kirasich et al., 2018). RF solves the problem by assembling results across many trees. Briefly speaking, RF generates multiple samples first based on the original data using resampling approaches such as bootstrapping and then creates a tree for each sample. The predicted values from multiple trees are averaged to create the final prediction. In the same way as MI-CART, the final predicted values are used for imputing missing values. As mentioned above, CART and RF do not rely on distributional assumptions or parametric models, so they both have the potential

to accommodate missing nonnormal data and nonlinear relationships (Doove et al., 2014; Shah et al., 2014).

Estimation methods used in the analysis phase

After imputation, the target analysis (SEM in this case) is applied to each imputed data set. Because the data are nonnormal, a robust ML estimator can be used to correct the bias in standard errors (SEs) due to nonnormality. Note that there are different versions of robust ML estimators. The one used in this study adopts the “sandwich” approach as in RFIML to adjust the standard errors (a.k.a., robust SEs or MLR SEs; Yuan & Bentler, 2000; Yuan & Hayashi, 2006). Specifically, the robust SEs are obtained from the asymptotic covariance matrix of the parameter estimates ($\hat{\theta}$):

$$nCov(\hat{\theta}) = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} \quad (2)$$

where the “bread” part $\mathbf{A} = -\sum_{i=1}^n \frac{\partial^2 \log l(\theta)_i}{\partial \theta \partial \theta'}$, and the “meat” part $\mathbf{B} = \sum_{i=1}^n \left(\frac{\partial \log l(\theta)_i}{\partial \theta} \right) \times \left(\frac{\partial \log l(\theta)_i}{\partial \theta} \right)'$. The derivatives in both \mathbf{A} and \mathbf{B} are evaluated at $\hat{\theta}$, and $l(\theta)_i$ is the log-likelihood for case i . When data are normal, $\hat{\mathbf{A}} = \hat{\mathbf{B}}$, and Eq. 2 can be reduced to $nCov(\hat{\theta}) = \hat{\mathbf{A}}^{-1}$, which is the ML asymptotic covariance matrix under the normality assumptions (i.e., the inverse of the observed information matrix; Yuan & Bentler, 2000). Lai (2018) found that this correction method was superior to other analytical techniques in estimating SEs and confidence intervals (CIs) of SEM parameters with complete nonnormal data.

Pooling procedures

In the pooling phase, outcomes from the analysis, such as point estimates and standard errors, are pooled into the final results, following Rubin's rules (Rubin, 1987). The final point estimates are obtained by taking the average across the imputations,

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m \quad (3)$$

where M denotes the number of imputed data sets, and $\hat{\theta}_m$ represents the parameter estimates for the m^{th} imputation. The pooled standard errors are the squared root of the sum of within-imputation variance (\mathbf{V}_W), between-imputation variance (\mathbf{V}_B), and a correction factor (\mathbf{V}_B/M),

$$\mathbf{V}_T = \mathbf{V}_W + \mathbf{V}_B + \frac{\mathbf{V}_B}{M} \quad (4)$$

where \mathbf{V}_W is computed as the average of the squared standard errors (for nonnormal data, they are robust standard errors)

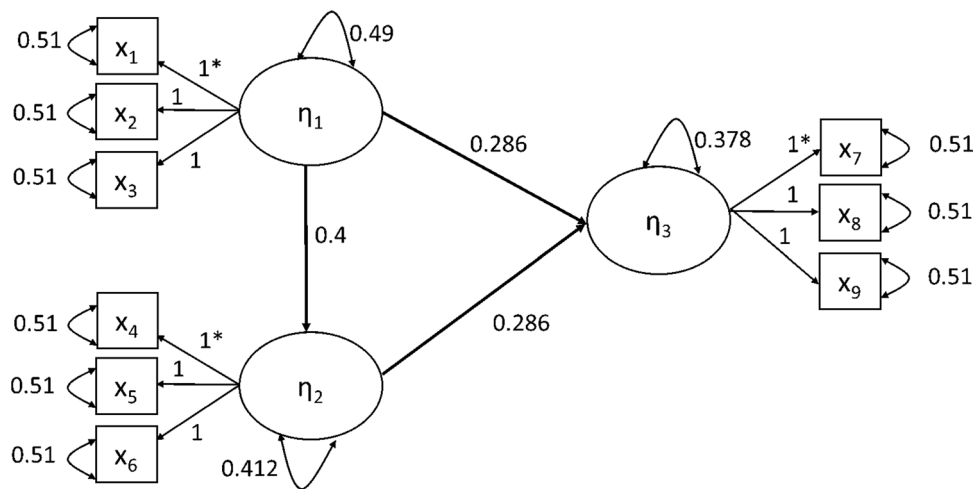


Fig. 1 The structural equation model for data generation

across M imputed data sets, and V_B is the variance of the M parameter estimates.

Simulation study

As mentioned above, this study examined five methods for dealing with missing nonnormal data in SEM: multivariate-normality-based MI (MI-NORM), MI with predictive mean matching (MI-PMM), MI with classification and regression trees (MI-CART), MI with random forest (MI-RF), and RFIML. The performances of these methods are compared in estimating model parameters.

Design

Data were generated based on a three-factor SEM model, in which factor η_2 was predicted by η_1 , and η_3 was predicted by both η_1 and η_2 . This type of model is commonly seen in the SEM literature (e.g., Bollen, 1989; Palomo et al., 2011). The data generation model is shown in Fig. 1. The population values were the same as in Fig. 1 in Enders (2001b). The values of the structural paths among the three variables were 0.4 ($\eta_1 \rightarrow \eta_2$), 0.286 ($\eta_1 \rightarrow \eta_3$) and 0.286 ($\eta_2 \rightarrow \eta_3$). Each factor was indicated by three manifest variables. The first factor loading of each factor was fixed to 1 for identification purposes, and the other loadings were all set to 1. The variance of η_1 was set to 0.490, and the residual variance of η_2 and η_3 was set to 0.412 and 0.378, respectively. The residual variance on the indicators was all set to 0.51. The indicators are all standard normal. We manipulated several factors in the data-generating process to create a wide range of conditions, including the degree of nonnormality, missing data proportion and mechanism, and sample size. The sample size (N) was set at two levels: small (300) and large (600).

Degree of nonnormality

We varied the degree of nonnormality at three levels. Nonnormality is typically reflected by the third standardized moment (skewness) and the fourth moment (kurtosis) around the mean of a distribution. The skewness describes the asymmetry of a distribution about its mean, and the kurtosis measures the “peakedness” of a distribution. For a univariate normal distribution, the skewness is 0 and the kurtosis is 3 (or excess kurtosis = kurtosis – 3 = 0). In this study, nonnormal continuous data were generated following the method proposed in Vale and Maurelli (1983) and Fleishman (1978). The levels of nonnormality were specified using three combinations of univariate skewness (S) and excess kurtosis (K): mild ($S = 1.5$, $K = 3$), moderate ($S = 2$, $K = 7$), and severe ($S = 3$, $K = 21$). The corresponding approximate multivariate kurtoses (Mardia, 1970) were 143, 187, and 314, respectively. These levels of nonnormality were reflective of the data observed in applied research (Curran et al., 1996) and were close to those used in Enders (2010) and Savalei and Falk (2014). For simplicity, all manifest variables had the same degree of nonnormality under each condition. The correlation matrix of the nonnormal data was consistent across the different levels of nonnormality.

Missing data conditions

We varied the missing data proportion (MP) as well as the missing data mechanisms. There are two levels of MP : small (15%) and large (30%). These levels are selected based on previous simulation studies and typical cases in SEM. We considered three missing data mechanisms: MCAR, MAR-Head, and MAR-Tail. Missing data were imposed on only two indicators for each factor (specifically, missing values

occurred on x_1 , x_2 , x_4 , x_5 , x_7 , and x_8 , see Fig. 1). These missing data were created as follows.

MCAR data were generated by randomly deleting the desired proportion of values on each incomplete variable. MAR data were imposed through the following procedure. We first ranked the values of each of the three fully observed variables (x_3 , x_6 , and x_9). We then used the percentile ranks to determine the probabilities of missingness on the other two manifest variables of the same latent factor. For MAR-Head, the probability of having missing data on x_1 was equal

to 1 minus the percentile rank% of x_3 . For instance, a case with the largest value on x_3 (100th percentile) would have a 0% chance of missing the value on x_1 , while the chance of having missing data on x_1 for a case at the 70th percentile on x_3 would be 30%. That is, the probability of missing an observation on x_1 increased as the x_3 value decreased. For each value on x_3 , we compare its probability of missingness with a randomly drawn value from a uniform distribution ranging from 0% to 100%. If the random value was greater than the probability, the case would have missing data on x_1 .

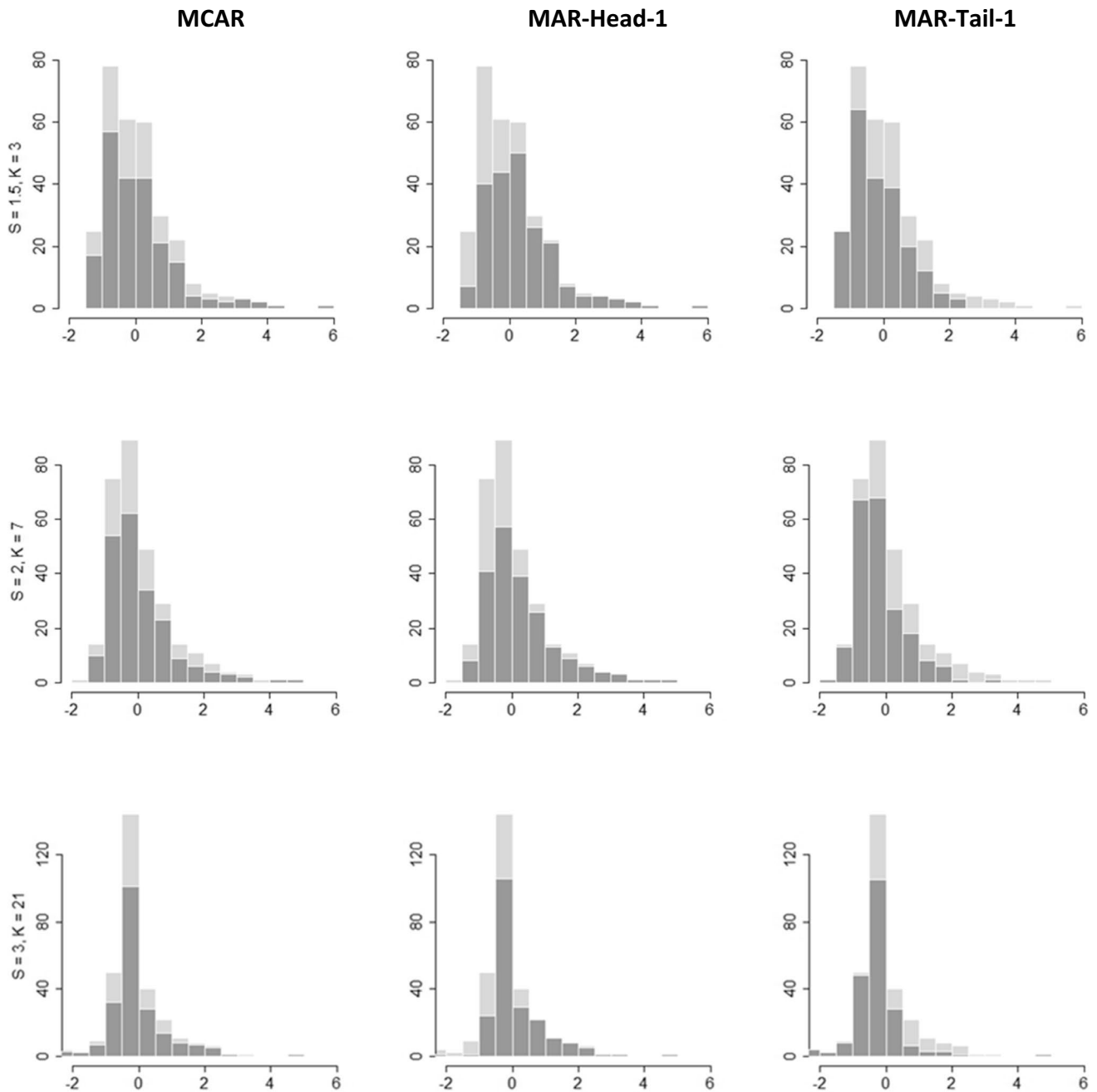


Fig. 2 Distributions of x_1 (continuous) for one replication with $N = 300$ before (light gray) and after (dark gray) imposing 30% missing data

This is done sequentially for other cases until the planned proportion of missing data (15% or 30%) is reached. Missing data on x_2 were imposed following the same rationale. Similarly, missing data on x_4 and x_5 were created based on the percentile ranks of x_6 , and missingness on x_7 and x_8 was determined by the percentile ranks of x_9 . Because all manifest variables were positively skewed and positively correlated, more missing data were imposed on the head of the distributions.

MAR-Tail data were generated in a similar fashion, except that the probability of having missing data on x_1 was simply equal to the percentile rank% of x_3 . For example, a case with the largest value on x_3 (100th percentile) would have a 100% chance of missing the value on x_1 , while the chance of having a missing value on x_1 would be 70% for a case at the 70th percentile on x_3 . That is, the probability of missing a value on x_1 decreased as the x_3 value decreased. Same rule applied to other manifest variables. Under MAR-Tail, more missing data were imposed on the heavy tail of the distributions.

To demonstrate the distributions of the incomplete variables under the different missing data mechanisms and levels of nonnormality, we selected one incomplete variable and visualized its distributions in Fig. 2 under various degrees of nonnormality and missing data mechanisms for one replication with $N = 300$ and $MP = 30\%$. As shown in Fig. 2, nonnormality could be less detectable with missing data in the tail than in the head of the distribution. For example, the average skewness and kurtosis based on complete cases decreased by 45% and 20%, respectively, after imposing missing data in the heavy tail of the distribution with severe nonnormality. When missing data occurred in the head of the distribution, in comparison, the skewness and kurtosis changed less (−8% and −13%, respectively).

In sum, there are 36 conditions (2 sample sizes \times 3 degrees of nonnormality \times 2 missing data proportions \times 3 missing data mechanisms). We generated 1000 replicated samples in each condition.

All data were generated through R (R Core Team, 2017) using the function `gen.nonnormal()` developed by Zopluoglu (2013). RFIML was implemented in lavaan (MLR; Rosseel, 2012), in which the convergence threshold (relative tolerance) is set at 10^{-10} . MI-NORM was implemented using the R package Amelia, which employs the EMB algorithm (Honaker et al., 2011). The convergence threshold for EM was equal to 10^{-4} . The FCS methods were implemented using the R package mice (van Buuren & Groothuis-Oudshoorn, 2011). The burn-in iterations were set at 20 for MI-PMM and MI-RF based on a preliminary simulation study using one of the most challenging conditions ($N = 300$, 30% missingness, MAR-Tail, and severe nonnormality). The number of donors for MI-PMM is set to 5 following Morris et al. (2014). For MI-RF, a minimum leaf size of 5 was used to create regression trees (Liaw & Wiener, 2002). The

number of bootstrap samples in RF (i.e., the number of trees) was set to 10 (Doove et al., 2014). The MLR estimator in the lavaan package was used to analyze all the imputed data. For the imputation methods, 50 imputed data sets were generated following the guidelines developed by White et al. (2011). A replication was deemed converged if the model converged for all 50 imputed data sets.

Evaluation criteria

The performance of the examined methods was evaluated based on relative bias in parameter estimates (Est bias), relative bias in standard errors (SE bias), and confidence interval coverage (CIC) rates. The relative bias of a parameter θ is calculated as the difference between the average parameter estimate across replications within a design cell ($\bar{\theta}_{est}$) and the population value (θ_0), divided by the true population value.

$$\text{Est bias} = \frac{(\bar{\theta}_{est} - \theta_0)}{\theta_0} \times 100\% \quad (5)$$

Following Muthén et al. (1987), we used $\pm 10\%$ as the acceptable cutoff points.

SE bias measures the accuracy of standard errors, which can be calculated as follows.

$$\text{SE Bias} = \frac{(\overline{\text{SE}} - \text{ESE})}{\text{ESE}} \times 100\% \quad (6)$$

where $\overline{\text{SE}}$ 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). We considered a SE bias acceptable if its absolute value was less than 10% (Hoogland & Boomsma, 1998).

The CIC of a model parameter is estimated as the percentage of replications in which the 95% CIs covered the population value. Ideally, a CIC should equal 95%. Following Bradley's (1978) "liberal criterion," we consider a CIC acceptable if it is between 92.5% and 97.5%.

Results

As mentioned above, the methods are compared on three outcomes: Est bias, SE bias, and CIC. Given that the structural path coefficients, capturing predicting relationships at the latent variable level, are often of the most interest to researchers, the primary parameters were the three structural path coefficients ($\beta_{2,1}$, $\beta_{3,1}$ and $\beta_{3,2}$). The secondary parameters are factor loadings and factor variances. For ease of

Table 1 Biases (%) in parameter estimates under MCAR

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	1.59	1.53	3.18	1.83	2.33	2.05	0.94	4.21	3.33	1.93	3.23	2.80	5.62	3.22	3.65
MI-MVN	-0.10	0.63	4.88	1.24	2.61	3.04	1.18	2.64	2.06	2.50	5.62	4.36	4.08	3.70	4.26
MI-PMM	1.66	2.16	3.27	3.23	0.49	3.00	1.86	5.10	5.61	-1.93	6.83	4.12	7.67	7.46	-5.03
MI-CART	1.28	1.55	3.14	3.37	-0.39	1.86	1.28	4.29	5.03	-2.20	5.29	3.69	7.12	6.55	-3.13
MI-RF	9.01	6.18	7.65	10.33	-13.61	10.16	6.83	9.70	12.80	-15.91	13.81	10.29	12.66	14.50	-17.79
<i>MP = 30%, N = 600</i>															
RFIML	1.15	0.46	1.50	0.87	0.84	1.34	0.92	1.86	1.16	1.01	2.37	0.60	3.95	2.48	1.05
MI-MVN	1.03	1.56	1.30	0.93	0.36	-1.00	-0.26	3.02	0.57	2.38	2.83	0.66	1.41	1.05	3.76
MI-PMM	1.15	0.73	1.39	1.62	-0.14	1.61	1.60	2.26	2.12	-0.54	4.32	2.44	4.93	4.86	-3.62
MI-CART	0.86	0.63	1.29	1.53	-0.68	1.11	1.23	1.86	2.01	-0.77	2.82	1.18	4.84	3.79	-1.98
MI-RF	6.49	4.13	5.08	7.48	-11.91	7.08	5.08	6.02	8.10	-12.67	9.90	5.86	8.48	10.38	-14.74
<i>MP = 15%, N = 300</i>															
RFIML	1.22	1.30	3.28	1.46	1.60	2.03	1.04	3.17	2.75	1.25	2.14	1.79	5.33	2.59	2.55
MI-MVN	-0.21	0.57	2.83	0.88	1.69	2.37	1.47	2.30	2.24	0.24	3.18	2.29	3.26	2.84	3.84
MI-PMM	1.40	1.63	3.07	2.17	0.61	2.44	1.28	3.31	3.74	-0.29	3.31	2.74	6.21	4.62	-1.18
MI-CART	1.50	1.38	3.21	2.47	-0.01	2.19	1.21	2.96	3.83	-0.68	3.13	2.46	5.79	4.48	-0.65
MI-RF	4.92	3.54	5.40	5.35	-5.93	6.07	3.67	5.62	7.08	-7.25	7.29	5.31	7.99	7.88	-7.35
<i>MP = 15%, N = 600</i>															
RFIML	0.99	0.45	1.16	0.60	0.64	1.17	0.95	1.53	0.90	0.62	1.81	0.45	3.42	1.97	1.03
MI-MVN	0.24	1.20	1.10	0.53	0.54	-0.74	-0.19	2.03	0.41	1.47	2.51	0.15	0.91	1.03	2.07
MI-PMM	1.10	0.65	1.14	0.96	0.09	1.41	1.39	1.62	1.32	-0.06	2.76	1.48	4.05	3.02	-1.07
MI-CART	1.09	0.60	1.01	1.11	-0.41	1.04	0.76	1.49	1.34	-0.22	2.15	0.31	3.89	2.84	-0.31
MI-RF	3.77	2.13	2.70	3.64	-5.49	3.96	2.66	3.29	4.16	-5.95	5.41	2.88	5.34	5.71	-6.56

Biases lower than -10% or larger than 10% are in boldface

presentation, we reported results for a representative parameter for each of them: the factor loading of x_3 (λ_3) and the variance of factor η_1 ($\varphi_{1,1}$). The two parameters were chosen because their result patterns were consistent with the others. In addition, given that the MCAR and MAR-Head results were similar, we only report MCAR and MAR-Tail results below. The results for all parameters under all missing data mechanisms are available upon request. In the following, we organized the results by evaluation criterion, missing data mechanism, and type of parameter.

Bias in parameter estimates

MCAR

As shown in Table 1, all methods produced negligible bias ($|\text{bias}| < 10\%$) for all types of parameters, except for MI-RF. MI-RF was acceptable when the missing data proportion was low (i.e., $MP = 15\%$). However, it yielded biased estimates for all types of parameters with a larger amount of missing data. Specifically, with 30% missing data, MI-RF

yielded negatively biased ($< -10\%$) factor variance regardless of the sample size and degree of nonnormality, and positively biased ($> 10\%$) loadings and path coefficients when the sample size was small ($N = 300$).

MAR-Tail

Path coefficients As shown in Table 2, when missing data occurred mainly on the heavy tail of the distribution, both non-parametric methods (MI-CART and MI-RF) produced over-estimated path coefficients ($|\text{bias}| > 10\%$), although estimates from MI-CART were slightly less biased. All other methods yielded negligible bias for path coefficients, except that RFIML produced slightly larger bias under severe nonnormality.

Loading and factor variance The design factors seemed to impact the parameter estimates for the representative factor loading and variance differently. The factor loading estimate mainly was influenced by missing data proportion. When $MP = 15\%$, all methods produced negligible bias (under 10%), with only a few exceptions. However, when $MP = 30\%$, almost all methods failed, except

Table 2 Biases (%) in parameter estimates under MAR-Tail

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	-5.43	-1.09	-1.46	10.12	7.01	-8.15	-5.28	-3.09	10.98	13.29	-8.46	-4.70	-5.16	13.12	14.12
MI-MVN	-0.91	-1.52	1.76	11.07	8.64	-3.75	-3.73	0.54	13.35	11.86	-3.52	2.23	0.43	17.23	8.64
MI-PMM	2.74	7.52	6.29	20.04	-14.84	3.34	3.27	7.73	21.95	-17.43	9.07	7.34	15.39	26.25	-26.72
MI-CART	17.62	19.04	17.94	8.93	-26.89	19.02	13.86	22.98	9.51	-28.00	23.54	17.85	22.50	12.92	-33.92
MI-RF	26.89	22.13	22.61	20.62	-46.90	30.83	19.87	25.61	21.30	-49.12	33.97	24.31	30.02	27.43	-53.77
<i>MP = 30%, N = 600</i>															
RFIML	-6.49	-5.58	-5.50	8.34	5.96	-11.43	-10.00	-6.94	9.45	11.86	-10.74	-6.32	-7.30	14.62	7.94
MI-MVN	-3.85	-2.93	-1.47	10.42	3.68	-6.84	-5.71	-3.26	11.17	9.14	-5.90	-2.18	-4.51	15.80	6.05
MI-PMM	-1.99	-1.13	-0.19	13.07	-9.94	-1.34	-1.78	2.62	16.18	-13.89	8.28	9.42	6.21	22.52	-26.00
MI-CART	11.62	8.74	9.34	2.66	-20.41	12.42	8.57	10.44	4.36	-23.37	19.19	17.15	14.90	8.21	-29.59
MI-RF	22.72	14.98	15.99	14.94	-43.76	23.45	14.23	19.09	16.31	-46.30	31.41	24.81	21.08	22.59	-52.67
<i>MP = 15%, N = 300</i>															
RFIML	-4.00	-0.40	-1.96	6.44	7.05	-7.10	-4.70	-3.28	8.26	12.24	-8.71	-6.09	-6.67	8.88	20.18
MI-MVN	-3.63	-3.08	1.34	6.44	7.80	-6.10	-4.54	-2.12	8.23	13.27	-7.26	-2.34	-4.85	12.02	17.20
MI-PMM	-0.66	3.04	2.96	9.83	-7.01	1.13	1.11	4.10	13.19	-10.04	6.11	4.69	7.34	15.25	-16.70
MI-CART	10.80	12.23	9.88	1.27	-20.80	13.20	11.48	12.09	3.06	-24.23	19.68	14.29	15.60	3.15	-29.13
MI-RF	17.01	15.49	13.22	4.19	-33.59	19.51	14.33	15.70	5.00	-36.33	26.34	19.17	19.67	7.80	-42.70
<i>MP = 15%, N = 600</i>															
RFIML	-4.97	-3.51	-3.27	5.45	6.32	-9.14	-8.17	-6.64	6.75	12.41	-11.47	-8.33	-7.90	9.83	16.72
MI-MVN	-4.24	-2.23	-1.21	6.42	4.77	-7.43	-6.04	-5.05	7.65	11.17	-8.29	-5.41	-8.17	10.49	14.72
MI-PMM	-1.72	-0.60	0.10	7.20	-5.76	-1.73	-1.84	0.12	9.19	-8.63	5.00	4.66	3.56	13.41	-16.45
MI-CART	9.31	7.99	7.04	-0.59	-18.66	10.35	7.69	7.43	-0.26	-20.55	17.66	14.40	12.82	1.97	-26.22
MI-RF	14.36	10.59	10.62	1.83	-31.53	16.19	10.63	11.24	2.62	-34.60	23.77	17.70	15.52	5.99	-41.33

Biases lower than -10% or larger than 10% are in boldface

for MI-CART. In comparison, the factor variance estimate was sensitive to the degree of nonnormality. With mild nonnormality, most methods were acceptable, except that MI-CART and MI-RF yielded negatively biased factor variance in some conditions. However, for moderate or severe nonnormality, all methods became problematic.

Bias in standard errors

MCAR

Path coefficients With mild and moderate nonnormality, the SE biases of path coefficients were all within the acceptable range ($|\text{bias}| < 10\%$; see Table 3). With severe nonnormality, RFIML and MI-NORM produced some negative SE biases for the path coefficients ($< -10\%$). In contrast, MI-PMM, MI-CART, and MI-RF produced acceptable SEs for path coefficients even with severely nonnormal data.

Loading and factor variance For the loading, all methods were acceptable except that the SEs from MI-NORM were underestimated under moderate nonnormality with $N = 300$ and severe nonnormality with both N s. For factor variance, the SE biases were most prominent under severe nonnormality, where almost all methods, including RFIML, resulted in underestimated standard errors.

MAR-Tail

Path coefficients As shown in Table 4, MI-RF had the worst performance for path coefficients when MAR missingness occurred mainly on the tail, especially when $N = 300$. MI-CART performed better than MI-RF when $MP = 15\%$ or $N = 600$. Other methods were generally better than MI-RF and MI-CART, although they produced a few unacceptable SEs for path coefficients when the sample size was small ($N = 300$), a large proportion of data were missing ($MP = 30\%$), or the data were severely nonnormal.

Table 3 Biases (%) in standard errors under MCAR

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	-4.93	-6.08	-1.37	-4.99	-3.45	-8.37	-3.70	-2.66	-10.27	-8.16	-10.01	-10.21	-8.60	-8.64	-19.49
MI-MVN	-2.59	-0.05	-4.79	-5.66	-5.72	-3.80	0.13	-6.51	-10.97	-10.44	-14.49	-9.41	-5.39	-15.42	-25.05
MI-PMM	-2.58	-3.19	2.08	-2.32	-2.30	-2.86	-0.42	3.47	-5.88	-4.18	-1.39	3.37	5.08	3.30	-8.34
MI-CART	-3.99	-3.26	1.75	-9.35	-6.90	-4.36	-2.16	0.82	-9.38	-7.83	-5.97	-1.78	0.27	-4.37	-28.78
MI-RF	4.36	3.09	7.93	11.25	6.13	1.64	5.15	9.06	6.37	2.60	1.91	4.39	5.69	8.98	-7.35
<i>MP = 30%, N = 600</i>															
RFIML	1.58	-0.90	0.25	-0.10	0.06	-2.55	-1.59	-5.86	-5.27	-1.47	-9.05	-6.01	-9.04	-7.01	-14.44
MI-MVN	-4.45	-3.80	-4.63	-8.55	-7.39	-1.11	-4.32	-4.14	-9.88	-10.59	-15.12	-12.88	-9.79	-16.85	-23.34
MI-PMM	3.12	0.51	0.71	1.33	-0.29	-1.33	-1.47	-5.45	-2.44	-0.80	-5.70	-1.03	-3.73	-1.28	-10.41
MI-CART	0.74	0.05	0.39	-4.71	-4.31	-1.97	-1.13	-6.35	-7.38	-5.22	-6.31	-1.02	-4.35	-6.33	-16.27
MI-RF	6.15	3.52	5.98	13.25	6.26	3.03	4.19	-1.72	7.23	4.79	-1.88	3.79	1.37	6.19	-9.90
<i>MP = 15%, N = 300</i>															
RFIML	-1.99	-5.19	-0.50	-3.75	-1.76	-6.80	-2.27	-1.26	-8.61	-6.28	-10.14	-6.95	-7.46	-7.48	-21.93
MI-MVN	-1.58	-1.40	-1.69	-4.58	-3.75	-5.20	-1.40	-4.43	-10.85	-7.25	-15.73	-9.77	-6.69	-18.69	-29.03
MI-PMM	-1.16	-3.92	1.91	-3.01	-1.57	-4.79	-0.05	0.05	-6.99	-5.53	-5.94	-2.32	-3.18	-3.06	-19.66
MI-CART	-1.91	-4.06	0.75	-6.64	-3.85	-4.93	-1.49	-0.82	-8.59	-7.57	-8.91	-3.30	-3.14	-7.66	-24.08
MI-RF	2.25	0.03	4.25	3.92	2.79	-1.69	2.40	3.03	-0.33	-1.31	-4.53	-0.20	-0.22	-0.82	-21.37
<i>MP = 15%, N = 600</i>															
RFIML	0.58	-1.75	0.42	-2.04	-2.24	-2.97	0.98	-6.03	-5.63	-1.51	-9.62	-6.80	-8.21	-7.37	-16.48
MI-MVN	-1.20	-2.99	-2.17	-5.89	-4.42	0.18	-1.28	-3.44	-7.08	-7.19	-11.87	-8.30	-8.00	-15.07	-19.12
MI-PMM	1.06	-1.51	0.67	-0.96	-2.28	-2.30	2.48	-5.23	-3.56	-0.56	-7.45	-3.92	-5.96	-4.66	-13.53
MI-CART	0.54	-2.48	-0.09	-4.44	-4.24	-2.92	2.98	-7.46	-6.57	-2.55	-7.99	-3.82	-6.28	-8.35	-17.77
MI-RF	3.41	0.93	3.37	5.38	2.16	-0.38	5.28	-3.18	1.40	2.99	-5.66	-1.57	-2.98	-1.73	-13.91

Biases lower than -10% or larger than 10% are in boldface

Loading and factor variance For the loading, MI-RF yielded biased SEs in all conditions; other methods generally performed well with only a few exceptions. For the factor variance, MI-PMM and MI-CART produced large negatively biased SEs in all conditions; RFIML and MI-NORM produced SE biases in all conditions with severe nonnormality, and MI-RF yielded biased SEs under nonnormality only when $MP = 30\%$.

Confidence interval coverage

MCAR

Path coefficients Table 5 shows the CICs from the examined methods under MCAR. The CICs of path coefficients from all methods were within the acceptable range (92.5% to 97.5%) under mild and moderate nonnormality, except that RFIML produced CICs below 92.5% for $\beta_{2,1}$ when the data were moderately nonnormal. When data were severely nonnormal, RFIML and MI-NORM yielded CICs below 92.5% for all three path coefficients in almost all conditions.

MI-PMM, MI-CART, and MI-RF, in comparison, tended to produce adequate CICs for the path coefficients across sample sizes, missing data proportions, and degrees of non-normality, with only a few exceptions.

Loading and factor variance The CICs for the loading showed different patterns. CICs from RFIML and MI-NORM could drop below or close to 92.5% with moderate or severe nonnormality, MI-PMM and MI-CART performed well across all conditions, and the CICs from MI-RF tended to be higher than 97.5% for $MP = 30\%$, especially with the small sample size ($N = 300$). For the factor variance, MI-RF produced CICs below 92.5% in all conditions. The other methods generally performed well under mild nonnormality but yielded unacceptable CICs in most conditions under moderate and severe nonnormality.

MAR-Tail

Path coefficients Under MAR-Tail, CICs for the path coefficients from RFIML were under 92.5% in most conditions

Table 4 Biases (%) in standard errors under MAR-Tail

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	-4.79	-9.12	-4.34	-4.40	-3.18	-9.48	-7.67	-2.40	-8.29	-5.03	-7.92	-12.34	-9.16	-7.76	-16.80
MI-MVN	5.12	4.92	-4.98	1.75	-2.32	3.80	8.68	2.07	2.00	-1.62	-4.76	-0.87	-6.57	-9.88	-5.84
MI-PMM	-0.70	2.93	5.87	4.17	-14.88	-3.05	10.04	11.54	2.99	-14.31	-3.85	9.37	4.27	5.94	-23.48
MI-CART	3.92	19.68	16.54	7.76	-13.89	1.19	19.50	15.78	4.56	-16.18	-4.00	20.92	17.24	-0.31	-31.30
MI-RF	13.83	26.51	26.84	25.29	2.51	16.27	38.42	34.43	29.45	6.94	16.36	44.06	50.21	25.27	-14.59
<i>MP = 30%, N = 600</i>															
RFIML	-2.02	-5.82	-3.59	-3.24	-1.44	-3.29	-1.65	-3.35	-0.69	-4.48	-8.71	-14.09	-7.51	-8.23	-11.21
MI-MVN	-2.37	-0.99	-2.35	-0.74	-4.34	-2.62	-3.92	-4.27	-4.97	-7.90	-8.43	-5.09	-7.01	-13.34	-14.91
MI-PMM	-9.56	-2.82	-2.22	-8.85	-22.98	-13.41	3.45	-4.01	-9.77	-23.00	-13.82	-6.12	1.70	-4.37	-27.65
MI-CART	-5.80	1.88	4.93	1.41	-15.66	-0.18	11.58	2.13	-2.18	-24.28	-7.76	3.32	6.70	2.21	-34.68
MI-RF	2.61	9.04	9.76	11.37	-4.95	5.70	16.76	11.53	13.83	-6.98	3.41	14.59	23.76	18.96	-15.27
<i>MP = 15%, N = 300</i>															
RFIML	-3.69	-5.00	-2.10	-2.91	-4.15	-3.48	-3.33	0.64	-7.08	-2.62	-10.84	-9.15	-3.76	-8.76	-10.61
MI-MVN	0.73	2.63	-3.70	-4.82	-4.74	-1.71	1.46	-2.32	-5.86	-5.13	-4.22	-5.57	-7.56	-13.59	-10.49
MI-PMM	-4.02	1.92	5.66	1.69	-14.98	-4.08	5.60	5.25	-3.35	-15.03	-0.61	6.13	6.94	7.16	-17.64
MI-CART	-4.44	2.61	4.95	2.87	-17.54	-2.60	5.15	11.74	-7.67	-12.03	-4.46	9.87	9.29	2.71	-27.39
MI-RF	7.52	15.39	16.38	18.79	6.90	10.47	19.06	23.03	15.87	6.98	9.10	21.66	26.49	20.06	-4.32
<i>MP = 15%, N = 600</i>															
RFIML	0.13	-0.30	-3.19	-1.95	-0.60	-0.05	0.29	-3.19	3.00	2.06	-5.47	-12.36	-8.11	-6.23	-10.17
MI-MVN	-1.16	-3.06	-2.93	-1.70	-3.64	-2.61	-0.51	-2.01	-2.12	-4.40	-9.18	-6.68	-4.38	-11.46	-12.36
MI-PMM	-7.83	2.80	-3.25	-3.29	-21.40	-3.95	5.33	-3.82	1.58	-16.60	-12.98	-4.52	0.75	0.79	-24.43
MI-CART	-4.62	3.93	3.20	-0.06	-13.56	-2.44	4.62	2.25	2.85	-16.54	-7.85	-1.80	0.63	-1.24	-27.05
MI-RF	4.19	13.63	11.13	12.99	3.09	9.24	17.00	10.51	17.89	7.00	4.77	11.08	13.52	16.31	-9.52

Biases lower than -10% or larger than 10% are in boldface

(see Table 6). The performance of MI-NORM became worse as the degree of nonnormality increased. Specifically, with severely nonnormal data, CICs for the path coefficients from MI-NORM dropped below 92.5% but were slightly better than RFIML. MI-RF led to greater than 97.5% CICs for path coefficients in all conditions. Among all the examined methods, MI-PMM and MI-CART were the best as they produced CICs closer to 95% in most conditions, regardless of sample size, missing data proportion, and nonnormality.

Loading and factor variance For the loading, RFIML and MI-NORM generally yielded acceptable CICs, except when data are severely nonnormal and $N = 600$. MI-PMM produced acceptable CICs, except that the CICs were above 97.5% when the sample size was small ($N = 300$) and the missing data proportion was large ($MP = 30%$). In comparison, MI-RF led to CICs greater than 97.5% in almost all conditions. Among all the methods, MI-CART appeared to produce the best CICs for the loading. The patterns were different regarding CICs for factor variance. The CICs from all donor-based methods tended to be too low ($< 92.5%$) in

all conditions. RFIML and MI-NORM worked well with mildly or moderately nonnormal data. However, they could not produce a sufficient CIC with severely nonnormal data, especially when the missing data proportion was large.

Empirical example

An empirical example is used to illustrate the examined methods. The data used in the example were from the longitudinal Fragile Families and Child Wellbeing Study (FFCWS; Reichman et al., 2001). The data were collected from mothers and fathers shortly after their children’s births between 1998 and 2000 and when the children were 1, 3, 5, 9, and 15 years old. Inspired by Marchand-Reilly and Yaure (2019), we built a structural equation model with three constructs: parents’ relationship at age 5 (η_1), child’s internalizing behaviors at age 5 (η_2), and child’s internalizing behaviors at age 15 (η_3).

In this model, $\beta_{2,1}$, $\beta_{3,1}$ and $\beta_{3,2}$ represent the effects of η_1 on η_2 , η_1 on η_3 , and η_2 on η_3 , respectively. The indicators

Table 5 Confidence interval coverages (CICs) under MCAR

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\varphi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	93.30	93.20	94.10	92.80	92.60	91.10	93.30	94.10	91.90	88.50	88.79	89.49	91.89	93.49	86.89
MI-MVN	93.40	95.23	94.52	92.59	92.99	94.40	94.19	93.67	92.32	90.87	90.34	92.51	93.00	88.77	83.45
MI-PMM	94.12	93.92	95.34	94.93	92.30	92.75	94.79	95.10	94.08	88.56	92.03	93.71	95.06	96.52	85.07
MI-CART	94.08	93.77	95.40	92.65	91.42	91.56	94.85	94.34	93.31	88.16	91.57	92.83	93.89	95.57	84.93
MI-RF	96.76	95.82	96.65	98.33	85.46	96.34	96.86	96.76	97.80	78.35	95.44	95.44	95.54	98.30	75.58
<i>MP = 30%, N = 600</i>															
RFIML	93.60	94.10	95.10	94.00	93.60	92.20	94.10	93.60	93.00	93.10	91.30	91.70	93.00	92.50	89.10
MI-MVN	95.10	94.20	94.30	93.10	92.70	93.80	93.40	93.70	92.20	91.90	89.27	91.17	90.57	89.67	87.16
MI-PMM	94.29	94.39	94.79	95.39	93.99	93.39	94.19	93.99	94.39	92.38	93.37	94.27	93.97	94.87	87.74
MI-CART	94.38	94.48	94.79	93.67	93.05	92.75	93.87	94.08	93.05	91.42	92.54	93.46	94.38	92.94	86.91
MI-RF	96.03	95.61	96.34	97.81	85.27	95.51	95.51	95.09	96.66	83.59	96.03	95.61	96.03	96.87	78.58
<i>MP = 15%, N = 300</i>															
RFIML	94.10	93.70	95.30	93.10	92.90	92.10	94.10	93.70	92.70	88.80	88.90	90.70	91.40	91.60	87.90
MI-MVN	93.30	93.60	95.00	92.90	93.90	93.49	92.69	93.89	92.19	89.99	87.77	90.60	90.50	87.46	85.04
MI-PMM	94.28	93.98	95.69	93.88	92.88	92.59	94.49	94.59	93.89	89.18	90.34	92.56	93.46	94.97	85.71
MI-CART	94.38	93.77	95.81	92.75	92.65	92.75	94.59	93.87	92.85	88.36	89.44	92.21	92.51	93.74	86.05
MI-RF	96.24	95.09	97.07	96.66	90.39	94.15	95.30	95.51	95.51	84.85	92.04	93.40	94.14	96.75	82.62
<i>MP = 15%, N = 600</i>															
RFIML	94.60	94.80	95.80	94.50	93.90	92.10	95.30	93.80	93.70	92.30	90.90	92.20	92.70	92.20	88.20
MI-MVN	94.80	93.80	95.20	92.90	93.60	93.80	94.10	95.50	92.70	92.80	91.30	92.30	91.10	90.00	87.40
MI-PMM	94.59	94.69	95.79	94.69	93.09	92.99	95.79	93.99	94.29	92.28	92.38	93.19	94.19	94.49	88.48
MI-CART	94.18	94.38	95.71	94.48	92.85	92.85	95.61	93.26	92.54	92.03	92.13	92.85	94.48	93.16	88.05
MI-RF	95.61	95.40	96.66	96.34	90.60	94.36	96.13	94.46	95.61	89.45	93.73	94.57	94.67	94.98	84.33

CICs lower than 92.5% are in boldface; CICs greater than 97.5% are in boldface and underscored

were scale scores (averaged across items) of mothers’ ratings on co-parenting quality and relationship quality for η_1 , and anxious/depressed and withdrawn/depressed for both η_2 and η_3 (see Dush et al., 2011; Marchand-Reilly & Yaure, 2019). Table 7 shows the descriptive statistics of the indicators, including minimum, maximum, skewness (ranged from -2.01 to 1.90), and excess kurtosis (ranged from 0.91 to 5.20). The skewness and kurtosis values indicate mild to moderate nonnormality. To demonstrate the impact of the missing data mechanism, we selected a complete subsample ($N = 940$) from the original data and imposed 15% missing data on one indicator of each construct based on the three missing data mechanisms (MACR, MAR-Head, and MAR-Tail) examined in the simulation study.

The parameter and standard error estimates obtained from the five missing data methods are shown in Table 8. The complete data results were also included in the table to evaluate the missing data methods. The parameter and SE estimates from all methods were slightly different than those from the complete data under MCAR and MAR-Head. However, larger differences were observed under MAR-Tail. Notably, the parameter and SE

estimates from MI-RF deviated furthest away from the complete data results under MAR-Tail. The result was consistent with the findings from the simulation study.

Discussion

The current study extended the past research on missing nonnormal data by investigating the performance of several MI methods in comparison with RFIML in recovering common parameters (i.e., structural path coefficients, loadings, and factor variances) in SEM. It considered a broad range of conditions, including various sample sizes, degrees of nonnormality, missing data proportions, and missing data generating mechanisms. It evaluated the methods using bias in point and SE estimates, as well as CIC. The results suggest that the design factors had a differential influence on the estimates of the different types of parameters. In general, the factor variance estimates were found to be more sensitive to the degree of non-normality than the path coefficient and loading estimates.

Table 6 Confidence interval coverages (CICs) under MAR-Tail

Method	Mild non-normality					Moderate non-normality					Severe non-normality				
	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$	$\beta_{2,1}$	$\beta_{3,1}$	$\beta_{3,2}$	λ_3	$\Phi_{1,1}$
<i>MP = 30%, N = 300</i>															
RFIML	88.40	90.90	90.50	96.20	92.70	84.75	89.37	90.77	95.39	93.88	84.13	84.02	86.35	92.92	90.29
MI-MVN	94.25	93.70	92.88	97.40	94.93	91.51	93.50	94.30	96.55	94.69	89.16	91.58	91.44	92.44	90.44
MI-PMM	93.50	93.75	95.34	98.28	77.57	91.08	95.18	96.51	98.07	75.30	91.89	93.56	95.75	98.20	62.29
MI-CART	95.85	97.11	97.11	96.23	70.10	94.76	96.71	98.29	96.83	64.56	94.97	96.49	97.08	95.67	52.63
MI-RF	98.71	98.61	99.36	99.36	39.34	97.72	99.35	99.24	99.24	38.33	99.11	98.67	99.44	99.11	29.67
<i>MP = 30%, N = 600</i>															
RFIML	89.60	91.40	90.80	94.90	94.00	84.30	87.30	89.00	93.70	94.70	83.87	86.27	87.58	88.88	91.28
MI-MVN	91.35	93.76	92.76	95.88	93.86	89.68	89.88	91.38	92.79	93.29	86.99	89.86	89.45	86.99	90.16
MI-PMM	89.34	92.25	91.55	95.27	78.57	87.81	93.05	92.75	95.77	73.72	93.23	91.71	93.02	95.35	58.95
MI-CART	95.77	95.15	95.05	94.12	72.78	95.15	95.36	95.77	94.63	64.81	95.37	93.82	96.19	94.64	52.63
MI-RF	97.07	98.12	97.91	99.06	28.32	96.66	98.12	98.64	98.96	26.75	97.81	98.12	98.64	98.64	21.11
<i>MP = 15%, N = 300</i>															
RFIML	90.70	93.90	92.60	95.50	93.90	88.49	89.99	90.99	94.29	95.50	83.75	85.46	88.06	92.88	93.48
MI-MVN	93.05	94.06	92.95	95.87	94.06	89.10	92.73	93.54	94.35	95.26	87.32	91.09	89.10	92.87	92.98
MI-PMM	91.41	94.64	95.25	97.07	83.72	92.00	95.45	94.43	97.06	79.66	92.59	93.74	95.09	97.08	72.34
MI-CART	94.23	96.80	95.26	93.71	72.68	94.94	96.90	96.28	94.11	67.05	95.50	95.40	96.03	93.72	59.21
MI-RF	98.21	98.42	98.32	97.79	58.61	97.38	98.53	98.85	97.28	53.19	98.31	98.00	98.63	97.47	43.31
<i>MP = 15%, N = 600</i>															
RFIML	91.30	94.70	91.70	95.30	95.30	87.60	90.10	90.80	94.40	96.30	84.80	85.80	88.50	92.10	95.20
MI-MVN	92.40	93.30	93.40	94.20	93.30	89.50	91.40	91.70	93.50	94.20	86.49	88.49	88.79	90.29	94.99
MI-PMM	91.68	94.69	92.79	95.59	83.87	90.28	93.69	92.59	96.69	79.96	91.87	92.77	94.38	95.08	69.88
MI-CART	94.48	96.63	95.71	93.56	71.09	95.60	95.81	95.19	93.66	67.69	94.27	94.78	95.19	93.86	55.78
MI-RF	96.45	97.49	97.91	97.39	45.87	96.97	98.12	97.91	98.22	41.59	96.76	97.81	97.28	98.54	33.75

CICs lower than 92.5% are in boldface; CICs greater than 97.5% are in boldface and underscored

Table 7 Empirical example: Descriptive statistics

Construct	Indicator	Min	Max	Skewness	Excess kurtosis	Imposed missingness
Parents' relationship at age 5	Co-parenting quality	2.17	4.00	-2.01	5.20	Yes
	Relationship quality	1.22	4.11	-0.93	0.91	No
Child's internalizing behaviors at age 5	Anxious/depressed	1.00	2.56	1.47	2.88	Yes
	Withdrawn/depressed	1.00	2.14	1.24	1.66	No
Child's internalizing behaviors at age 15	Anxious/depressed	1.00	3.00	1.90	4.40	Yes
	Withdrawn/depressed	1.00	3.00	1.78	2.93	No

This study also revealed similarities and discrepancies of the methods under the various conditions examined in the study and provided valuable insights about their empirical performance. Regarding the two parametric methods (RFIML and MI-NORM), we found that RFIML generally performed well under MCAR or MAR, except when MAR occurred mainly on the heavy tail of the data distribution (i.e., MAR-Tail), and the proportion of missing data was

large. These findings are consistent with previous research (e.g., Enders, 2001b; Savalei & Falk, 2014). The overall performance of MI-NORM was similar to that of RFIML, particularly under mild and moderate nonnormality. Under MAR-Tail, MI-NORM could even outperform RFIML by producing slightly better CICs for path coefficients.

Regarding the semi-parametric and nonparametric methods, we expected MI-PMM, MI-CART, and MI-RF to

Table 8 Empirical example: Comparison of missing data methods

	$\beta_{2,1}$		$\beta_{3,1}$		$\beta_{3,2}$	
	Est.	SE	Est.	SE	Est.	SE
Complete data						
MLR	-0.169	0.040	-0.277	0.066	0.362	0.091
MCAR						
RFIML	-0.178	0.043	-0.254	0.069	0.364	0.103
MI-NORM	-0.175	0.042	-0.249	0.067	0.377	0.103
MI-PMM	-0.177	0.043	-0.250	0.070	0.374	0.107
MI-CART	-0.170	0.045	-0.250	0.069	0.359	0.100
MI-RF	-0.183	0.051	-0.260	0.079	0.364	0.110
MAR-Head						
RFIML	-0.170	0.041	-0.292	0.071	0.359	0.098
MI-NORM	-0.164	0.041	-0.298	0.069	0.360	0.099
MI-PMM	-0.160	0.041	-0.292	0.069	0.371	0.095
MI-CART	-0.164	0.043	-0.282	0.072	0.373	0.094
MI-RF	-0.159	0.047	-0.278	0.080	0.387	0.108
MAR-Tail						
RFIML	-0.193	0.049	-0.297	0.086	0.372	0.100
MI-NORM	-0.202	0.050	-0.276	0.081	0.381	0.109
MI-PMM	-0.199	0.061	-0.293	0.091	0.351	0.111
MI-CART	-0.227	0.068	-0.230	0.084	0.411	0.119
MI-RF	-0.342	0.135	-0.424	0.236	0.533	0.289

outperform MI-NORM in dealing with missing nonnormal data because they rely less on the normality assumption. However, the results did not fully support the expectation. When missing data were MCAR or MAR-Head, MI-PMM was comparable to or, in some conditions, even better than MI-NORM or RFIML. For example, it produced better SEs and CICs under moderate and severe nonnormality. However, it could result in more biased point and SE estimates than MI-NORM or RFIML under MAR-Tail, especially when the sample size was small and the missing data proportion was large. MI-CART performed similarly to MI-PMM, except it was more sensitive to small sample sizes under MAR-Tail. The overall performance of MI-RF was poor across conditions, especially under MAR-Tail. As mentioned above, studies found that MI-CART and MI-RF were able to deal with interactions or nonlinearities adequately (Doove et al., 2014; Shah et al., 2014). These studies, however, were not conducted in the SEM context and did not involve nonnormality in data. Our findings shed light on the performance of these methods in a broader range of situations. The combination of these results implies that the nonparametric methods may have difficulty handling certain types of nonnormality, for example, nonnormality that is not due to nonlinear relationships among observed variables.

Based on these findings, we offer the following recommendations to substantive researchers. Note that these

recommendations are only limited to the conditions examined in the study. RFIML generally performed well in dealing with missing data and nonnormality, except that it yielded lower CICs (< 90%) under MCAR and MAR-Head with moderate and severe nonnormality, or under MAR-Tail regardless of the degree of nonnormality. MI-NORM was in general comparable to RFIML; thus, it could serve as an alternative to RFIML if MI is to be adopted. Although MI-PMM showed some advantages over RFIML and MI-NORM under moderate and severely nonnormality when missing data were MCAR or MAR-Head, it generally had problems estimating parameters when the missing data mechanism was MAR-Tail. Thus, we only recommend it when the missing data mechanism is not MAR-Tail. MI-CART was comparable to MI-PMM under MCAR or MAR-Head, but it could yield more severely biased point and standard error estimates than MI-PMM did under MAR-Tail. Since MI-CART is not better than MI-PMM, and MI-RF was worse than the other methods in many conditions, we do not recommend either approach. It is important to note that when the nonnormality is severe, all these methods could fail, especially with small sample size and a large proportion of missing data.

Given that the performance of the examined methods is highly contingent on the missing data mechanism, it would be helpful to explore the missing data mechanism of the data at hand. Researchers could utilize available prior

knowledge on the distribution of the target variable to determine whether the missing data likely occurred on the tail or head of the distribution. This could shed some light on the possible degree and direction of potential bias. In addition, it appears that a larger sample size would mitigate the influence of missing data and nonnormality. Thus, if possible, researchers should plan a relatively large sample size for SEM, accounting for potential missingness and nonnormality in the data.

Limitations

Several limitations of the study are worth mentioning. First, we only examined a three-factor SEM model where missing data were imposed on two indicators of each factor. To explore whether the examined methods perform differently when applied to other kinds of SEM models, we provided an additional empirical example in the Appendix with a Multiple Indicator Multiple Cause (MIMIC) model. Similar to the empirical example described above, MI-CART and MI-RF yielded the most different results compared with those from the complete data. This is also consistent with the findings from the simulation. More models could be investigated in the future.

Second, we only considered univariate unconditional nonnormality, and the degree of nonnormality was set to be constant across all variables. In practice, nonnormality could occur after conditioning on other variables (i.e., conditional nonnormality). The degree of nonnormality could vary across variables, or nonnormality may come from categorical—such as ordinal—indicators. We refer readers to Jia and Wu (2019) for methods to deal with missing ordinal data. More conditions could be examined in future research.

Third, the MAR on each indicator was only determined by another indicator of the same factor. In reality, the MAR mechanism could be much more complex. Future research may examine other MAR data that are determined by a combination of variables.

Fourth, as donor-based methods, the capability of MI-PMM, MI-CART, and MI-RF is dependent on the availability of suitable donors in the sample. Thus, they require a larger sample size than the other examined MI methods, especially when a larger proportion of data are missing. For example, Lee and Carlin (2017) used $N = 1000$ for estimating marginal means; Doove et al. (2014) also generated 1000 observations for multiple regression and logistic regression models. To preliminarily explore the sample size issue, we selected the worst scenario in the simulation ($MP = 30\%$, severe nonnormality, and MAR-Tail) and used it to examine whether the MI-RF performance would change if we increased the sample size to 1000 or 10,000. We found that the accuracy of parameter estimates improved as the sample size increased. Nevertheless, even with $N = 10,000$, large bias for path coefficients (13–20%) and latent factor variance

(41–45%) were still observed. We did not consider sample sizes larger than 10,000 due to the limit of time and computational power. It would be interesting to conduct a full-scale simulation study to thoroughly examine the effect of sample size on the donor-based MI methods.

Fifth, the performance of the recursive partitioning methods, MI-CART and MI-RF, could be affected by the settings of hyperparameters such as the number of trees (for RF) and the size of leaves. In this study, we used the hyperparameter values found reliable in past research (e.g., Doove et al., 2014; Shah et al., 2014). Per the suggestion of a reviewer, we conducted a small ad-hoc simulation to explore the impact of hyperparameters on the performance of MI-RF. We did not find noticeably different results in the conditions we chose: different numbers of trees (10, 50, and 100 trees), and different leaf sizes (5, 10, and 20 donors in each leaf). Strategies for optimizing hyperparameters for RF (e.g., random search and sequential model-based optimization) can be found in the machine learning literature (Probst et al., 2019). However, these strategies have not been fully implemented in behavioral studies, especially when combined with missing data imputation. Further research is warranted.

Finally, we did not include model fit evaluation in the current study. The robust ML method will produce a corrected (correct for nonnormality) chi-square test statistic for model fit. To our knowledge, there is not yet a clear solution for pooling corrected chi-square test statistics across imputations (Enders & Mansolf, 2018). Future work is needed to identify/develop an appropriate pooling method, based on which the performance of MI methods for missing nonnormal data in model fit evaluation could be examined.

Appendix

This is an additional empirical example to demonstrate the differences among the examined missing nonnormal data methods. Inspired by Fan et al. (2010), we used the data from Educational Longitudinal Study of 2002 (National Center for Education Statistics, 2002) to examine the covariance between two latent constructs (students' motivation and parent-school communication concerning poor performance), and the effect of two observed variables (socioeconomic status [SES] and gender) on them. In this Multiple Indicator Multiple Cause (MIMIC) model, students' motivation was measured by three composite scores: Math self-efficacy, English self-efficacy, and general effort and persistence. Parent-school communication concerning poor performance had two indicators: frequencies of school contacted parent about poor performance, and frequencies of parent contacted school about poor performance.

We chose a complete subsample ($N = 1287$) from the original data and imposed 15% missing data on all the three

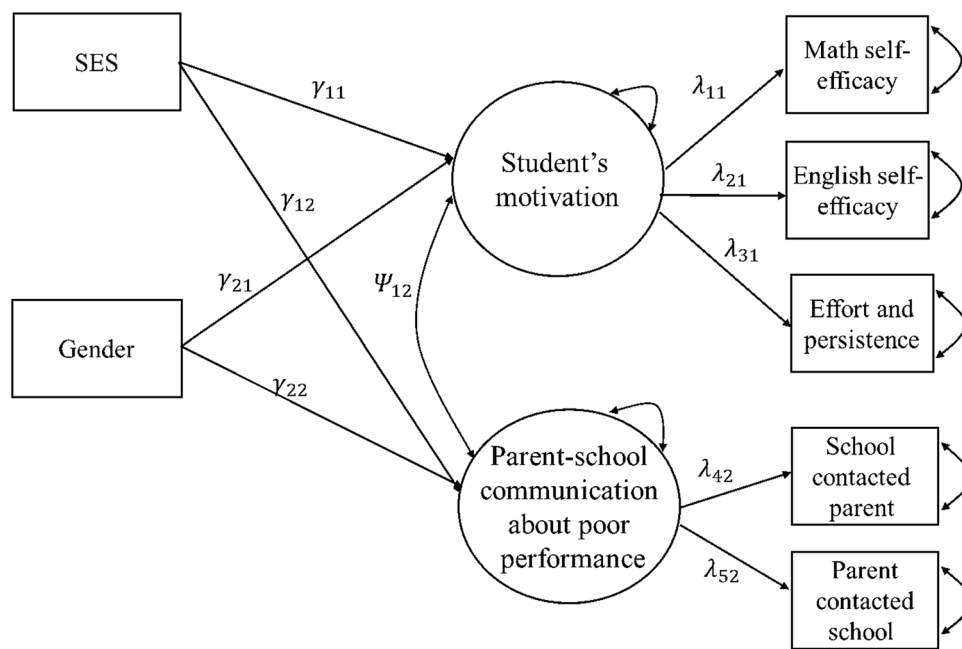


Fig. A1. MIMIC Model

Table A1 Point and standard errors estimates of selected parameters

	γ_{11}	γ_{21}	Ψ_{12}
Complete			
MLR	0.088 (0.024) *	0.028 (0.024)	-0.024 (0.006) *
MCAR			
RFIML	0.082 (0.026) *	0.037 (0.026)	-0.023 (0.007) *
MI-NORM	0.083 (0.026) *	0.036 (0.026)	-0.023 (0.007) *
MI-PMM	0.087 (0.026) *	0.036 (0.026)	-0.023 (0.007) *
MI-CART	0.086 (0.026) *	0.033 (0.026)	-0.023 (0.007) *
MI-RF	0.083 (0.026) *	0.037 (0.025)	-0.022 (0.007) *
MAR-Head			
RFIML	0.078 (0.033) *	0.030 (0.026)	-0.022 (0.007) *
MI-NORM	0.073 (0.031) *	0.028 (0.026)	-0.021 (0.006) *
MI-PMM	0.087 (0.031) *	0.028 (0.026)	-0.021 (0.007) *
MI-CART	0.007 (0.027)	0.011 (0.026)	-0.020 (0.007) *
MI-RF	0.055 (0.029)	0.025 (0.027)	-0.023 (0.007) *
MAR-Tail			
RFIML	0.092 (0.030) *	0.035 (0.027)	-0.028 (0.007) *
MI-NORM	0.093 (0.032) *	0.035 (0.027)	-0.029 (0.007) *
MI-PMM	0.074 (0.031) *	0.040 (0.027)	-0.028 (0.007) *
MI-CART	-0.016 (0.027)	0.079 (0.026)	-0.027 (0.007) *
MI-RF	0.059 (0.029) *	0.046 (0.026)	-0.028 (0.007) *

Standard errors are in parentheses. The * symbol indicates significance at $\alpha = 0.05$

indicators of student’s motivation, based on the three missing data mechanisms: MACR, MAR-Head, and MAR-Tail.

In both MAR conditions, we used SES to determine the probabilities of missingness on those indicators.

The parameter and standard error estimates obtained from the five missing data methods in comparison with complete data results are shown in Table A1. Under MCAR, all missing data methods yielded comparable results with that of the complete data, while under MAR-Head and MAR-Tail, largest differences were found to be associated with the effect of SES on student’s motivation (γ_{11}). Specifically, for the point estimate, MI-PMM performed the best under MAR-Head, while underestimated γ_{11} under MAR-Tail. RFIML and MI-NORM yielded smaller γ_{11} under MAR-Head and overestimated γ_{11} under MAR-Tail. The estimates obtained from MI-CART and MI-RF in both MAR conditions were drastically smaller than the complete data results. All methods yielded inflated standard errors of γ_{11} to a certain degree in both MAR conditions.

References

Allison, P. D. (2000). *Missing data*. Sage.
 Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40–64. <https://doi.org/10.1111/j.1751-5823.2010.00103.x>
 Asparouhov, T., & Muthén, B. (2010). Multiple imputation with Mplus. *Technical Report*. Retrieved September, 18, 2021, from: <https://www.statmodel.com>
 Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. Taylor & Francis.
- 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. <https://doi.org/10.1111/j.2044-8317.1984.tb00789.x>
- Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, 44(2), 347–357. <https://doi.org/10.1111/j.2044-8317.1991.tb00966.x>
- 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. <https://doi.org/10.1037/1082-989X.6.4.330>
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16–29. <https://doi.org/10.1037/1082-989X.1.1.16>
- Demirtas, H. (2009). Multiple imputation under the generalized lambda distribution. *Journal of Biopharmaceutical Statistics*, 19(1), 77–89. <https://doi.org/10.1080/10543400802527882>
- Demirtas, H., & Hedeker, D. (2008). Imputing continuous data under some non-Gaussian distributions. *Statistica Neerlandica*, 62(2), 193–205. <https://doi.org/10.1111/j.1467-9574.2007.00377.x>
- Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1), 69–84. <https://doi.org/10.1080/10629360600903866>
- Di Zio, M., & Guarnera, U. (2009). Semiparametric predictive mean matching. *ASta Advances in Statistical Analysis*, 93(2), 175–186. <https://doi.org/10.1007/s10182-008-0081-2>
- Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92–104. <https://doi.org/10.1016/j.csda.2013.10.025>
- Dush, C. M. K., Kotila, L. E., & Schoppe-Sullivan, S. J. (2011). Predictors of supportive coparenting after relationship dissolution among at-risk parents. *Journal of Family Psychology*, 25(3), 356. <https://doi.org/10.1037/a0023652>
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26.
- Enders, C. K. (2001a). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1), 128–141. https://doi.org/10.1207/S15328007SEM0801_7
- Enders, C. K. (2001b). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4), 352–370. <https://doi.org/10.1037/1082-989X.6.4.352>
- Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5
- Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, 23(1), 76–93. <https://doi.org/10.1037/met0000102>
- Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement*, 58(5), 701–735. <https://doi.org/10.1177/0013164498058005001>
- Fan, W., & Williams, C. M. (2010). The effects of parental involvement on students' academic self-efficacy, engagement and intrinsic motivation. *Educational Psychology*, 30(1), 53–74.
- Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2), 87–107. <https://doi.org/10.1080/10705519709540063>
- Fleishman, A. I. (1978). A method for simulating nonnormal distributions. *Psychometrika*, 43(4), 521–532. <https://doi.org/10.1007/BF02293811>
- Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, 47(1), 1–25. <https://doi.org/10.1080/00273171.2012.640589>
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>
- Hayes, T., & McArdle, J. J. (2017). Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. *Computational Statistics & Data Analysis*, 115, 35–52. <https://doi.org/10.1016/j.csda.2017.05.006>
- He, Y., & Raghunathan, T. E. (2009). On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics: Simulation and Computation*, 38(4), 856–883. <https://doi.org/10.1080/03610910802677191>
- Heitjan, D. F., & Little, R. J. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society C*, 40(1), 13–29. <https://doi.org/10.2307/2347902>
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1–47. <https://doi.org/10.18637/jss.v045.i07>
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling An overview and a meta-analysis. *Sociological Methods & Research*, 26(3), 329–367. <https://doi.org/10.1177/0049124198026003003>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
- Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior Research Methods*, 51(5), 2337–2355.
- Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3), 9.
- Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, 42(4), 371–404. <https://doi.org/10.3102/1076998616687084>
- Koller-Meinfelder, F. (2010). Analysis of incomplete survey data—multiple imputation via bayesian bootstrap predictive mean matching. PhD thesis, Otto-Friedrich-University, Bamberg. Retrieved November 5, 2019, from: <https://www.fis.uni-bamberg.de/handle/uniba/213>
- Lai, K. (2018). Estimating standardized SEM parameters given non-normal data and incorrect model: Methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 600–620.
- Lee, K. J., & Carlin, J. B. (2017). Multiple imputation in the presence of nonnormal data. *Statistics in Medicine*, 36(4), 606–617. <https://doi.org/10.1002/sim.7173>

- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18–22.
- Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296. <https://doi.org/10.1080/07350015.1988.10509663>
- Little, T., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the Items versus Parcels Controversy Needn't Be One. *Psychological Methods*, 18(3), 285–300. <https://doi.org/10.1037/a0033266>
- Marchand-Reilly, J. F., & Yaure, R. G. (2019). The Role of Parents' Relationship Quality in Children's Behavior Problems. *Journal of Child and Family Studies*. <https://doi.org/10.1007/s10826-019-01436-2>
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530. <https://doi.org/10.1093/biomet/57.3.519>
- Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, 42(4), 432–466. <https://doi.org/10.3102/1076998617690869>
- Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1), 75. <https://doi.org/10.1186/1471-2288-14-75>
- Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3), 431–462. <https://doi.org/10.1007/BF02294365>
- National Center for Education Statistics. (2002). Education longitudinal study of 2002 (ELS:2002). U.S. Department of Education. [Data file]. Retrieved March 23, 2022, from https://nces.ed.gov/surveys/els2002/avail_data.asp
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7(4), 557–595.
- Palomo, J., Dunson, D. B., & Bollen, K. (2011). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models*. Elsevier. <https://doi.org/10.1016/B978-044452044-9/50011-2>
- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/widm.1301>
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation Statistical Computing. Retrieved August 19, 2019, from <http://www.R-project.org/>
- Reichman, N., Teitler, J., Garfinkel, I., & McLanahan, S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, 23(4–5), 303–326. [https://doi.org/10.1016/S0190-7409\(01\)00141-4](https://doi.org/10.1016/S0190-7409(01)00141-4)
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons.
- 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. V. Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.
- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 477–497.
- Savalei, V., & Falk, C. F. (2014). Robust Two-Stage Approach Outperforms Robust Full Information Maximum Likelihood With Incomplete Nonnormal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 280–302. <https://doi.org/10.1080/10705511.2014.882692>
- Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from FIML. *Structural Equation Modeling*, 19, 477–494. <https://doi.org/10.1080/10705511.2012.687669>
- Savalei, V., & Rhemtulla, M. (2017). Normal Theory Two-Stage ML Estimator When Data Are Missing at the Item Level. *Journal of Educational and Behavioral Statistics*, 42(4), 405–431. <https://doi.org/10.3102/1076998617694880>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.
- Schafer, J. L. (2010). *Analysis of incomplete multivariate data*. CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), 425–446. [https://doi.org/10.1016/0167-9473\(95\)00057-7](https://doi.org/10.1016/0167-9473(95)00057-7)
- 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(6), 764–774. <https://doi.org/10.1093/aje/kwt312>
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, 48(3), 465–471. <https://doi.org/10.1007/BF02293687>
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. <https://doi.org/10.1080/10629360600810434>
- von Hippel, P. T. (2005). TEACHER'S CORNER: How Many Imputations Are Needed? A Comment on Hershberger and Fisher (2003). *Structural Equation Modeling*, 12(2), 334–335. https://doi.org/10.1207/s15328007sem1202_8
- von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1), 105–138.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399. <https://doi.org/10.1002/sim.4067>
- Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30(1), 165–200. <https://doi.org/10.1111/0081-1750.00078>
- Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, 59(2), 397–417. <https://doi.org/10.1348/000711005X85896>
- Yuan, K. H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions.

Sociological Methods & Research, 41(4), 598–629. <https://doi.org/10.1177/0049124112460373>

Zopluoglu, C. (2013). Generating multivariate nonnormal variables [Computer program]. Retrieved October 21, 2014, from <http://sites.education.miami.edu/zopluoglu/software-programs>

Open practices statement This paper is based on a simulation study. No data is available, and no experiment was preregistered.

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.