Behavior Research Methods (2024) 56:1229–1243
https://doi.org/10.3758/s13428-023-02088-3

# Pooling test statistics across multiply imputed datasets for nonnormal items

Fan Jia[1]

Accepted: 9 February 2023 / Published online: 27 March 2023

## Abstract

In structural equation modeling, when multiple imputation is used for handling missing data, model fit evaluation involves pooling likelihood-ratio test statistics across imputations. Under the normality assumption, the two most popular pooling approaches were proposed by Li et al. (*Statistica Sinica, 1*(1), 65–92, 1991) and Meng and Rubin (*Biometrika, 79*(1), 103–111, 1992). When the assumption of normality is violated, it is not clear how well these pooling approaches work with the test statistics generated from various robust estimators and multiple imputation methods. Jorgensen and colleagues (2021) implemented these pooling approaches in their R package semTools; however, no systematical evaluation has been conducted. In this simulation study, we examine the performance of these approaches in working with different imputation methods and robust estimators under nonnormality. We found that the naïve pooling approach based on Meng and Rubin (*Biometrika, 79*(1), 103–111, 1992; $D_{3SN}$) worked the best when combining with the normal-theory-based imputation and either MLM or MLMV estimator.

**Keywords** Structural equation modeling · Robust estimator · Nonnormality · Missing data · Multiple imputation · Pooling · Test statistic

Model fit evaluation is a critical aspect in structural equation modeling (SEM). In recent decades, model fit evaluation with difficult data, such as nonnormal data or incomplete data, have received extensive attentions (e.g., Satorra & Bentler, 1994; Yuan & Hayashi, 2006). When data are skewed or kurtotic, the normality assumption of the popular estimators (e.g., maximum likelihood, a.k.a., ML) in SEM is violated, resulting in biased standard errors and inappropriate likelihood ratio test statistics and therefore distorted model fit indices that are based on the test statistic (e.g., Browne, 1984; Curran et al., 1996; Olsson et al., 2000). A well-known solution is to use one of the robust ML-based estimators to mitigate the impact of nonnormality (e.g., MLM, MLMV, MLR; Satorra & Bentler, 1994; Yuan & Bentler, 2000; Asparouhov & Muthén, 2005; Asparouhov & Muthén, 2010; Maydeu-Olivares, 2017). These robust estimators use a sandwich-like covariance matrix of parameters to obtain corrected standard errors; they also produce adjusted test statistics for more accurate statistical inferences.

Most of these robust estimators were developed based on complete data. The presence of missing data adds another layer of difficulty. For missing nonnormal data, only MLR can be directly applied. Studies show that MLR (a.k.a, robust full information maximum likelihood, robust FIML, or RFIML) can generally yield unbiased point and standard error estimates with certain types of missing nonnormal data (e.g., Enders, 2001). However, it has been found that MLR could generate inflated type I errors in many conditions (e.g., Liu & Sriutaisuk, 2020; Savalei & Bentler, 2009). Missing nonnormal data can also be handled using multiple imputation (MI). A typical MI takes three steps. In the first step, it generates multiply imputed data sets, in which missing values are filled in using a certain imputation model (i.e., imputation step). Second, it runs the analysis model with each imputed data set (i.e., analysis phase). In the end, the results from all imputed data sets are pooled to produce the final results (i.e., pooling step). One advantage of MI is that it fills in missing values in multiply imputed data so that the complete-data-based estimation methods can be applied. This allows the use of other robust estimators, such as MLMV,

✉ Fan Jia
fjia3@ucmerced.edu

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

which was found to be superior to MLR with complete data (Maydeu-Olivares, 2017). However, the challenge of using MI with robust estimators is that even though the point and standard errors estimates can be easily pooled after imputation (Rubin, 1987), the pooling approaches for the chi-square test statistics remain less clear in the literature.

Two pooling approaches, from Li et al. (1991) and Meng and Rubin (1992) for the chi-square test statistics, have received some attention in recent years. These two methods are referred to as $D_2$ and $D_3$, respectively, in the missing data literature (e.g., Enders, 2010; Schafer, 1997). Enders and Mansolf (2018) examined the performance of $D_3$ in the missing normal data context. They found that the values and type I error rates of chi-square test statistics from $D_3$ were comparable to those from FIML with certain types of missing data, but $D_3$ tended to have lower power. Liu and Sriutaisuk (2020) focused on the performance of $D_2$ for ordinal data with least-squares estimators. They revealed that $D_2$ could work appropriately for ordinal variables if the analysis model also contained other completely observed variables, or variables with little missingness that correlated with the incomplete ordinal variables. Jorgensen et al. (2021) implemented these pooling approaches and their variations in the R package semTools[1]. However, none of these approaches has been well evaluated for missing nonnormal data with robust ML-based estimators.

This article aims to fill the gap in the literature and provide guidance to substantive researchers in the case when MI is used for missing data handling and when adjustment for nonnormality is necessary. We organize the rest of the article in the following manner. We first introduce the background of the study, including three adjusted test statistics from different robust ML-based estimators with complete nonnormal data; two imputation strategies for missing nonnormal data; and four pooling approaches and their implementations with adjusted test statistics. Next, we describe the design and results of a simulation study that compared 24 combinations of robust estimators, imputation methods and pooling approaches (3×2×4), under a variation of conditions. We conclude the article with a discussion of the results and guidelines for substantive researchers.

## Adjusted test statistics

### MLM and MLMV

The likelihood ratio test statistic of ML follows a chi-square distribution under the assumption of normality. Let $\hat{\theta}$ and $\tilde{\beta}$ denote the parameter estimates of the structure model and saturated model, respectively, then the likelihood ratio test statistic is defined as

$$T = -2\left[l(\hat{\theta}) - l\left(\tilde{\upbeta}\right)\right] \tag{1}$$

where $l(\hat{\theta})$ is the maximized log-likelihood under the structured model, and $l\left(\tilde{\beta}\right)$ is the corresponding log-likelihood of the saturated model (Savalei & Bentler, 2005). When the normality assumption is violated, the chi-square distribution can no longer be used as the reference for the likelihood ratio test statistic. Satorra and Bentler (1994) proposed a robust estimator that produces a mean-adjusted test statistic for nonnormal data. We refer it to as MLM to keep consistence with the term used in popular SEM software programs, such as Mplus (Muthén & Muthén, 1998–2017) and lavaan (Rosseel, 2012), even though the default settings of MLM in these programs may differ (Savalei & Rosseel, 2021). In lavaan, the MLM adjusted chi-square test statistic is given by

$$T_{\mathrm{MLM}} = \frac{d}{tr\left(\hat{U}_E\tilde{\Gamma}\right)}T \tag{2}$$

where $d$ is the degree of freedom, $\hat{U}_E = \hat{W} - \hat{W}\hat{\Delta}\left(\hat{\Delta}'\hat{W}\hat{\Delta}\right)^{-1}\hat{\Delta}'\hat{W}$, in which $\hat{W}$ is the estimate of the complete data expected information matrix for the structured model (Mplus uses saturated model here) and $\hat{\Delta} = \frac{\partial\sigma(\theta)}{\partial\theta'}\Big|_{\hat{\theta}}$. $\tilde{\Gamma}$ is the asymptotic covariance matrix under the asymptotic distribution free (ADF) assumption (Browne, 1984). With this adjustment, the mean of $T_{MLM}$ is equal to the mean of a $\chi^2_d$. The distribution of $T_{MLM}$ is not exactly $\chi^2_d$, but it works well approximately (Savalei, 2014).

Asparouhov and Muthén (2010) described a variation of $T_{MLM}$, which is known as the mean-and-variance adjusted test statistic or scaled-and-shifted test statistic (denoted $T_{MLMV}$), and can be obtained from popular software packages, such as Mplus and lavaan, with the MLMV estimator. $T_{MLMV}$ adjusts $T$ with a scale parameter (a) and a shift parameter (b).

$$T_{MLMV} = aT + b \tag{3}$$

where $a = \sqrt{\frac{d}{tr\left(\left(\hat{U}_E\tilde{\Gamma}\right)^2\right)}}$, and $b = d - \sqrt{\frac{d\left(tr\left(\hat{U}_E\tilde{\Gamma}\right)\right)^2}{tr\left(\left(\hat{U}_E\tilde{\Gamma}\right)^2\right)}}$.

MLM and MLMV only deal with complete nonnormal data. MI creates the possibility of MLM and MLMV being implemented and evaluated in the context of missing nonmoral data.

[1] These functionalities will be moved to the new package *lavaan.mi* written by the same authors at the time of publication or soon after.

### MLR

Another popular robust estimator in SEM is termed MLR (a.k.a. robust full information maximum likelihood, robust FIML, or RFIML in missing data literature). MLR can be seen as an extension of FIML. It is widely available in SEM software packages (e.g., Mplus and lavaan) and can be directly applied to data with or without missingness. The MLR test statistic is a mean-adjusted test statistic and can take a variety of forms (Savalei & Rosseel, 2021). One most used form (default in Mplus and lavaan) is based on Asparouhov and Muthén (2005). Let's use $\hat{\beta} = \beta\left(\hat{\theta}\right)$ to denote the vector including all non-redundant elements in the covariance matrix and the means under the structure model, and $\beta$ represents the vector of those elements under the saturated model. Then, $\hat{A}_\beta$ and $A_\beta$ denote the negative "second derivative of the log-likelihood" ($A_\beta$, see Eq. 4) evaluated at $\hat{\beta}$ and $\tilde{\beta}$, respectively.

$$A_\beta = -\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 1_i(\beta)}{\partial\beta\partial\beta'} \tag{4}$$

We also use $\hat{B}_\beta$ and $\tilde{B}_\beta$ to denote the outer product of the "first derivative of log-likelihood" with itself ($B_\beta$, see Eq. 5) evaluated at $\hat{B}$ and $\tilde{B}$, respectively.

$$B_\beta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\beta)}{\partial\beta}\frac{\partial l_i(\beta)}{\partial\beta'} \tag{5}$$

Then the MLR test statistic is written as

$$T_{MLR} = \frac{d}{tr\left(\tilde{B}_\beta\tilde{A}_\beta^{-1}\right) - tr\left(\hat{B}_\beta\hat{A}_\beta^{-1}\right)}T \tag{6}$$

where $d$ is the degrees of freedom. This form is asymptotically equivalent to the $T_2^*$ statistic in Yuan and Bentler (2000). Recent research examined the performance of $T_{MLR}$ with complete nonnormal data (Maydeu-Olivares, 2017) and missing nonnormal data (Liu & Sriutaisuk, 2020). They both found that $T_{MLR}$ tended to produce inflated type I error rates in various conditions.

## Multiple imputation strategies for missing data

### Missing data mechanisms

Data can be missing through different processes. Rubin (1976) summarized these processes into three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR and MAR are also known as ignorable mechanisms, under which missing data on a variable is unrelated to any variables in the data (completely at random) or only related to the observed variables (at random). Under ignorable missing data mechanisms, the missing information can be well recovered using modern missing data techniques, such as FIML and MI. MNAR, on the other hand, is the non-ignorable mechanism, for missing data on a variable is determined by the unobserved values of the missing data themselves; and to recover missing information, one needs to explicitly model the missing data generation process. We assume ignorable missingness in this study.

### Multiple imputation based on normality (MI-NORM)

Let's focus on MI. For nonnormal continuous data, one imputation strategy is to ignore nonnormality and use regular MI methods as if the data are normal. We refer to these types of methods as normal-theory-based MI or MI-NORM. MI-NORM can be implemented using either of two algorithms: joint modeling (JM; Schafer, 1997) and expectation-maximization with bootstrapping (EMB; Honaker et al., 2011). The two algorithms are theoretically equivalent as they both impute missing values on multiple variables simultaneously based on multivariate normality. We employed EMB in this study. EMB first uses the bootstrapping technique to repeatedly draw samples of the same size (with replacement) from the original sample and obtains the ML estimates of the mean and the covariance matrix through the expectation-maximization algorithm in each bootstrap sample. These estimates are then used to impute missing data. Interested readers are referred to Dempster et al. (1977) and Efron and Tibshirani (1986) for more information about the EM algorithm and bootstrapping.

### Multiple imputation based on predictive mean matching (MI-PMM)

We can also account for nonnormality by using MI strategies that rely less on the normality assumption. One such strategy is to impute missing values using the predictive mean matching (PMM) method implemented through the so-called MICE algorithm. MICE (a.k.a., multiple imputation by chained equations; van Buuren, 2007; van Buuren & Groothuis-Oudshoorn, 2011) does not impute missing data on multiple variables simultaneously, instead, it can individually specify the imputation model for each incomplete variable, and iteratively predict missing values on each variable conditional on the current values of the other variables. The flexibility of MICE makes it possible to adopt a variation of imputation models to accommodate missing data with different distributions. Predictive mean matching (PMM) is one of these imputation models. Iteratively, it predicts all values on each incomplete variable based on the current values of the other variables, then fills in the missing values on the variable by randomly drawing a “nearest” neighbor (i.e., one of candidate donors) on the same variable. PMM can have different versions depending on its computational settings, such as how to predict values on incomplete variable, and how to determine the “nearest” neighbors and the number of candidate donors. PMM has been found to be a reasonable MI method with nonnormal data in various scenarios (e.g., Di Zio & Guarnera, 2009; Kleinke, 2017; Morris et al., 2014), however, no study has examined its performance in producing the test statistic in SEM.

## Pooling approaches for test statistic

### D$_2$ and its variations

Li et al. (1991) proposed an approach of pooling $m$ Wald chi-square tests to generate a final significance test statistic. The pooled statistic produced from this approach is referred to as the $D_2$ statistic in the missing data literature (e.g., Enders, 2010; Schafer, 1997). Let $\overline{T}_W$ be the arithmetic average of $m$ Wald test statistics ( $T_{Wt}$, $t = 1, 2, \ldots$m):

$$\overline{T}_W = \frac{1}{m}\sum_{t=1}^{m} T_{Wt} \tag{8}$$

Then, the $D_2$ statistic is given by:

$$D_2 = \frac{\overline{T}_W k^{-1} - (m+1)(m-1)^{-1} ARIV_1}{1 + ARIV_1} \tag{9}$$

where $ARIV_1$ is an estimate of the average relative increase in variance with $k$ degrees of freedom and can be computed as follows.

$$ARIV_1 = \left(1 + m^{-1}\right)\left[\frac{1}{m-1}\sum_{t=1}^{m}\left(\sqrt{T_{Wt}} - \overline{\sqrt{T_w}}\right)^2\right] \tag{10}$$

where $\overline{\sqrt{T_w}}$ is the average of $m$ square root of $T_{Wt}$. The $D_2$ statistic approximates an F distribution with $k$ degrees of freedom for the numerator and $v$ degrees of freedom for the denominator, where

$$v = k^{-\frac{3}{m}}(m-1)\left(1 + \frac{1}{ARIV_1}\right)^2 \tag{11}$$

Li et al. (1991) mentioned that the type 1 error rate of the $D_2$ statistic can be quite sensitive to the fraction of missing information (FMI), which is a measure of the impact of missing data on the sampling variability of a parameter estimate (Enders, 2010).

$D_2$ can be easily implemented with SEM models, as it can directly pool $m$ Wald-like chi-square test statistics from $m$ imputed data sets (Jorgensen et al., 2021). Therefore, when a robust estimator is used in SEM analysis, the adjusted chi-square test statistics can be directly pooled using the $D_2$ approach. Liu and Sriutaisuk (2020) investigated the performance of $D_2$ for confirmatory factor analysis (CFA) models with ordinal items. They found that the $D_2$ method could adequately pool the adjusted chi-square test statistics unless all items are incomplete in the CFA model. The most influential factors to its performance were the number of response categories, factor loadings, and sample size. However, to our knowledge, no published study has examined how $D_2$ performs with on nonnormal continuous items in SEM. In this study, we examined $D_2$ and two of its variations. One is denoted by $D_{2A}$ (i.e., Asymptotic D$_2$), which is simply the product of $D_2$ and its numerator degree of freedom that transforms an F statistic to a chi-square asymptotically, so it could be further used to compute model fit indices. As it is an asymptotical method, we expect that it requires a large sample size to perform well. The other is $D_{2ASN}$ (i.e., asymptotic scaled naïve $D_2$), which pools the naïve (unadjusted) test statistic across $m$ imputations using $D_{2A}$ first and apply the average scale parameter (and shift parameter) to the pooled naïve test statistic. It is also an asymptotical method and can be directly used for computing model fit indices. Note that $D_{2ASN}$ is only available as chi-square because the scale/shift parameters cannot be applied to $F$.

### D$_3$ for nonnormal data

Another approach of pooling the test statistics is known as the Meng and Rubin’s (1992) approach or $D_3$ in the missing data literature (e.g., Enders, 2010; Schafer, 1997). The $D_3$ approach involves three steps. It first takes the average of the likelihood ratio test statistics across $m$ imputations.

$$\overline{T}_{LR} = \frac{1}{m}\sum_{t=1}^{m} T_{LRt} \tag{12}$$

where the likelihood ratio test statistic ($T_{\mathrm{LR}}$) compares the log-likelihood values from the structured and saturated models. In the second step, the two models are re-estimated with their model parameters constrained to the pooled values, and the averaged likelihood ratio test statistic across imputations is computed again ($\overline{T}_{Constrained}$). Finally, the pooled likelihood ratio test statistic is calculated as follows.

$$D_3 = \frac{\overline{T}_{Constrained}}{k\left(1 + ARIV_2\right)} \tag{13}$$

where $k$ is the number of parameter constraints, and $ARIV_2$ is another estimate of the average relative increase with $k$ degrees of freedom.

$$ARIV_2 = \frac{m+1}{k(m-1)}\left(\overline{T}_{LR} - \overline{T}_{Constrained}\right) \tag{14}$$

$D_3$ assumes normality. Enders and Mansolf (2018) conducted a comprehensive simulation to evaluate the application of $D_3$ in SEM with normal items. They found the pooled test statistics produced from $D_3$ were comparable to those from FIML in terms of type I error rate, even though $D_3$ was less efficient. However, it was still not clear in the literature how $D_3$ would perform with nonnormal missing items. $D_3$ can be also applied to nonnormal items. One can pool the naïve (unscaled) test statistic across $m$ imputations using $D_3$ first and apply the average scale parameter (and shift parameter) to the pooled test statistic. We refer to this approach as $D_{3SN}$ (i.e., scaled naïve D$_3$). Similar to $D_{2ASN,}$ $D_{3SN}$ is only available as chi-square. This approach has not been systematically evaluated in SEM.

## Simulation study

We conducted a simulation study to compare the performance of 24 strategies that involve the choices of robust estimators (MLR, MLM, and MLMV), imputation methods (MI-NORM and MI-PMM), and test statistic pooling approaches ($D_2$, $D_{2A}$, $D_{2ASN}$, and $D_{3SN}$) to deal with missing nonnormal data.

### Population model

We generated data from a three-factor SEM model (Fig. 1). Similar models are commonly seen in the SEM literature (e.g., Bollen, 1989; Enders, 2001; Enders & Mansolf, 2018; Palomo et al., 2007). In this population model, the three paths between latent variables were set at 0.4, 0.286, and 0.286, respectively. The variance/residual variances of the three latent variables $\eta 1$, $\eta 2$, and $\eta 3$ were set at 0.490, 0.412, and 0.378, respectively. All loadings were 1 and all residual variances on the items were set at 0.51.

### Design factors

Several factors were considered to vary in this simulation design, including sample size, missing data proportion, missing data mechanism, degree of nonnormality, and misspecification. We considered two analysis models: correct (i.e., non-misspecified) and misspecified models. The correct model was the same as the population model. In the misspecified model, we constrained the path coefficient from $\eta_1$ to $\eta_3$ at zero, following Enders and Mansolf (2018). Levels of sample size and missing data proportion were selected based on previous simulation studies and cases commonly seen in SEM. Specifically, we examined three sample sizes ($n$): small (150), medium (300) and large (600); and two missing data proportions ($mp$): small (15%) and large (30%). The other factors are explicitly described below.

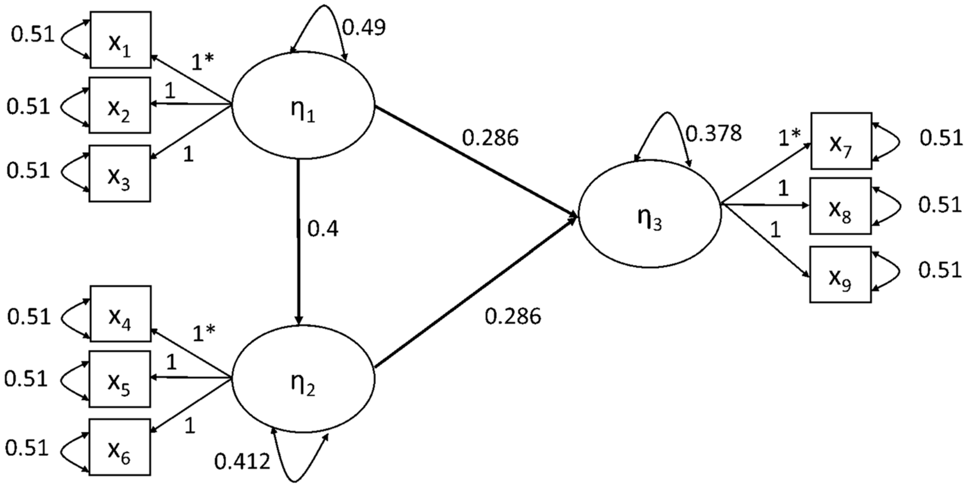

**Fig. 1** The structural equation model for data generation

### Nonnormality

For a univariate normal distribution, nonnormality can be measured using skewness ($S$) and kurtosis ($K$). The skewness of a normal distribution equals zero; a non-zero skewness indicates the asymmetry of a distribution about its mean. A kurtosis that deviates from 3 (or an excess kurtosis that deviates from 0) implies a flatter or more peaked distribution than a normal distribution. In this study, we generated nonnormal items following the method proposed by Foldnes and Olsson (2016) at three levels: mild ($S$ = 1.5, $K$ = 3), moderate ($S$ = 2, $K$ = 7), and severe ($S$ = 3, $K$ = 21). These levels of nonnormality were commonly seen in both simulation studies (e.g., Enders, 2010; Savalei & Falk, 2014) and applied research (Curran et al., 1996). The same level of nonnormality was applied to all items in each replication.

### Missing data mechanism

Missing data were generated on the first two items for each latent factor (i.e., $X_1$, $X_2$, $X_4$, $X_5$, $X_7$ and $X_8$) under one of three missing data mechanisms: MCAR, MAR-Head, and MAR-Tail. Under MCAR, missing data were randomly imposed so that every individual data point has the same probability of being missing. For MAR, we assumed that different missing data patterns were determined by different combinations of the complete items (i.e., weighted sum of $X_3$, $X_6$, and $X_9$; weights = 0 or 1 for). Specifically, for a subject who had incomplete items that loaded onto only one factor (e.g., $\eta_1$), the probability of missingness was only determined by the value of the complete item from the same factor (e.g., $1 \times X_3 + 0 \times X_6 + 0 \times X_9 = X_3$); for a subject who had incomplete items from more than one factors (e.g., $\eta_1$ and $\eta_2$), the probabilities of missingness were determined by the sum scores of the complete items from those factors (e.g., $1 \times X_3 + 1 \times X_6 + 0 \times X_9 = X_3 + X_6$). MAR-Head and MAR-Tail data were generated based on two different logistic distribution functions associated with the weighted sum scores. For MAR-Tail missingness, subjects with higher weighted sum scores had a higher probability to have missing values. In contrast, under MAR-Head, higher weighted sum scores resulted in lower probability of missingness.

For each of the 24 strategies, we examined their performance in 108 fully crossed conditions (2 models × 3 sample sizes × 2 missing data proportions × 3 degrees of nonnormality × 3 missing data mechanisms). In each condition, we generated 1000 replicated samples, and evaluated empirical type I error rate and power for the correct and misspecified models, respectively. For the correct model, we computed the type I error rate as the proportion of replicated samples with a significant test statistic ($p < 0.05$). Conventionally, a type I error rate below 0.1 has been considered "acceptable". In addition, following Bradley's (1978) "liberal criterion", we considered a type I error rate "accurate" if it is between 0.025 and 0.075. Empirical power was computed in the same manner for the misspecified model. We only report the values of empirical power in the conditions where the type I error rates from the correct model fell within the "accurate" range.

Data were generated using R (R Core Team, 2022) following the Foldnes & Olsson, (2016) method. MI-NORM and MI-PMM were implemented using the R packages Amelia (Honaker et al., 2011) and mice (van Buuren & Groothuis-Oudshoorn, 2011; with 20 burn-in and five donors), respectively. For both imputation methods, 20 imputed data sets were generated. All pooling procedures were implemented through R package semTools (Jorgensen et al., 2021)[2].

## Results

### Correct model

### Complete data

We first assessed the type I error rates produced by different estimation methods given data were complete. We expected that these results could later help us partial-out the impact of estimation methods when examining the impacts of imputation method and pooling approaches with incomplete data. Figure 2 shows the average type I error rates across replications generated from three estimation methods (columns of the panels), three degrees of nonnormality (rows of the panels), and three sample sizes (on x-axis in each panel). The solid horizontal line represents the nominal type I error rate of 0.05. The two dotted horizontal lines represent the limits of the "accurate" range of type I error rate, 0.025 and 0.075. We found that MLR test statistic only worked marginally well with the large sample size (600) and mild nonnormality. It generated severely inflated type I error rates in all other conditions. In contrast, type I error rates from MLM and MLMV were more appropriate across sample sizes and degrees of nonnormality.

### Incomplete data

**Overall patterns** Figure 3 contains violin plots and shows average type I error rates across replications with incomplete data generated from three estimation methods (columns of the panels), four pooling approaches (rows of the panels) and two imputation methods (on $x$-axis in the panels). The violin

[2] The imputation-related functionality in *semTools* package will be moved to the new package *lavaan.mi* written by the same authors at the time of publication or soon after.

plot is analogous to the boxplot that illustrates the central tendency and variability of data. In addition, the violin plot depicts the frequency density of data distribution, that is, the thicker parts indicate that data occur more frequently, i.e., higher frequency; and the thinner parts reflect lower frequency. In Fig. 3, each dot is the average type I error rate across replications within one design cell. Each "violin" shows the distribution of 54 dots (3 sample sizes × 2 missing data proportions × 3 degrees of nonnormality × 3 missing data mechanisms). The same as in Fig. 2, the solid horizontal line and two dotted horizontal lines represent type I error rates of 0.05, 0.025 and 0.075, respectively. Because the poor performance of MLR with complete data has been already observed, it is not surprising that MLR type I error rates with missing data also demonstrated much larger variability than those produced from MLM and MLMV, regardless of the imputation methods or pooling approaches. For both MLM and MLMV, the two scaled-naïve pooling approaches (i.e., $D_{2ASN}$ and $D_{3SN}$) show smaller variabilities (see the four panels in the bottom-right corner in Fig. 3), implying more stable performance than the other two $D_2$ methods (i.e., $D_2$ and $D_{2A}$). Comparing the two imputation methods, we found that MI-PMM tended to generate lower type I error rates than MI-NORM. Next, we will further look inside the four bottom-right panels.

**MLM with $D_{2ASN}$ and $D_{3SN}$** To further examine the type I error rates produced by the pooled MLM test statistics, we organized the results in two nine-panel figures. Figure 4 contains results for MLM test statistics pooled through $D_{2ASN}$. The rows of the panels represent the three degrees of nonnormality. The columns are three missing data mechanisms, MCAR, MAR-Head, and MAR-Tail. The *X* and *Y* axes in each panel are sample size with missing data proportion and type I error rate, respectively. The two missing imputation methods are displayed using two different shapes in the panels. We found the type I error rates based on MI-NORM were most inflated with mild nonnormality (> 0.1) and to a lesser extent with moderate nonnormality (> 0.075). As data became more severely nonnormal, the inflated type I

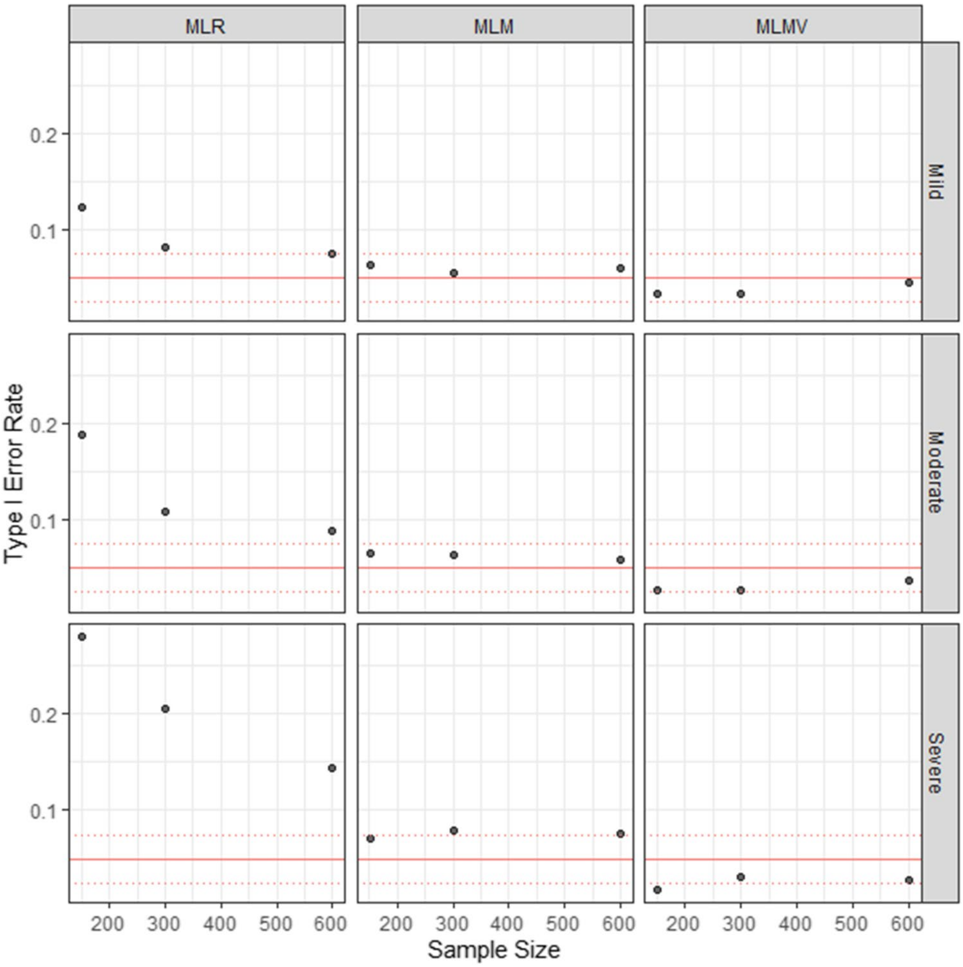

**Fig. 2** Type I error rate with complete data

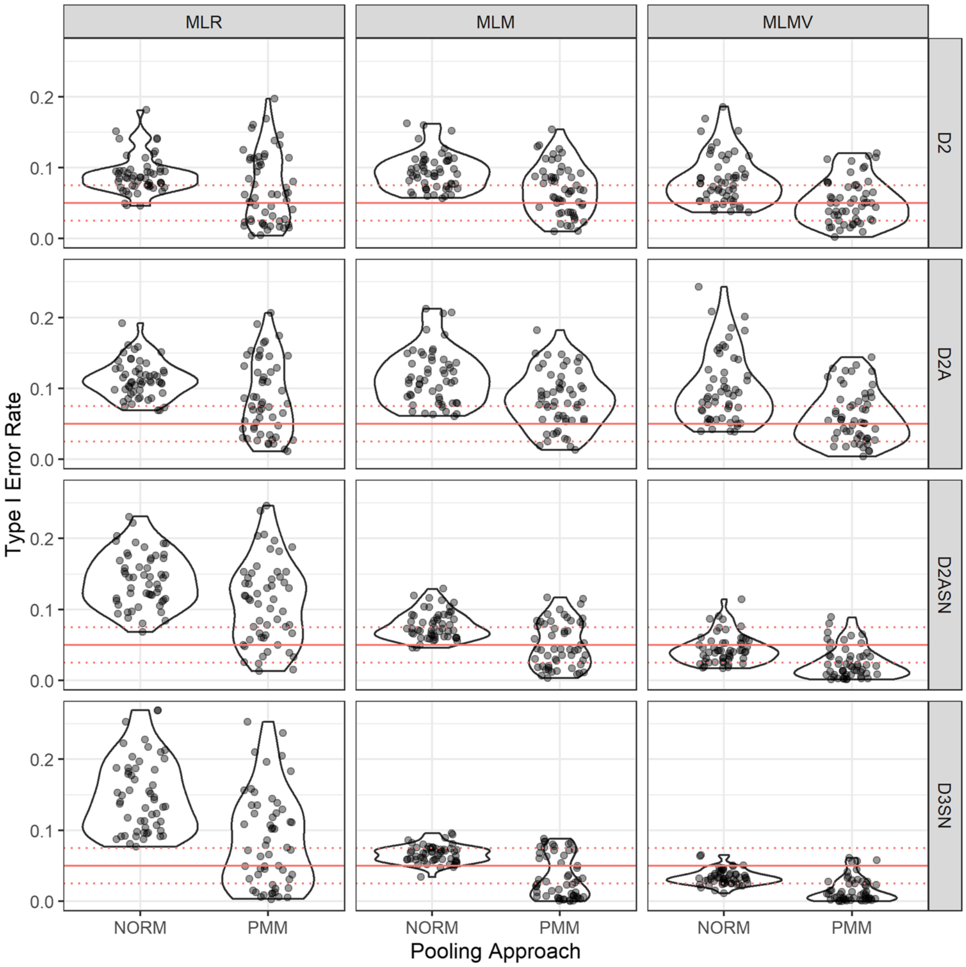

**Fig. 3** Type I error rate with incomplete data

error rates shrunk to a reasonable range. MI-PMM worked the best when missing data were MCAR. Under MAR, its performance fluctuated largely across conditions. Generally, it tended to yield deflated type I error rates under MAR-Head and inflated type I error rates under MAR-Tail. These results indicate that the MLM with $D_{2ASN}$ is not an ideal approach, unless data were missing completely at random (MCAR) with non-severe nonnormality, and MI-PMM was used for imputation.

We depict results of MLM test statistics pooled through $D_{3SN}$ in Fig. 5, using the same structure as Fig. 4. The MI-NORM results were more reasonable. The type I error rates generally hovered around the upper limit of the "accurate" range (i.e., 0.075), and all of them fall within the "acceptable" range (< 0.1). The performance of MI-PMM was dependent on the missing data mechanism. Under MCAR, the type I error rates hovered around the lower limit of the "accurate" range (i.e., 0.025); under MAR-Head, all type I error rates dropped below the 0.025 cutoff. When missing data were MAR-Tail, the type I error rates were comparable to those from MI-NORM, which hovered around the upper limit of the "accurate" range (i.e., 0.075). The results in Fig. 5 indicates that MLM with $D_{3SN}$ was more desirable when working with the normality-based MI, especially when sample size was medium to large and missing data were not MAR-Tail. MI-PMM tended to produce more "marginally accurate" or "lower-than-accurate" type I error rates.

**MLMV with $D_{2ASN}$ and $D_{3SN}$** With MLMV, the performance of imputation methods and the pooling approaches show different patterns than those with MLM. Figures 6 and 7 are used to illustrate these new patterns. Figure 6 contains the results for MLMV test statistics pooled through $D_{2ASN}$. MI-NORM generally worked adequately, especially with moderate nonnormality; with mild nonnormality, type I error rates could exceed the upper limit of the "accurate" range (0.075) and even the "acceptable" limit (0.1) when $n$ = 600 and $mp$ = 0.3; when data were severely nonnormal, type I error rates decreased and fell around the lower limit of the acceptable range (0.025). MI-PMM only worked

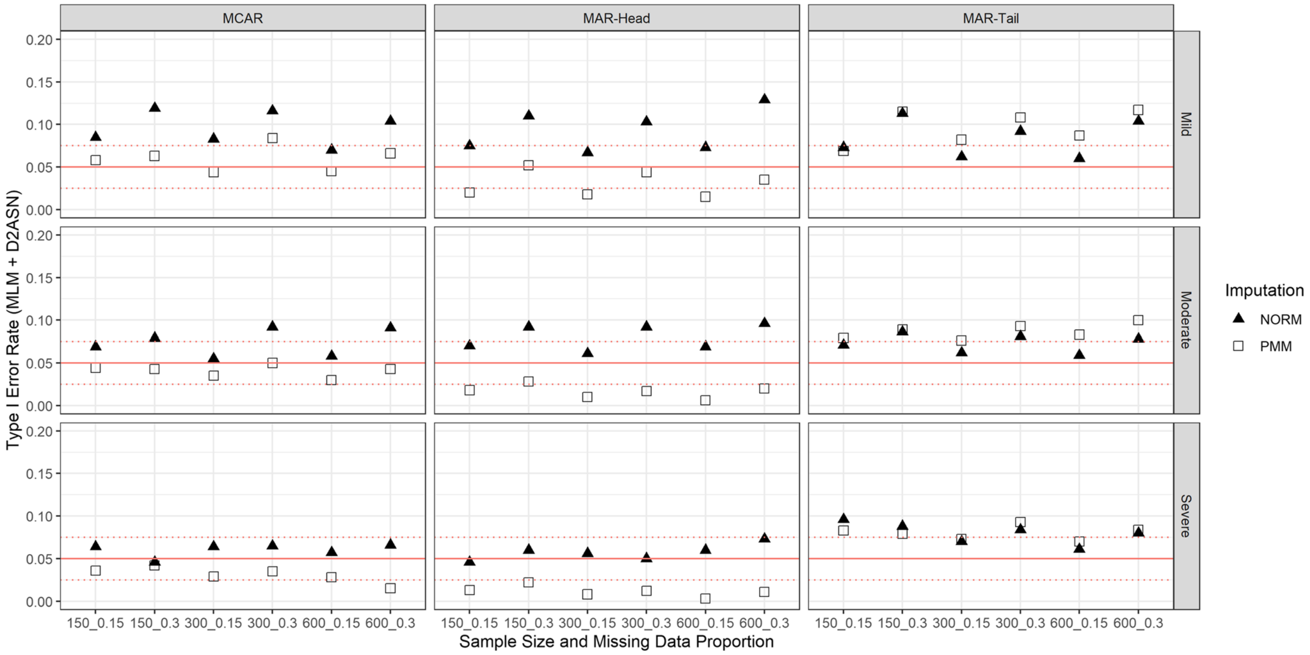

Fig. 4 Type I error rate produced from MLM and $D_{2ASN}$

well in a small set of conditions, in which nonnormality was mild under MCAR, or under MAR-Tail when missing date proportion was small.

Figure 7 shows the type I error rates generated from MLMV with $D_{3SN}$. Similar to the findings in the previously discussed situation, MI-NORM was generally superior to MI-PMM. Different from the patterns found with $D_{2ASN}$, however, in this case MI-NORM worked the best with mildly nonnormal data or under MAR-Tail. With moderate and severe nonnormality, type I error rates from MI-NORM could be marginally "accurate" or slightly below 0.025 under MACR or MAR-Head, typically with small to medium sample sizes. In contrast, MI-PMM worked well only for MAR-Tail, when nonnormality was mild, sample size was large, or missing data proportion was small.

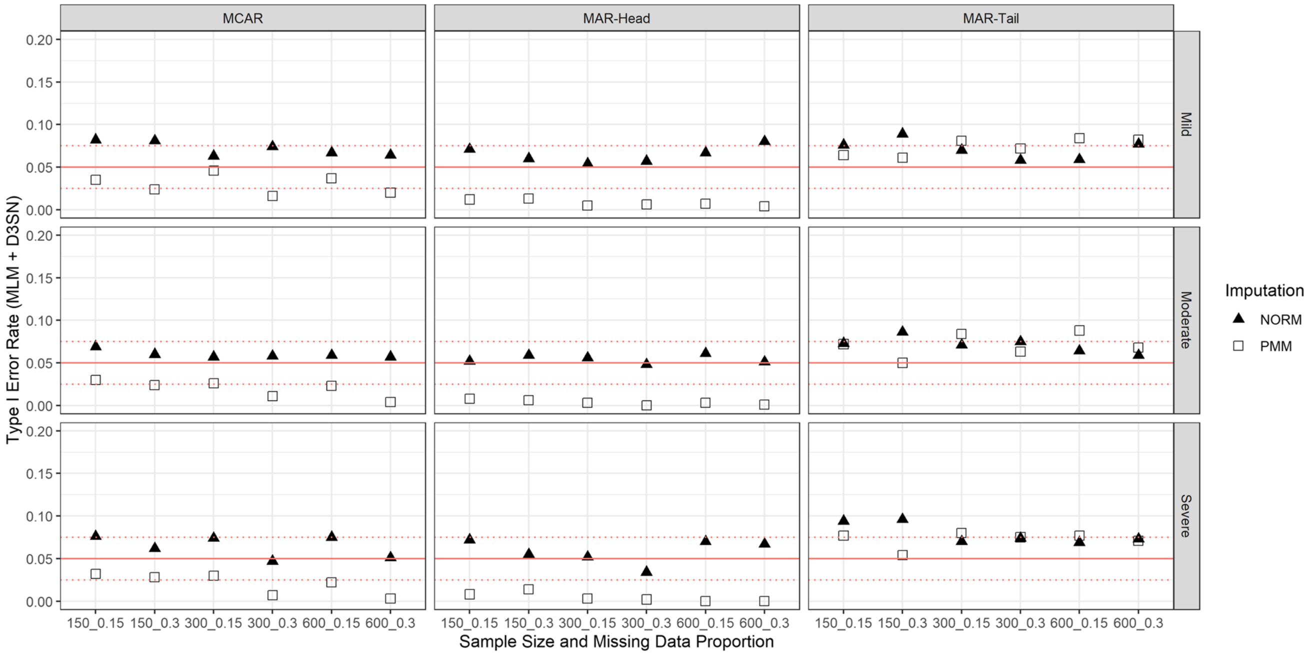

Fig. 5 Type I error rate produced from MLM and $D_{3SN}$

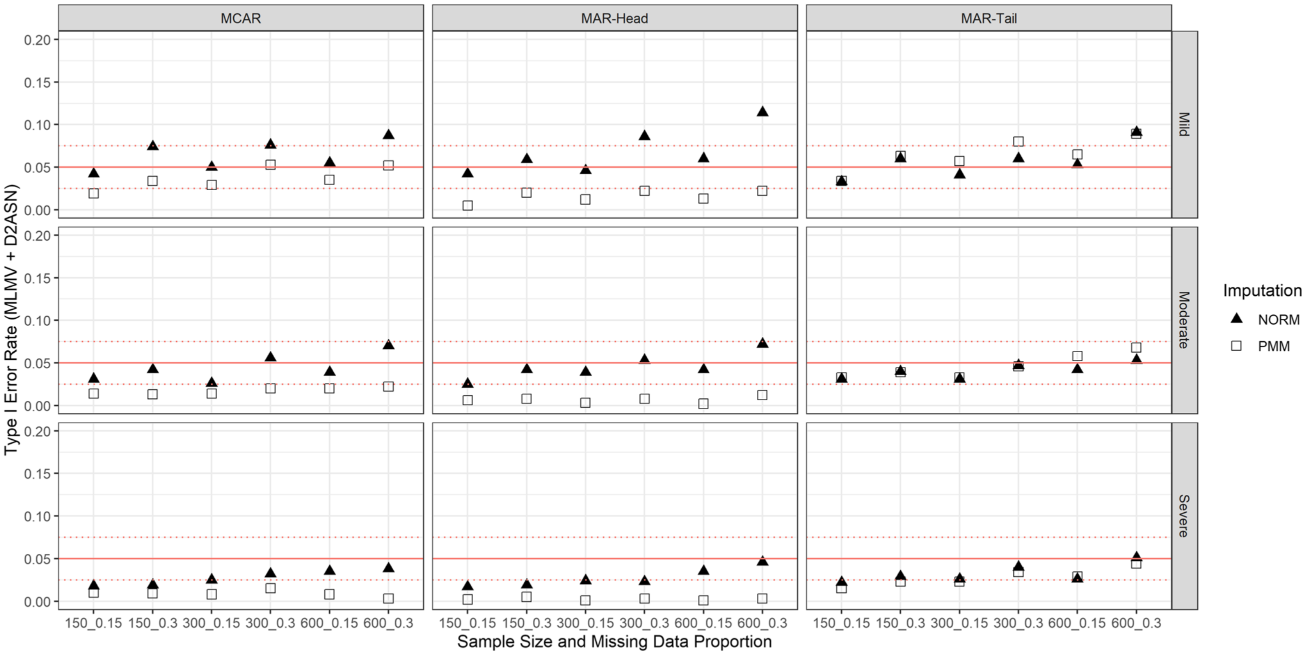

Fig. 6 Type I error rate produced from MLMV and $D_{2ASN}$

## Misspecified model

### Complete data

As noted, when there is misspecification in the analytic model, the empirical power of the test statistic is defined as the proportion of replications in which the misspecified model is rejected. In Table 1, we report values of empirical power with complete data only for MLM and MLMV, as MLR failed to produce "accurate" type I error rates in most conditions (see Fig. 2). It shows that MLMV generally required larger sample sizes than MLM to achieve a

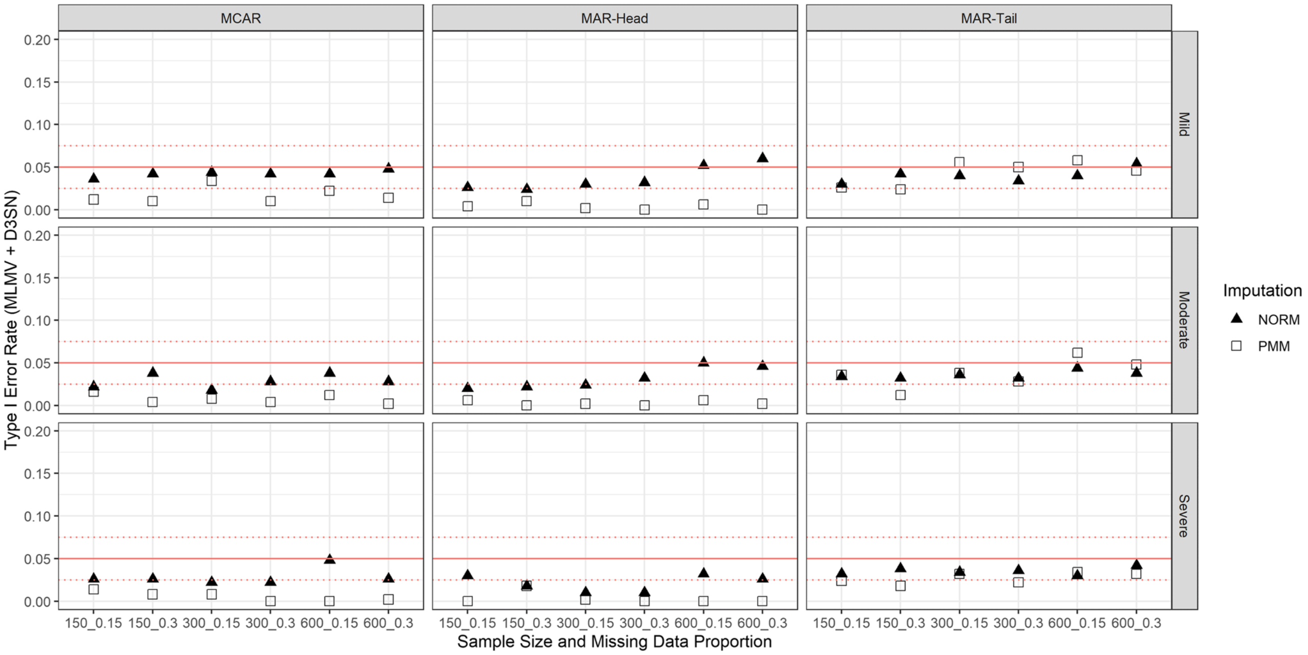

Fig. 7 Type I error rate produced from MLMV and $D_{3SN}$

**Table 1** Power to detect misspecification with complete data

| Nonnormality and *n* | | MLM | MLMV |
|---|---|---|---|
| Mild | 150 | 0.23 | 0.13 |
| | 300 | 0.44 | 0.36 |
| | 600 | 0.79 | 0.76 |
| Moderate | 150 | 0.23 | 0.10 |
| | 300 | 0.42 | 0.26 |
| | 600 | 0.81 | 0.74 |
| Severe | 150 | 0.24 | -- |
| | 300 | -- | 0.23 |
| | 600 | 0.75 | 0.58 |

The empty cells indicate conditions with inaccurate type I error rates

desirable power (i.e., 0.8), however, the differences between their power values became smaller as the sample size increased, or as the nonnormality became less severe.

### Incomplete data

The same as in compete data results, we examined empirical power of the strategies only in the conditions where they could generate "accurate" type I error rates for the correct model. Given the results we obtained earlier, there were only four strategies that produced "accurate" type I error rates in a majority of conditions: MLM or MLMV combining with $D_{2ASN}$ or $D_{3SN,}$ using MI-NORM. Table 2 shows that the MLM generally has higher empirical power than MLMV; and $D_{2ASN}$ almost always possessed higher empirical power than $D_{3SN}$. The power values of all strategies were most impacted by sample size, followed by missing data proportion and degree of nonnormality, but they did not vary too much across missing data mechanisms. Larger sample size not only raised power but also reduced differences in power across conditions/methods. The smallest power values ($< 0.1$) were found with small sample size (150) and large missing date proportion (30%), with moderate or severe nonnormality. Higher power values were found with large sample size, less amount of missing data, and less severe nonnormality. Compared to complete data results, there were noticeable decreases in power even with a small amount of missingness.

### Additional simulation results with plain ML

The results obtained from the robust ML estimators (MLM/MLMV) with missing data show an interesting pattern that the type I error rates were more inflated with mild nonnormal data than those with more severe nonnormal data, especially for MI-NORM and $D_{2ASN}$ (see Figs. 4 and 6). This calls into question whether it would be wiser to pool the naïve (unscaled) test statistics without any robust adjustment with

**Table 2** Power to detect misspecification with missing data through MI-NORM with MLMV

| Nonnormality | | MCAR | | | | MAR-Head | | | | MAR-Tail | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MLM | | MLMV | | MLM | | MLMV | | MLM | | MLMV | |
| | | $D_{2ASN}$ | $D_{3SN}$ | $D_{2ASN}$ | $D_{3SN}$ | $D_{2ASN}$ | $D_{3SN}$ | $D_{2ASN}$ | $D_{3SN}$ | $D_{2ASN}$ | $D_{3SN}$ | $D_{2ASN}$ | $D_{3SN}$ |
| Mild | $n = 150$, mp = 0.15 | - | - | 0.13 | 0.11 | 0.20 | 0.18 | 0.12 | 0.09 | 0.24 | - | 0.15 | 0.12 |
| | $n = 150$, mp = 0.3 | - | - | 0.16 | 0.09 | - | 0.12 | 0.14 | 0.06 | - | - | 0.15 | 0.10 |
| | $n = 300$, mp = 0.15 | - | 0.35 | 0.34 | 0.29 | 0.39 | 0.34 | 0.32 | 0.26 | 0.39 | 0.37 | 0.32 | 0.29 |
| | $n = 300$, mp = 0.3 | - | 0.27 | - | 0.20 | - | 0.23 | - | 0.18 | - | 0.29 | 0.30 | 0.23 |
| | $n = 600$, mp = 0.15 | 0.61 | 0.64 | 0.63 | 0.62 | 0.61 | 0.61 | 0.61 | 0.59 | 0.62 | 0.63 | 0.63 | 0.64 |
| | $n = 600$, mp = 0.3 | - | 0.51 | - | 0.49 | - | - | - | 0.44 | - | - | - | 0.51 |
| Moderate | $n = 150$, mp = 0.15 | 0.21 | 0.19 | 0.09 | - | 0.18 | 0.16 | 0.09 | - | 0.23 | 0.20 | 0.09 | 0.10 |
| | $n = 150$, mp = 0.3 | - | 0.11 | 0.09 | 0.04 | - | 0.09 | 0.07 | 0.03 | - | - | 0.10 | 0.08 |
| | $n = 300$, mp = 0.15 | 0.34 | 0.32 | 0.26 | - | 0.35 | 0.31 | 0.25 | 0.20 | 0.36 | 0.35 | 0.24 | 0.23 |
| | $n = 300$, mp = 0.3 | - | 0.21 | 0.22 | 0.13 | - | 0.20 | 0.20 | - | - | 0.27 | 0.22 | 0.18 |
| | $n = 600$, mp = 0.15 | 0.62 | 0.63 | 0.61 | 0.57 | 0.62 | 0.63 | 0.62 | 0.57 | 0.64 | 0.65 | 0.63 | 0.61 |
| | $n = 600$, mp = 0.3 | - | 0.50 | 0.48 | 0.43 | - | 0.44 | 0.46 | 0.35 | - | 0.55 | 0.51 | 0.48 |
| Severe | $n = 150$, mp = 0.15 | 0.18 | - | - | - | 0.14 | 0.15 | - | - | - | - | 0.08 | 0.08 |
| | $n = 150$, mp = 0.3 | 0.10 | 0.09 | - | 0.04 | 0.09 | 0.08 | - | 0.03 | - | - | 0.07 | 0.06 |
| | $n = 300$, mp = 0.15 | 0.31 | 0.32 | 0.18 | 0.15 | 0.27 | 0.24 | - | - | 0.36 | 0.36 | 0.20 | 0.17 |
| | $n = 300$, mp = 0.3 | 0.23 | 0.21 | 0.13 | - | 0.15 | 0.11 | - | - | - | 0.29 | 0.17 | 0.14 |
| | $n = 600$, mp = 0.15 | 0.54 | 0.54 | 0.46 | 0.44 | 0.52 | 0.53 | 0.46 | 0.42 | 0.59 | 0.61 | 0.50 | 0.48 |
| | $n = 600$, mp = 0.3 | 0.43 | 0.45 | 0.36 | 0.29 | 0.35 | 0.33 | 0.28 | 0.22 | - | 0.50 | 0.42 | 0.38 |

The empty cells indicate conditions with inaccurate type I error rates

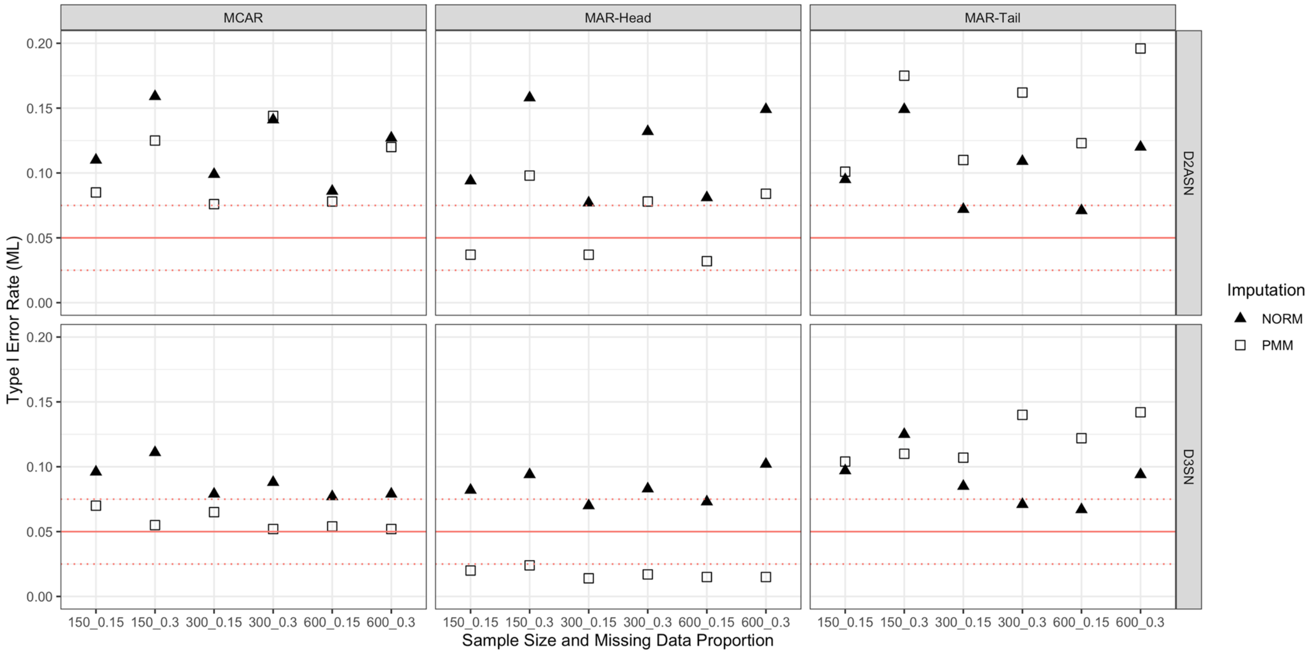

**Fig. 8** Type I error rates produced from plain ML with $D_{2ASN}$ and $D_{3SN}$ for mildly nonnormal data only

mild nonnormality. Therefore, we conducted additional simulations to examine how the plain (i.e., non-robust) ML work when $S = 1.5$, and $K = 3$. Figure 8 depicts the type I error rates obtained from plain ML with MI-NORM or MI-PMM combining with $D_{2ASN}$ or $D_{3SN}$, across different missing data mechanisms, sample sizes, and missing data proportions. We found that with plain ML, type I error rates were "accurate" only when $D_{3SN}$ were used under MCAR (see the bottom left panel of Figure 8). In the conditions where the type I errors were more likely to be inflated with MLM/MLMV (Figs. 4 and 6), using plain ML did not seem to be a better choice.

## Empirical example

To illustrate these strategies, we used a data set retrieved from the Fragile Families and Child Well-Being Study (FFCWS; Reichman et al., 2001). Inspired by Marchand-Reilly & Yaure (2019), we focused on the relations among three latent variables that were similar to the population model in the simulation study: parents' relationship at child's age 5 ($\eta_1$) predicted child's internalizing behaviors at age 5 ($\eta_2$), and both $\eta_1$ and $\eta_2$ predicted child's internalizing behaviors at age 15 ($\eta_3$). Table 3 shows the descriptive statistics of the indicators. We selected a subsample of 940 subjects with no missing values, and then imposed 15% missingness on one indicator of each latent variable under MAR-Tail (i.e., more missing data occurred in the tail). The correct model was fitted to this incomplete data set.

The chi-square values produced from the examined strategies are shown in Table 4. We did not include $D_2$ because it produces F statistic rather than chi-square. We found that MI-PMM generally underestimated the chi-square statistic. The combination of MI-NORM, MLMV, and $D_{3SN}$ produced the chi-square statistic closest to that from the complete data.

**Table 3** Empirical example: Descriptive statistics

| Construct | Indicator | Min | Max | Skewness | Excess Kurtosis | Missingness |
|---|---|---|---|---|---|---|
| Parents relationship at age 5 ($\eta_1$) | Co-parenting quality ($x_1$) | 2.17 | 4 | –2.01 | 5.2 | Yes |
| | Relationship quality ($x_2$) | 1.22 | 4.11 | –0.93 | 0.91 | No |
| Child's internalizing behaviors at age 5 ($\eta_2$) | Anxious/depressed ($x_3$) | 1 | 2.56 | 1.47 | 2.88 | Yes |
| | Withdrawn/depressed ($x_4$) | 1 | 2.14 | 1.24 | 1.66 | No |
| Child's internalizing behaviors at age 15 ($\eta_3$) | Anxious/depressed ($x_5$) | 1 | 3 | 1.9 | 4.4 | Yes |
| | Withdrawn/depressed ($x_6$) | 1 | 3 | 1.78 | 2.93 | No |

**Table 4** Empirical example: Chi-square values from the examined strategies under MAR-Tail

| | Complete data | $D_{3SN}$ | | $D_{2ASN}$ | | $D_{2A}$ | |
|---|---|---|---|---|---|---|---|
| | | NORM | PMM | NORM | PMM | NORM | PMM |
| MLR | 28.76 | 29.32 | 23.76 | 33.84 | 22.20 | 34.22 | 22.74 |
| MLM | 28.45 | 33.28 | 24.49 | 27.13 | 24.83 | 28.10 | 24.88 |
| MLMV | 27.93 | 27.90 | 19.06 | 26.32 | 13.73 | 27.78 | 14.19 |

*NORM* multiple imputation based on normality; *PMM* multiple imputation based on predictive mean matching

## Discussion

We conducted a simulation study to investigate the performance of 24 combinations of different robust estimators, missing data imputation methods, and pooling approaches. Our focus was the empirical type I error rate and power of the pooled test statistics with nonnormal items. The goal of the study is to uncover the optimal combination of these techniques that can help substantive researchers make a series of decisions when dealing with missing nonnormal data in SEM.

With complete nonnormal data, we found that MLR produced largely inflated type I error rates in almost all conditions. In contrast, MLM and MLMV performed quite well in controlling type I error rate. These results echo the findings of Maydeu-Olivares (2017), which inspired us to further investigate the viability of MLM and MLMV in the missing data context. Unlike MLR, the two other estimators MLM and MLMV do not directly work with missing data. To circumvent this obstacle, we incorporated another technique, multiple imputation (MI), which can help make use of MLM and MLMV for missing data.

With missing nonnormal data, several options are available to implement the three steps in MI: imputation, analysis, and pooling. Comparing between the two imputation methods, we found that the normal-theory-based imputation method (MI-NORM) generally produced larger type I errors than MI-PMM. Their comparative performance, however, was dependent on the selections of the other approaches in the following steps. In the analysis step, when MLR was used, neither MI-NORM nor MI-PMM worked well, as large variabilities in type I error rates were observed across conditions, and the type I error rates were drastically affected by factors such as sample size, missing data mechanism and missing data proportion, no matter which pooling approach was used. We exclude MLR in further discussion for this reason. MLM and MLMV, on the other hand, led to smaller variabilities in type I error rates across conditions or both MI-NORM and MI-PMM, especially when they were in conjunction with $D_{2ASN}$ or $D_{3SN}$ in the pooling step. Comparing between MLM and MLMV, we found that MLM generally produced "acceptable" type I error rates (< 0.1) with a few exceptions, while MLMV yield more "accurate" type I error rates (0.25–0.75). MLM, however, produce higher values power in general, regardless of the pooling approaches. This finding also aligned with the complete data results. When it came to the comparison among the four pooling approaches with MLMV, we found that the approaches that directly pooled the robust (adjusted) test statistics ($D_2$ and $D_{2A}$) were not viable options, as the type I error rates still largely fluctuated across conditions. The naïve pooling approaches ($D_{2ASN}$ and $D_{3SN}$) tended to be superior, as their type I error rates were more clustered around the "accurate" region (0.25–0.75). More specifically, with $D_{2ASN}$ and $D_{3SN}$, MI-NORM produced more accurate type I error rates; while MI-PMM tended to yield nearly zero type I error rates, implying that it would likely lack the power to detect model misspecifications.

Among the 24 examined strategies, two clearly outperformed the others in controlling the type I error rate within a reasonable range. These strategies use MI-NORM for imputation, $D_{3SN}$ for pooling, and either MLM or MLMV in the analysis phase. The strategy with MLM generally produced accurate type I errors with medium or larger samples; and inflated type I errors could be observed with small samples. Using MLMV instead, it would be less likely to produced inflated type I errors in all conditions; and the type I error rates could be too low for moderate or severe nonnormality, and therefore this strategy possessed lower statistical power. However, as we have discovered, larger sample size could reduce the differences in power among conditions/methods. Under conditions with accurate type I error rates, one might expect that the power difference would become negligible as sample size increases.

This simulation study explored how the current available options in semTools (Jorgensen et al., 2021) work in pooling the test statistics with different robust estimators and MI methods. To keep our study in a manageable scope, we focused on continuous nonnormal data with three ML-based robust estimators and two MI methods. Future research could compare strategies for ordinal data, e.g., those with ML- vs. least-squares-based estimators for ordered five-category data, to examine a common debate in practice for choosing whether to treat ordinal variables as continuous or make a latent-response assumption. Moreover, we only used one population model and investigated a single type of misspecification (same as in Enders & Mansolf, 2018). The empirical power of these methods could be further explored by considering more complex model and different types and locations of misspecification. In addition, we only examined one version of MI-PMM. Change in computational settings may alter the performance of MI-PMM. Other types of MI methods may also be worth considering in future studies. As noted in Enders and Mansolf (2018), a restrictive imputation method (using a structured model) may be helpful to improve power. Finally, practical fit indices, such as RMSEA, CFI and TLI, were not included in this study. The pooling approaches for these indices involve more methodological details and warrant further research.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02088-3.

## References

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. In: *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference, Arlington, VA*. https://www.statmodel.com/download/2005FCSM.pdf. *Accessed 1 Feb 2022.*

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Mplus Technical Appendix*. Retrieved February 1, 2022, from https://www.statmodel.com/download/WLSMV_new_chi21.pdf.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

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

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

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.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Di Zio, M., & Guarnera, U. (2009). Semiparametric predictive mean matching. *AStA Advances in Statistical Analysis, 93*(2), 175–186.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

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

Foldnes, N., & Olsson, U. H. (2016). A Simple Simulation Technique for Nonnormal Data with Prespecified Skewness, Kurtosis, and Covariance Matrix. *Multivariate Behavioral Research, 51*(2–3), 207–219.

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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling. Retrieved fromhttps://CRAN.R-project.org/package=semTools. Accessed 20 Feb 2022.

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.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*(1), 65–92.

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

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

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 383–394.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103–111.

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), 1–13.

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

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. (2007). Bayesian structural equation modeling. In Lee, S. Y. (Ed.), *Handbook of latent variable and related models* (pp. 163–188). Elsevier

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

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

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 399–419). Thousand Oaks.

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

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling*, *12*(2), 183–214.

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 sata. *Structural Equation Modeling: A*

*Multidisciplinary Journal*, *21*(2), 280–302. https://doi.org/10.1080/10705511.2014.882692

Savalei, V., & Rosseel, Y. (2021). Computational options for standard errors and test statistics with incomplete normal and nonnormal data in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2021.1877548

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

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

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

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.

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.

**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.