Behavior Research Methods (2022) 54:1063–1077
https://doi.org/10.3758/s13428-021-01582-w

Evaluating FIML and multiple imputation in joint ordinal-continuous measurements models with missing data

Aaron J.-M. Lim[1] · Mike W.-L. Cheung[1]

Accepted: 17 March 2021 / Published online: 20 September 2021

Abstract

Missing data is a common occurrence in confirmatory factor analysis (CFA). Much work had evaluated the performance of different techniques when all observed variables were either continuous or ordinal. However, few have investigated these techniques when observed variables are a mix of continuous and ordinal variables. This study investigated the performance of four approaches to handling missing data in these models: a joint ordinal-continuous full information maximum likelihood (FIML) approach and three multiple imputation approaches (fully conditional specification, fully conditional specification with latent variable formulation, and expectation-maximization with bootstrapping) combined with the weighted least squares with mean and variance adjustment (WLSMV) estimator. In a Monte-Carlo simulation, the FIML approach produced unbiased estimations of factor loadings and standard errors in almost all conditions. Fully conditional specification combined with WLSMV was second best, producing accurate estimates when the sample size was large. However, FIML encountered slight non-convergence issues when certain ordinal categories have extremely low frequencies, which is typical of skewed data. If the sample is large, fully conditional specification combined with weighted least squares is recommended when the FIML approach is not feasible (e.g., non-convergence, impractical computation durations, and variables that predict missingness are not of interest to the analysis).

Keywords Missing data · Full information maximum likelihood · Multiple imputation · Joint ordinal continuous · Confirmatory factor analysis

Confirmatory factor analysis (CFA) is widely used in psychological research for its ability to estimate relationships between unobservable latent constructs (factors) and observable indicators (Kline, 2016). Validation of a scale's factor structure (Borkenau & Ostendorf, 1990; Caprara et al., 1993; Digman, 1997), estimation of reliability (Leite et al., 2010), testing of measurement invariance (Meredith, 1993) or assessing the measurement model as part of structural equation modelling (SEM; Nagengast et al., 2011) are some common but vital applications of CFA. Missing data is another common occurrence in psychological research, and methods to address missing data in CFA are of interest and relevant to many applied researchers. Numerous studies have investigated the performance of different missing data techniques in CFA (e.g., Chen et al., 2019; Enders & Bandalos, 2001; Jia & Wu, 2019; Rosseel, 2012; Shi et al., 2019, 2020). However, most studies only investigated measurement models containing either continuous or ordinal indicators, but not both.

Models containing both ordinal and continuous indicators are of interest as researchers integrate different measurements of the same construct. This paper will refer to these models as joint ordinal-continuous (JOC) models for brevity. Applications of JOC models include situations when a construct is measured by both behavioural (objective) and self-reported (subjective) indicators and researchers wish to integrate both forms of measurement. Some examples include recovery from physical injury (MacDermid et al., 2000). Recovery can be measured behaviourally and on a continuous scale by grip strength, range of motion, and dexterity of the affected areas. But there are also subjective and self-reported aspects of recovery, such as inconvenience to daily functioning, which are measured using Likert scales. Other examples include the study of pain (Loggia et al., 2011), which can be measured with neuroscientific equipment

✉ Mike W.-L. Cheung
mikewlcheung@nus.edu.sg

[1] Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570, Singapore

such as heart rate and skin conductance and with self-reported pain intensity and unpleasantness scales. Cognitive ability measures using cognitive tasks also present an interesting case, as both response time and response accuracy measure cognitive ability (Vandierendonck, 2017) but are recorded on a continuous and binary scale, respectively. Situations where indicators of one construct are a mix of both parcelled and individual Likert scales (e.g., Duncan et al., 2001) also motivate JOC models.

Existing missing data approaches can be adapted to handle JOC models. Full information maximum likelihood (FIML) and multiple imputation are two families of techniques that are considered the best in the field of missing data (Schafer & Graham, 2002). However, the performance of existing missing data techniques in JOC models is less studied. The inclusion of both continuous and ordinal variables presents challenges to some missing data techniques, e.g., multiple imputation approaches that impute from a joint distribution have to define the joint distribution which describes the mix of continuous and ordinal variables. In this paper, we focus on a recently proposed approach to implementing FIML in JOC models by Pritkin, Brick, and Neale (Pritikin et al., 2018), which circumvents the computational limits from existing FIML approaches when applied to models with ordinal indicators. This new approach is implemented in R through the OpenMx package (Neale et al., 2016). We also focused on three implementations of multiple imputation: expectation-maximization with bootstrapping (EMB) as implemented in the Amelia package (Honaker et al., 2011), fully conditional specification as implemented via chained equations in the mice package (FCS; van Buuren & Groothuis-Oudshoorn, 2011), and fully conditional specification with a latent variable formulation as implemented in the Blimp software (FCSLV; Enders et al., 2018). The multiple imputation approaches were combined with the weighted least squares with mean and variance adjustment (WLSMV) estimator from the lavaan package (Rosseel, 2012). These packages are available in the R environment (R Development Core Team, 2018), which is free and easily accessible. Blimp is free for MacOS and Windows operating systems at http://www.appliedmissingdata.com/multilevel-imputation.html.

Objectives

This study aims to evaluate the performance of four different approaches to analysing JOC models with missing data. Using a simulation study, we manipulated factors of the dataset across approaches such that recommendations can be tailored to specific conditions. The study also tests three additional approaches combining multiple imputation and FIML, such that we can attribute performance differences to either the missing data method or estimator. The eighth approach was an ad-hoc missing data method, combining pairwise deletion (PD) with WLSMV. The eight approaches are summarized in Table 1.

Applied researchers who wish to analyse JOC models may not have a strong basis for adopting certain approaches when there is missing data. We seek to provide guidelines by addressing the following research questions:

> **Question 1:** What are the effects of sample size, missingness, number of categories, and distribution of response categories on convergence rates, factor loadings, standard errors, and accuracy?
> **Question 2:** Amongst the approaches studied, which performs the best?
> **Question 3**: Are differences in performance driven by differences in missing data approaches or CFA estimators?

The paper begins with an introduction to the weighted least squares with mean and variance adjustment (WLSMV) estimator for CFA. Next, the paper will introduce FIML as implemented by Pritkin et al. (Pritikin et al., 2018), followed by an introduction to multiple imputation in mice, Amelia, and Blimp. We briefly compare FIML versus multiple imputation and present a short empirical illustration. The design of the simulation study and performance measures will be reviewed. Lastly, the paper will end with a result and discussion section summarizing the simulation results, provide recommendations, and lay out future research directions.

Weighted least squares with mean and variance adjustment

Maximum likelihood (ML) is the default estimator for CFA in many programs. But ML assumes that indicators were continuous and multivariate normal, which is violated by ordinal indicators. A theoretically sound approach to handling ordinal indicators is the weighted least squares (WLS; Muthén, 1984), consisting of three stages. WLS handles the non-normality of categorical variables by calculating polychoric correlations and thresholds using ML in the first two stages (Muthén, 1984). This assumes that a normal and continuous distribution underlies the indicators. The polychoric correlations and thresholds are used in minimizing the fit function to obtain parameter estimates in the third stage,

$$F_{WLS} = (\boldsymbol{s}_{WLS} - \boldsymbol{\sigma}(\theta))^T \, W(\boldsymbol{s}_{WLS} - \boldsymbol{\sigma}(\theta)), \quad (1)$$

where W is the weight matrix of $\boldsymbol{s}_{WLS}$, $\boldsymbol{s}_{WLS}$ is a vector containing non-duplicated elements of estimated sample statistics that include the threshold and polychoric correlation values, and $\boldsymbol{\sigma}(\theta)$ is a vector of the model implied correlations. The weight matrix contains information about the kurtosis and covariance to correct for the non-normal distribution.

Table 1 Approaches evaluated in the study

Approach	Description	Software/Packages
FIML	A recently proposed approach to FIML for JOC models by Pritikin, Brick, & Neale (Pritikin et al., 2018)	OpenMx
EMB-WLSMV	Imputation by EMB (Honaker et al., 2011), followed by WLSMV (Muthén & Muthén, 2017) in the analysis phase	Amelia and lavaan with the WLSMV estimator
FCS-WLSMV	Imputation by FCS (van Buuren, 2007), followed by WLSMV (Muthén & Muthén, 2017) in the analysis phase	mice and lavaan with the WLSMV estimator
FCSLV-WLSMV	Imputation by FCSLV (Enders et al., 2018), followed by WLSMV in the analysis phase	Blimp and lavaan with the WLSMV estimator
EMB-ML	Imputation by EMB (Honaker et al., 2011), followed by FIML in the analysis phase	Amelia and OpenMx
FCS-ML	Imputation by FCS (van Buuren, 2007), followed by FIML in the analysis phase	mice and OpenMx
FCSLV-ML	Imputation by FCSLV (Enders et al., 2018), followed by FIML in the analysis phase	Blimp and OpenMx
PD-WLSMV	Pairwise deletion followed by the WLSMV estimator in the analysis phase	lavaan with the WLSMV estimator

Diagonal weighted least squares (DWLS) is a special case of WLS when only the diagonal elements of W in WLS are used in minimizing the fit function. Corrections for standard errors and the chi-square statistic are applied to compensate for the potential misestimation when the full weight matrix is not computed (Asparouhov & Muthén, 2010). DWLS with the abovementioned corrections is referred to as weighted least squares with mean and variance adjustment (WLSMV). Simulation studies found that WLSMV produced unbiased estimates for ordinal indicators even when responses were asymmetrically distributed and had few categories (Bandalos, 2008; DiStefano et al., 2019). Beauducel and Herzberg (2006) found that WLSMV clearly outperformed ML estimation (assuming categorical indicators were continuous) when the categorical indicators had two or three categories and did not require a larger sample size. Parameter estimates and chi-square statistics were underestimated and overestimated respectively by ML, while WLSMV produced less biased parameter estimates and chi-square statistics. When the underlying distribution of the ordinal indicators was normally distributed, the least-squares-based estimators outperformed even robust ML (treating ordinal indicators as continuous) in generating unbiased factor loadings in CFA (Rhemtulla et al., 2012). But the performance of the least-squares-based estimators was not perfect, as Li (2016) noted that WLSMV produced biased estimates in small samples when the underlying distribution of the categorical indicators was non-normal.

Full information maximum likelihood (FIML)

For continuous and multivariate normal data, FIML is an extension of the normal theory ML estimator, which accommodates missing data as a by-product of using raw data as inputs. In ML, the likelihood function is usually computed assuming all observations have complete responses and observations with missing values are discarded. FIML accommodates missing responses by using the non-missing information to calculate the likelihood instead of discarding missing cases (Enders, 2010). FIML uses all observed data as input, while ML uses the covariance matrix of the dataset as input to minimize the fit function. FIML uses the raw data to calculate case-wise log-likelihoods (multiplied by −2 for convenience),

$$-2LL_i = k_i log(2\pi) + log\left|\sum(\theta)_i\right| + \left(\boldsymbol{Y_i}-\boldsymbol{\mu}(\theta)_i\right)^T \sum(\theta)_i^{-1} \left(\boldsymbol{Y_i}-\boldsymbol{\mu}(\theta)_i\right), \quad (2)$$

where LL represents the log-likelihood, k_i is the number of variables with non-missing values for the ith case, and $\boldsymbol{Y_i}$ is a vector containing the non-missing scores for the ith case. The subscript i implies that each case can have different patterns of missingness, and the relevant vectors will be adjusted to only contain non-missing estimates. $\sum(\theta)_i$ and $\boldsymbol{\mu}(\theta)_i$ are the model-implied covariance matrix and the model-implied vector of the means of the joint distribution of the non-missing variables for the ith case, respectively. The $-2LL$ of each case is summed up to form the $-2LL$ of the entire dataset,

$$-2LL_{Data} = \sum_{i=1}^{n} -2LL_i. \quad (3)$$

Maximizing the log-likelihood of the data will produce parameter estimates that describe the model being tested. FIML is available in most programs such as Mplus, Lisrel, lavaan, OpenMx, EQS, and Amos and is widely used to handle missing data.

Additional modification is required to use FIML with JOC models or models with categorical indicators. Applying continuous FIML to ordinal SEM models with missing data produced

biased parameter estimates and standard errors (Teman, 2012). Probit and logit FIML can accommodate missing ordinal data (e.g., Asparouhov & Muthén, 2016) but are more commonly utilized in the domain of Item Response Theory (IRT). Wirth and Edwards (2007) provide an informative overview of IRT's relationship with SEM and the limitations of probit/logit FIML. In testing measurement invariance, Chen et al. (2019) found that combining FIML with probit or logit links produced relatively unbiased factor loadings, standard errors, and type I error rates during model comparison. One practical limitation is that probit and logit FIML approaches become computationally intensive when the number of latent variables (dimensions) increases due to the numerical integration involved. Pritikin et al. (2018) proposed a novel approach to FIML in JOC models, implemented in the OpenMx package (Neale et al., 2016). Based on Lee et al. (1990), the multivariate probit distribution is used to model ordinal variables. Standard errors and the likelihood ratio tests are available and covariances between ordinal and continuous variables can be freed. Unlike probit and logit FIML, performance is not limited by the number of latent variables but by the number of ordinal variables in the model. Rather than estimating the joint likelihood directly, the authors utilized the axioms of conditional probability to break down the joint likelihood as

$$\begin{aligned} &P(Ordinal \cap Continuous) \\ &\quad = P(Ordinal|Continuous)\ P(Continuous) \\ &\quad = P(Continuous|Ordinal)\ P(Ordinal). \end{aligned} \tag{4}$$

The conditional likelihood and marginal likelihood are first calculated, and their product forms the joint likelihood. The choice of conditioning on the continuous or ordinal indicators is determined by an algorithm to maximize processing speed. Further technical details of its implementation and the mathematical underpinnings of this approach can be found in Pritkin et al. (Pritikin et al., 2018). The authors found that this method produced relatively accurate estimates in small and moderate sample sizes of 250 and 500, respectively, than the estimates produced from applying multiple imputation and WLS in their studied conditions. This held true when the method was tested with both a small and a large factor model containing five and nine indicators, respectively. However, the authors were limited in their evaluation as they did not systematically vary missingness proportion, the number of categories, and the threshold of the ordinal variables. The effect of these conditions on the performance of FIML will be evaluated in the current study.

Multiple imputation (MI)

MI is split into three phases. Firstly, missing data is imputed multiple times to produce multiple sets of data in the imputation phase. Each set of data is analysed, and the sets of results are pooled using the formulas developed by Rubin (1987) in the pooling phase. MI methods differ mainly in their imputation algorithm; we introduce three imputation algorithms in this study.

Expectation-maximization with bootstrapping (EMB) Missing cases are imputed with values drawn from a distribution estimated by the EMB algorithm (Honaker et al., 2011). Firstly, multiple bootstrapped samples are generated from incomplete data. In the Expectation-Maximization (EM) algorithm, each bootstrapped sample goes through the Expectation (E) stage, calculating the expected likelihood of the model using assumed model parameters. In the Maximization (M) stage, updated model parameters are estimated such that they maximize the expected likelihood generated in the E stage. One cycle consists of an E and M stage. The updated model parameters are utilized in the E stage of the next cycle to estimate the likelihood. Cycles are repeated until convergence, where updated model parameters do not differ significantly from the previous cycle. Missing cases are imputed by values drawn from the distribution described by the updated model parameters. Each bootstrapped sample is imputed and used in the analysis phase.

The above is a simplified explanation of the EM algorithm, as illustrated in Honaker et al. (2011). The EMB algorithm runs much faster compared to FCS as convergence is more easily assessed and each bootstrapped sample is independent of each other. However, EMB assumes that the data form a multivariate-normal distribution, and imputed values are continuous. Amelia II implements the EMB algorithm and imputes ordinal variables by transforming imputed continuous values. For the imputation of an ordinal variable with k categories, the continuous value is divided by k, and this resultant value is fixed within the range between zero and 1. The resultant value is used as the probability to describe a binomial distribution with k trials. An ordinal value, which is the count of successful draws from the binomial distribution, is drawn and imputed. MI with EMB is currently implemented in R from the Amelia II package (Honaker et al., 2011).

Wu et al. (2015) found that EMB in Amelia II produced problematic estimates when the sample was small, and the ordinal responses were asymmetrically distributed in regression analysis. Similarly, Jia and Wu (2019) found that treating ordinal indicators as continuous variables in a CFA and imputing them with EMB produced high non-convergence rates when the responses were binary, asymmetrically distributed, and the sample was small. However, acceptable estimates were obtained in other conditions with no convergence issues. The above studies were based on models where indicators were purely ordinal. EMB is studied in the current paper to investigate whether its performance would be acceptable when a JOC model was analysed.

Fully conditional specification (FCS) FCS is another approach to multiple imputation which imputes missing data on a

variable-by-variable basis. FCS excels at ease and flexibility in creating constraints such as boundaries, skip patterns, interactions, and bracketed responses (van Buuren, 2018). In exchange, FCS is computationally slower compared to EMB, as missing values on each variable are imputed separately. FCS's most unique feature is the ability to specify imputation models for each variable. Specifying an appropriate imputation model generates imputations that are more likely to be drawn from the same population as the observed data (van Buuren & Groothuis-Oudshoorn, 2011; Wu, Jia, & Enders, 2015).

FCS in mice In mice, FCS is implemented using the multiple imputation by chained equations (MICE) algorithm, a Markov chain Monte Carlo (MCMC) method developed by van Buuren (2007). Missing values are initially imputed by randomly selected observed values. In the first step, model parameter estimates ($\widehat{\theta}$) that characterize the missing data are drawn from a posterior distribution derived from other variables, non-missing data of that variable, and the missingness. In the second step, imputations are drawn from another posterior distribution, derived from the same information and the parameter estimates from the first step. Missing values on the variable are imputed in the second step. The two steps above are iterated k times, ending at the kth variable. One cycle is complete when all variables have gone through these two steps. Cycles are repeated until the imputed values are stable and convergence is reached. The final imputed dataset is saved as one imputation to be used as input in the analysis phase. More details can be found in van Buuren (2018). For continuous variables with missing values, the default imputation model is predictive mean matching (van Buuren, 2007). For ordinal variables, the default imputation model is the proportional odds model (McCullagh, 1980). When an ordinal variable possesses only two categories, it is imputed with logistic regression. Imputed values generated for missing ordinal or binary values will be ordinal or binary, respectively. MICE is available in R from the mice package (van Buuren & Groothuis-Oudshoorn, 2011).

Jia and Wu (2019) applied MICE with the proportional odds model and the WLSMV estimator to analyse a three-factor ordinal model. FCS-WLSMV encountered high non-convergence in MAR[1] (with missingness occurring more frequently on the tail end of the distribution), small sample sizes, moderate missingness, and asymmetric distribution of responses. In other conditions, FCS-WLSMV produced acceptable parameter estimates.

FCS in Blimp (FCSLV) Blimp was primarily developed for the imputation of incomplete ordinal and nominal variables in multilevel modelling but is also usable for non-multilevel analysis. In Blimp, FCS was modified and extended to impute ordinal variables using a latent variable formulation in multilevel modelling. Conceptually, Blimp uses the same FCS algorithm from mice and adopts a cumulative probit model as the imputation model for incomplete ordinal variables, which posits that a latent variable underlies each incomplete ordinal variable. Continuous latent scores are modelled and discrete imputations for ordinal variables are created by applying estimated thresholds from the probit model. Detailed information about multilevel imputation by FCS can be found in van Buuren (2018) and Enders et al. (2018). More information about Blimp and the latent variable formulation can be found in Enders et al. (2018). Blimp was included in this study, as an anonymous reviewer highlighted the compatibility of the latent variable formulation in Blimp with the formulation of latent variables in CFA.

Comparing FIML and MI

FIML and multiple imputation have been compared in a variety of analyses. In SEM models with only ordinal indicators, Jia and Wu (2019) found that the combination of FIML with robust corrections by Yuan and Bentler (2000) performed better compared to a variety of multiple imputation methods in convergence and the generation of unbiased parameter estimates. These authors also found that it was least sensitive to different factors of the dataset compared to the multiple imputation methods. Pritikin et al. (2018) found that FIML performed better than combining FCS and WLS for estimating a simple polychoric correlation between two ordinal variables that are MAR. In latent growth models with MCAR time-invariant covariates, Cheung (2007) found that FIML performed better compared to multiple imputation, as multiple imputation underestimated standard errors and produced inaccurate model fit statistics when missingness was large. Overall, the literature suggests that the performance of FIML was superior.

But there are also practical considerations when choosing between FIML and multiple imputation. One of these considerations is the ease of implementation of the desired analysis. FIML integrates the handling of missing data and the estimation of the model into a single step, which makes the process seamless. But the trade-off is that it is much harder to incorporate FIML into analysis or software that does not support it. The modular nature of multiple imputation makes it such that it is simple to integrate into the existing analysis. It is even possible and almost trivial to perform the different phases in different software by exporting imputed datasets. Also, FIML approaches can be computationally intensive due to the difficulty of numerical integration involved, which may prevent normal users from utilizing FIML with more sophisticated models.

[1] Missingness on an indicator was dependent on other indicators measuring the same factor.

An empirical illustration

We present a short empirical illustration demonstrating the four approaches (EMB-WLSMV, FCS-WLSMV, FCSLV-WLSMV, and FIML) as applied to a simple JOC model using data from the National Health and Nutrition Survey (NHANES). Due to space constraints, the full empirical illustration was written in the supplemental materials. Table 2 displays the factor loading estimates and standard errors of the four approaches. It is uncertain which of the four approaches produced trustworthy results in this illustration, as the true factor structure is unknown. Thus, a simulation was conducted to assess the performance of each approach in different datasets.

A simulation study

Different factors were manipulated across datasets to investigate their effects on the performances of each approach. The results will be used to create recommendations for applied researchers.

Method

The simulation was conducted in R 3.5.0 (R Development Core Team, 2018) on the high-performance computing cluster. The R codes used can be found at https://github.com/Aaron0696/FIML_MI_JOC_MISSINGDATA.

Data generation

The population model used to generate the data was a one-factor CFA model with six indicators. Three of the indicators were ordinal indicators (o1, o2, and o3), and three were continuous indicators (c1, c2, and c3). Continuous and multivariate-normal data for all six indicators were first generated from the lavaan package (Rosseel, 2012). The factor loadings for o1, o2, o3, c1, c2, and c3 were 0.3, 0.5, 0.7, 0.7, 0.8, and 0.85, respectively, following the general pattern of the model tested by Pritikin et al. (2018). Error variances of the ordinal indicators and factor variance were fixed to 1 (Kline, 2016). Error variances of the continuous indicators were fixed to the difference between 1 and the square of their respective factor loadings. Ordinal indicators were created from the continuous data by categorizing the scores according to thresholds. MCAR missing data was created by randomly deleting data from each variable according to the desired missingness proportion. Figure 1 displays the diagram for the population model with the relevant parameters.

Manipulated factors in the simulation

Number of categories Ordinal indicators with two, three, five, and seven categories were studied in the simulations. All ordinal variables had the same number of categories in the same condition. These values were adapted from Wu et al. (2015), Chen et al. (2019), and Jia and Wu (2019).

Distribution of responses across categories The distribution of responses across categories was manipulated by using different thresholds adapted from previous simulation studies (Jia & Wu, 2019; Rhemtulla et al., 2012; Wu et al., 2015). Thresholds were classified as symmetric, moderately asymmetric, and severely asymmetric. The thresholds corresponding to the respective number of categories and the degree of asymmetry are shown in Table 3.

Missingness proportion Existing studies have investigated widely different missingness proportions. Wu et al. (2015) and Chen et al. (2019) used 30% and 50%, Jia and Wu (2019) used 15% and 30%, Shi et al., (2019, 2020) used 15%, 25% and 50%, and Enders (2001) used 0%, 5%, 10%, 15%, 20% and 25%. Based on the values from previous studies, the current paper uses 10%, 20%, and 40% missingness on each variable to represent small, moderate, and large missingness proportions. MCAR missing data was created by randomly deleting data from each variable according to the missingness proportion.

Sample size Existing simulation studies (Chen et al., 2019; Shi et al., 2019, 2020) have used sample sizes of 200 and 1000 to

Table 2 Factor loading estimates (Est) and standard errors (SE) from the empirical illustration

Indicator	EMB-WLSMV		FCS-WLSMV		FCSLV-WLSMV		FIML	
	Est	SE	Est	SE	Est	SE	Est	SE
CGrip	6.68	0.51	7.61	0.47	8.14	0.56	7.92	0.60
LiftD	−1.34	0.12	−2.11	0.20	−2.23	0.26	−2.20	0.24
GraspD	−0.58	0.04	−0.84	0.05	−0.87	0.06	−0.87	0.05
PshPlD	−1.31	0.11	−1.90	0.15	−1.91	0.17	−1.95	0.18

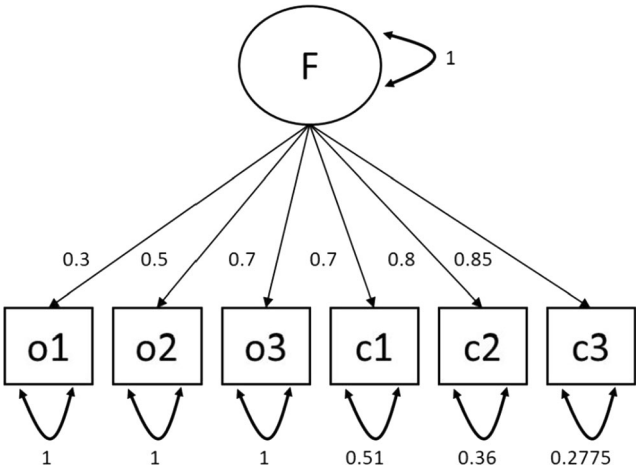

Fig. 1 Simulation population model. *Note.* Factor loadings and error variances are displayed in the diagram

represent small and large sample sizes in structural equation models. Following these studies, small and large sample sizes corresponded to 200 and 1000, respectively. Crossing the abovementioned factors created 72 (4×3×3×2) different conditions to be analysed by seven different missing data approaches. Figure 2 visualizes the distribution of responses across different conditions with sample size and missingness proportion constrained at 1000 and 20%, respectively.

Other specifications

For all approaches, the mean and variance of the latent factor was fixed to zero and 1, respectively, and error variances of the ordinal variables were fixed to 1. For the approaches involving MI, 50 imputations were generated for every imputation phase. Since both MI and FIML were used, the simulation studies were computationally very intense. We used 800 replications in each condition to balance accuracy and computational time.

Performance measures

The following measures were considered in evaluating the performance of the different approaches: Proportion of non-convergence, relative bias of parameter estimates, and relative bias of standard errors.

Proportion of non-convergence Following Jia and Wu (2019), replications that produced improper solutions or did not converge were counted as non-convergence replications. Improper solutions included the following cases: Standard errors or factor loadings that were ten standard deviations above or below the mean standard error or factor loading for that condition, standard errors greater than 10, or negative standard errors.

For FIML, a replication is considered converged if the status code in OpenMx reflects that the estimator converged without issue. In some cases, certain standard errors may not be estimated even though convergence was attained. This simulation treats these replications as non-converged and improper solutions, and they are excluded in further calculations. For the approaches that utilized multiple imputation, the replication is considered non-convergent if the pooled estimate is an improper solution or less than half of the imputations converged. Non-converged imputations were excluded from the pooling phase and from all subsequent calculations.

Relative bias of factor loadings (FL bias) For each factor loading in the model, the relative bias was calculated as the proportion of the raw bias (numerator) to the true parameter value (denominator) from the population model, $FL\ bias = \frac{(\overline{\lambda}-\lambda_{Pop})}{\lambda_{Pop}} * 100\%$, where $\overline{\lambda}$ denotes the average factor loading for a condition with 800 replications and λ_{Pop} denotes the factor loading of the population model. Hoogland and Boomsma (1998) considered a relative bias of

Table 3 Thresholds for generating ordinal variables

Degree of asymmetry	Number of categories	Thresholds in Z-scores					
Symmetric	2	0.00					
	3	-0.83	0.83				
	5	−1.50	-0.50	0.50	1.50		
	7	−1.79	−1.07	-0.36	0.36	1.07	1.79
Moderate asymmetry	2	0.36					
	3	-0.50	0.76				
	5	-0.70	0.39	1.16	2.05		
	7	−1.43	-0.43	0.38	0.94	1.44	2.54
Severe asymmetry	2	1.04					
	3	0.58	1.13				
	5	0.05	0.44	0.84	1.34		
	7	-0.25	0.13	0.47	0.81	1.18	1.64

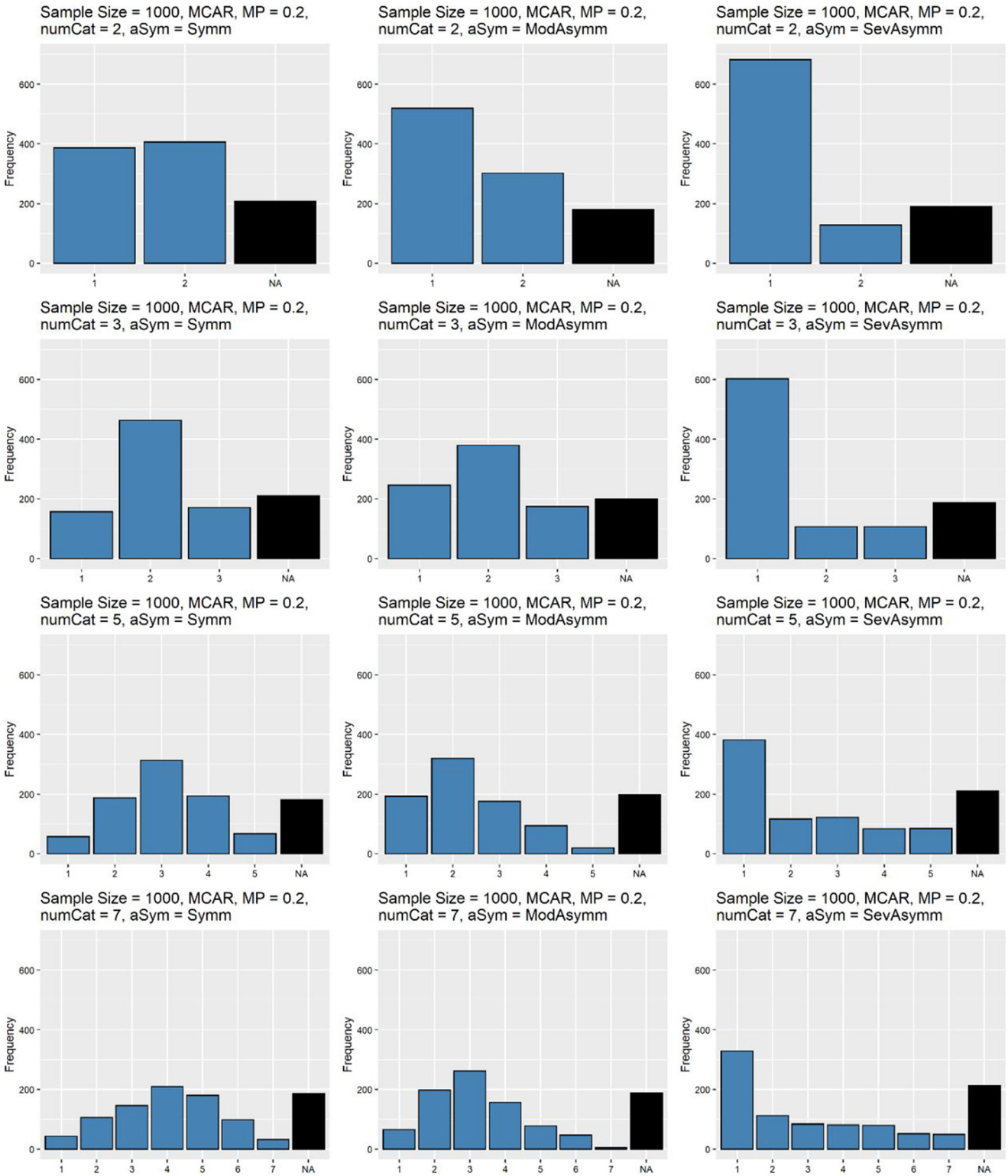

Fig. 2 Distribution of o1. *Note.* The distribution of responses is visualized across the different conditions after data is deleted when the sample size is 1000. The height of the bars (y-axis) reflects the frequency of each response category (x-axis). Black bars represent the amount of missing data. The title above each graph displays the sample size, missing mechanism, missingness proportion (MP), number of categories (numCat), and degree of asymmetry (aSym)

±5% as acceptable, but Muthén et al. (1987) posit that bias within ±10% was acceptable. The current paper considers ±10% as an acceptable range for the relative bias of factor loadings.

Relative bias of standard errors (SE bias) The relative bias in the standard error is the proportion of the raw bias (numerator) to the true empirical standard error (ESE; denominator) across 800 replications[2], $SE\ bias = \frac{(\overline{SE}-ESE)}{ESE} * 100\%$, where $\overline{SE}$ is the average standard error across all 800 replications of a condition, and *ESE* is the standard deviation of the parameter estimate across 800 replications. For SE bias, a relative bias within ±10% and below is considered acceptable (Hoogland & Boomsma, 1998).

Root-mean-squared-error (RMSE) The mean-squared-error (MSE) is a measure of overall accuracy, summarizing the bias and variability in estimation (Wu et al., 2015). For ease of interpretation, the current study takes the square root of the MSE such that it is on the same scale as the factor loadings. $RMSE = \sqrt{\sum_{k=1}^{K}(FL_k - FL_{True})^2}/K$, where K = 800 is the number of replications for each condition. Smaller values of RMSE closer to zero are indicative of less error and greater accuracy.

Results and discussion

The results and discussion section will answer each research question by referencing results from the simulation. Rounded values of the performance measures are included in the supplementary materials and displayed as heatmaps (Figures S2 to S8). Table 4 summarizes the trends in FL bias, SE bias and RMSE within each approach.

> **Question 1:** What are the effects of sample size, missingness, number of categories, and distribution of response categories on the convergence rates, factor loadings, standard errors and RMSE?

Non-convergence rates

Overall, approaches across all conditions had non-convergence rates of less than 10% (80 out of 800 replications) except for the FCSLV approaches. FCSLV-WLSMV had large non-convergence rates between 37% and 63% when the sample size was small, missingness was large, the number of categories was seven, and the distribution was moderately asymmetric. This specific combination of data factors produced the smallest possible observed frequencies, where a certain response category of the ordinal variables has very small raw counts. FIML also had increased non-convergence when small observed frequencies were present, but non-convergence was below the 10% threshold. FCSLV-ML had non-convergence rates between 20% and 35% when the sample was small, missingness was high, and the number of categories was two. All other methods performed adequately well when the convergence rate was assessed and had no major problems with convergence.

FL and SE bias

All approaches produced biased estimates in some conditions except for FIML, which produced unbiased factor loading and standard error estimates for all tested conditions. We observed that higher rates of missingness appear to be the most important data factor affecting the performance, as it was associated with greater bias in almost all approaches for all estimated parameters. Smaller sample size, lower number of categories, and more asymmetric distributions were also associated with greater bias across all approaches and parameters, but were less prevalent compared to higher missingness.

RMSE

All approaches produced similar trends in RMSE for ordinal factor loadings. Within small samples, RMSE was greater when missingness was higher, number of categories was lower, and distributions were more asymmetric. RMSE within large samples were lower and were less sensitive to the degree of missingness, number of categories, and distribution. RMSE for continuous factor loadings produced two distinct trends for the WLSMV and FIML-based approaches. RMSE for continuous factor loadings in the WLSMV approaches had a similar trend as the RMSE for ordinal factor loadings. RMSE for continuous factor loadings in the FIML approaches were higher when the sample size was small, but did not interact with the degree of missingness, number of categories, and distribution.

> **Question 2:** Amongst the approaches studied, which performs the best?

Overall, FIML would be the most preferred approach as it was able to produce unbiased estimates of factor loadings and standard errors in all tested conditions. RMSE values were less than 0.1 for continuous factor loadings in all conditions. RMSE values for ordinal factor loadings were also less than

[2] Complete-data SEM with ML utilizes a *z*-test for hypothesis testing. When multiple imputation is involved, the estimated SE follows a *t*-distribution with degrees of freedom that is dependent on the missing information. The degrees of freedom involved in the current study are large and exceed 50. Thus, this measure of relative bias remains appropriate as the *t*-distribution approximates a *z*-distribution with large degrees of freedom.

Table 4 Summary of FL bias, SE bias, and RMSE across approaches

			Missing data methods			
			EMB	FCS	FCSLV	PD
CFA estimation methods	WLSMV	Ordinal factor loadings	Biased in higher missingness, lower number of categories and more asymmetric distributions, regardless of sample size	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
		Standard errors of ordinal factor loadings	Biased in higher missingness, lower number of categories, and more asymmetric distributions, regardless of sample size	Biased in moderate or high missingness, especially if sample size is small	Biased in moderate or high missingness, especially if sample size is small	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
		RMSE of ordinal factor loadings	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
		Continuous factor loadings	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
		Standard errors of continuous factor loadings	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Biased in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
		RMSE of continuous factor loadings	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions
	FIML*	Ordinal factor loadings	Biased in higher missingness, lower number of categories and more asymmetric distributions, regardless of sample size	Unbiased	Unbiased	Not applicable
		Standard errors of ordinal factor loadings	Biased when missingness is moderate or high, regardless of other factors	Biased when missingness is moderate or high, regardless of other factors, or when small observed frequencies are present	Biased when missingness is moderate or high, regardless of other factors, or when small observed frequencies are present	Not applicable

Table 4 (continued)

	Missing data methods			
	EMB	FCS	FCSLV	PD
RMSE of ordinal factor loadings	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Greater in smaller samples, higher missingness, lower number of categories and more asymmetric distributions	Not applicable
Continuous factor loadings	Unbiased	Unbiased	Unbiased	Not applicable
Standard errors of continuous factor loadings	Biased when missingness is moderate or high, regardless of other factors	Biased when missingness is moderate or high, regardless of other factors, or when small observed frequencies are present	Biased when missingness is moderate or high, regardless of other factors, or when small observed frequencies are present	Not applicable
RMSE of continuous factor loadings	Greater in smaller samples and higher missingness	Greater in smaller samples and higher missingness	Greater in smaller samples and higher missingness	Not applicable

*When either EMB or FCS was used to handle missing data, there will be no missing data during estimation. Utilizing FIML to address missing data and conduct CFA estimation produced unbiased factor loadings and standard error in all conditions, RMSE trends were identical to the other FIML approaches

0.1 when the sample size was large, but ranged from 0.12 to 0.26 when sample size was small. Despite the lower accuracy in small samples when estimating ordinal factor loadings, FIML was one of the most accurate amongst all tested approaches and conditions for both ordinal and continuous factor loadings. FIML was prone to moderate rates of non-convergence when small observed frequencies were present, but its accuracy in estimating parameters relative to other approaches and across most conditions outweighs this disadvantage. If FIML is not feasible or does not converge, the second-best approach would be the combination of FCS and WLSMV, as it was able to produce unbiased and accurate estimates consistently when the sample size was large. The next section will answer the third research question by comparing the remaining missing data methods when the same CFA estimator was used to ascertain the effect of each missing data method on the bias produced.

> **Question 3:** Are differences in performance driven by differences in missing data approaches or CFA estimators?

Answering the above question would require us to compare across missing data methods when the same estimator is used and vice versa. We first compare missing data methods while holding the CFA estimator constant. Differences in performance between the approaches can be attributed to differences between the missing data methods.

EMB-WLSMV, FCS-WLSMV, FCSLV-WLSMV, and PD-WLSMV

Performance was similar when estimating the factor loadings of the continuous variables and their corresponding standard errors. Differences were observed when estimating ordinal factor loadings and their corresponding standard errors. When estimating ordinal factor loadings with a small sample, all four approaches were similarly vulnerable to higher missingness, lower number of categories, and more asymmetric distributions. For FCS, FCSLV, and PD, these data factors had little to no effect when the sample size was large. EMB was marginally worse as biases were still observed in large samples. When estimating the standard errors of the factor loadings of ordinal variables, the four approaches all performed differently. EMB produced biased standard errors in both small and large samples when the missingness was high, the number of categories was low, and the distribution was more asymmetric. FCS and FCSLV produced biased standard errors when missingness was moderate or high, and the bias was persistent across all numbers of categories and sample sizes, which worsens when the distribution was asymmetric. PD produced biased standard errors in small samples when the missingness was high, the number of categories was low, and the distribution was more asymmetric. The RMSE trends

within the four WLSMV approaches were similar, with greater RMSE in conditions with smaller samples, higher missingness, lower number of categories and more asymmetric distributions. Overall, FCSLV, FCS and PD performed best, as they were able to consistently provide unbiased factor loadings and standard errors when the sample size was large. However, FCSLV's high rate of non-convergence is concerning and the FCS imputation method would be preferred when paired with the WLSMV estimator. While PD performed well in this study, and even outperformed some of the missing data approaches on certain performance measures, it is not recommended that researchers utilize PD over other missing data approaches. This is because PD may produce biased estimates when missingness is not MCAR, which is likely in applied research. A multiple imputation approach would produce less biased estimates in non-MCAR conditions, as it utilizes information from other variables. van Ginkel et al. (2020) provides a nice summary about the other advantages of multiple imputation over PD.

EMB-ML, FCS-ML, and FCSLV-ML

When the FIML estimator is used, all three missing data approaches produced unbiased estimates of continuous factor loadings in all conditions. FCS and FCSLV produced unbiased estimates of ordinal factor loadings in all conditions, while EMB produced bias estimates when missingness was high, the number of categories was low, and the distribution was asymmetric. In estimating ordinal factor loadings, the FCS and FCSLV imputation methods appear superior. When estimating standard errors of both continuous and ordinal factor loadings, all three imputation methods produced bias estimates when missingness was moderate or high, regardless of other data factors. Performance and trends on RMSE were similar as well, as the three conditions produced higher RMSE in smaller samples regardless of other factors. EMB has a slight advantage over the other two imputation methods as the imputed datasets were not as susceptible to small observed frequencies like the FCS and FCSLV methods.

FCS-WLSMV, FCS-ML, and FIML

FIML and FCS-WLSMV were the two best approaches from the simulation study. FIML produced unbiased estimates even in small samples, whereas the FCS-WLSMV approach produced biased estimates when missingness was high, the number of categories was low, and distribution was asymmetric. Both FIML and FCS-WLSMV produced greater RMSE in smaller samples, but FCS-WLSMV became less accurate when missingness increased. It is difficult to ascertain whether these differences in performance were due to the missing data treatment or choice of CFA estimator, as FIML and FCS-WLSMV differed in both. The FCS-ML approach is used to ascertain the effect of the missing data approach by comparing the performance between FCS-ML and FIML. We also compare FCS-ML and FCS-WLSMV to ascertain the effect of the estimator.

FCS-ML and FIML The main performance difference between these two approaches lies in the estimation of standard errors. While FIML produced unbiased standard errors in all conditions, FCS-ML produced biased standard errors when missingness was moderate or high. The FCS imputation method appears to be responsible for the bias in standard errors when missingness was moderate or high, which is when more data is imputed, and the impact of imputation is larger. Thus, the bias in standard errors can be attributed to the missing data treatment.

FCS-ML and FCS-WLSMV FCS-ML was able to produce unbiased estimates of factor loadings in all tested conditions, while WLSMV produced bias estimates in smaller samples. In smaller samples, RMSE for continuous factor loadings were greater when missingness was higher in WLSMV, while missingness and RMSE were independent in FIML. The ability to produce unbiased factor loadings and accurate estimates in small samples can be attributed to the choice of estimator. Both approaches also produced biased standard errors in both large and small samples, especially when missingness was high. But they differed in the trend and magnitude of the bias. FCS-WLSMV produced biased standard errors in fewer conditions when the sample size was large compared to FCS-ML, which produced biased standard errors in both large and small samples. FCS-ML also produced completely untrustworthy standard estimates when the data had small observed frequencies. While this and the former section have ascertained that the FCS missing data method is responsible for the bias in standard errors, it appears that the trend of the bias interacts with the choice of CFA estimator.

Recommendations

Based on the results discussed above, FIML is recommended as the approach to handling missing data in JOC models if there is no issue of non-convergence or if the sample is small. FIML is the preferred choice due to its ability to generate acceptable estimates in almost all conditions in the simulation study. However, caution should be applied when ordinal categories have low raw frequencies, which could lead to non-convergence. FIML for JOC models also becomes computationally intensive as the number of ordinal indicators increase, which makes it unfeasible for large models.

If the sample size is large, FCS-WLSMV is recommended as the alternative if non-convergence persists, if there are low raw frequencies in the ordinal variables or computation takes

too much time. FCS-WLSMV performed best in large samples, producing relatively unbiased estimates of factor loadings and standard errors when the sample was large. If FIML is not feasible and the sample size is small, researchers should ensure that the dataset has low missing proportions and should have five or more categories in their ordinal variables with symmetric distributions before applying the WLSMV-based approaches. If such conditions are not available, applied researchers could consider collapsing categories with small observed frequencies while using FIML.

In light of the simulation findings, it would appear that the estimates from FIML in the empirical illustration would be the most trustworthy; it was unbiased in all conditions in the simulation. The estimates from EMB-WLSMV differed the most from the other methods; it appeared to consistently underestimate the factor loadings and agreed with the results from the simulation study. FCS-WLSMV, FCSLV-WLSMV, and FIML did not differ much, possibly because the data from the empirical illustration were rather well-behaved. It had a large sample size and the ordinal variables had four categories.

Limitations and future directions

The present simulation study only investigated a small subset of possible models and datasets and would likely not generalize exactly to other simulated or actual datasets. Thus, caution is advised in over-interpreting specific effects, and researchers should focus more on the general trends.

Applied researchers with datasets that exhibit more extreme factors than the factors in this study should apply the recommendations cautiously. It is possible that there are other properties of the dataset that may present a challenge to the tested approaches. One such property was the small observed frequencies, which may have caused greater rates of non-convergence and inaccurate model rejection rates. Future studies should also investigate other factors that may compromise the performances of the approaches. Some recommendations include other mechanisms of missingness, patterns of missingness, observable frequencies, and degree of non-normality of the distribution that underlies the ordinal variables. The involvement of both ordinal and continuous variables also brings with it the possibility of varying their relationship as design factors in future simulation studies. One suggestion would be to investigate conditions where the missingness of an ordinal and continuous variable was yoked to each other.

The current study only studied JOC models in the context of CFA, using a simplistic one-factor model. Future studies should expand the models to investigate the performance of the various techniques in estimating models with structural paths (SEM) or within exploratory factor analysis where these models are different from CFA. Applying JOC models to SEM is straightforward, as structural paths can be easily added in OpenMx; extensions to EFA are also possible if methods for performing factor rotations are available. Furthermore, the model studied used an equal number of ordinal and continuous indicators. Future studies could investigate whether the proportion of continuous and ordinal indicators has any effect on the performance of each approach. It is plausible that the joint distribution of the observed variables would more closely approximate a multivariate normal distribution if there were more continuous variables in relation to ordinal variables, where the maximum likelihood estimators may perform better. The current study is also limited in studying the different imputation models within multiple imputation. The mice package alone contains more than 15 imputation methods, but the current study only investigated the more popular defaults of predictive mean matching and proportional odds models. Some imputation models (e.g. mice.impute.norm) are more compatible with the data-generating model used in the simulation; future studies can evaluate whether a match between the imputation and data-generating model results in better performance.

The current paper is not exhaustive in testing all methods of handling JOC models with missing data, such as Bayesian methods, which are freely available in programs such as Stan (https://mc-stan.org/), and which were mentioned by an anonymous reviewer to be capable of model estimation with small samples. Future studies may compare the pros and cons of the frequentist versus Bayesian approaches.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01582-w.

References

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

Asparouhov, T., & Muthén, B. O. (2016). IRT in Mplus. Mplus Technical report. Retrieved from: http://www.statmodel.com/download/MplusIRT.pdf

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling, 15*, 211–240.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*(5), 515-524.

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The "Big Five Questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences*, *15*(3), 281–288.

Chen, P., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2019). Testing measurement invariance with ordinal missing data: A comparison of estimators and missing data techniques. *Multivariate Behavioral Research, 55*(1), 87–101. https://doi.org/10.1080/00273171.2019.1608799.

Cheung, M. W. L. (2007). Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random. *Organizational Research Methods, 10*(4), 609–634

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246.

DiStefano, C., McDaniel, H. L., Zhang, L., Shi, D., & Jiang, Z. (2019). Fitting large factor analysis models with ordinal data. *Educational and Psychological Measurement, 79*(3), 417–436. https://doi.org/10.1177/0013164418818242

Duncan, S. C., Duncan, T. E., & Strycker, L. A. (2001). Qualitative and quantitative shifts in adolescent problem behavior development: A cohort-sequential multivariate latent growth modeling approach. *Journal of Psychopathology and Behavioral Assessment, 23*(1), 43-50.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*, 352–370.

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

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*(2), 298.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software, 45*(7), 1-47. Retrieved from http://www.jstatsoft.org/v45/i07/

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

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

Kline, R. B. (2016). *Methodology in the social sciences. Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters, 9*(1), 91–97.

Leite, W. L., Svinicki, M., & Shi, Y. (2010). Attempted validation of the scores of the VARK: Learning styles inventory with multitrait–multimethod confirmatory factor analysis models. *Educational and Psychological Measurement, 70*(2), 323–339.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949.

Loggia, M. L., Juneau, M., & Bushnell, M. C. (2011). Autonomic responses to heat pain: Heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity. *PAIN®, 152*(3), 592–598.

MacDermid, J. C., Richards, R. S., Donner, A., Bellamy, N., & Roth, J. H. (2000). Responsiveness of the short form-36, disability of the arm, shoulder, and hand questionnaire, patient-rated wrist evaluation, and physical impairment measurements in evaluating recovery after a distal radius fracture. *The Journal of Hand Surgery, 25*(2), 330–340.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(2), 109–127.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132.

Muthén, L.K. & Muthén, B.O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

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

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K. T., & Trautwein, U. (2011). Who took the "×" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science, 22*(8), 1058–1066.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*(2), 535–549.

Pritikin, J. N., Brick, T. R., & Neale, M. C. (2018). Multivariate normal maximum likelihood with both ordinal and continuous variables and data missing at random. *Behavior Research Methods, 50*, 490–500.

R Development Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/.

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

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

Shi, D., Lee, T., Fairchild, A. J., & Maydeu-Olivares, A. (2019). Fitting ordinal factor analysis models with missing data: A comparison between pairwise deletion and multiple imputation. *Educational and Psychological Measurement* https://doi.org/10.1177/0013164419845039

Teman, E. D. (2012). *The performance of multiple imputation and full information maximum likelihood for missing ordinal data in structural equation models* (Doctoral dissertation, University of Northern Colorado, Greeley, Colorado). Available from ProQuest Dissertations Publishing (UMI No. 3555133).

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465–471.

van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall.

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 Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment, 102*(3), 297–308.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods, 49*(2), 653–673.

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

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate*

Behavioral Research, 50, 484–503. https://doi.org/10.1080/00273171.2015.1022644

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165–200.