Behavior Research Methods (2020) 52:2567–2587
https://doi.org/10.3758/s13428-020-01415-2

Addressing missing data in specification search in measurement invariance testing with Likert-type scale variables: A comparison of two approaches

Po-Yi Chen[1] · Wei Wu[2] · Holger Brandt[3] · Fan Jia[4]

Published online: 3 June 2020

Abstract
In measurement invariance testing, when a certain level of full invariance is not achieved, the sequential backward specification search method with the largest modification index (SBSS_LMFI) is often used to identify the source of non-invariance. SBSS_LMFI has been studied under complete data but not missing data. Focusing on Likert-type scale variables, this study examined two methods for dealing with missing data in SBSS_LMFI using Monte Carlo simulation: robust full information maximum likelihood estimator (rFIML) and mean and variance adjusted weighted least squared estimator coupled with pairwise deletion (WLSMV_PD). The result suggests that WLSMV_PD could result in not only over-rejections of invariance models but also reductions of power to identify non-invariant items. In contrast, rFIML provided good control of type I error rates, although it required a larger sample size to yield sufficient power to identify non-invariant items. Recommendations based on the result were provided.

Keywords Specification search · Partial invariance model · Ordinal missing data · Measurement invaraince · Modification index

Introduction

Measurement equivalent/invariance (ME/I) concerns whether the relationship between the targeted latent variable and the observed items are identical across groups (Millsap, 2012). ME/I is an important psychometric property, which ensures that the results obtained from a construct measure are comparable and generalizable across groups (Brown, 2014). Failure to achieve ME/I can lead to biased comparisons or selections across groups (Chen, 2008; Millsap & Kwok, 2004; Steinmetz, 2013).

Evaluating ME/I properties of a scale is a complex issue. There are various basic levels of ME/I, representing more or less restricted levels of invariance (explained in detail below). Thus, ME/I is typically examined through a series of Chi-square difference tests ($\Delta\chi^2$ tests) between nested invariance models (representing different levels of ME/I) using multiple group confirmatory analysis (MG-CFA) (Vandenberg, & Lance, 2000). If a certain basic level of ME/I is not achieved, indicating that one or more items are not invariant across groups, researchers can follow up with a specification search for the problematic items and release implausible constraints correspondingly. In the current study, we refer to this kind of specification search as sequential backward specification search (SBSS).

The appropriate methods for specification search vary depending on the nature of the item-level data. Considering the popularity of Likert-type scales in social and behavioral sciences, the items are often ordinal in nature. In addition, missing data are likely to occur due to nonresponse or planned missing data designs. To our knowledge, although specification search methods have been studied with ordinal complete data (Oort, 1998; Yoon & Kim, 2014), no research on those methods has considered missing ordinal data.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01415-2) contains supplementary material, which is available to authorized users.

✉ Po-Yi Chen
poyi.chen@utrgv.edu

[1] Department of Psychological Science, University of Texas Rio Grande Valley, Edinburg, TX, USA

[2] Department of Psychology, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

[3] Department of Psychology, University of Zürich, Zürich, Switzerland

[4] Department of Psychological Science, University of California Merced, Merced, CA, USA

In practice, three estimation methods have been used to conduct SBSS when ordinal missing data present. These methods include full information maximum likelihood (FIML), robust full information maximum likelihood (rFIML), and the mean and variance adjusted weighted least-squared method paired with pairwise deletion (WLSMV_PD) (e.g., Beam, Marcus, Turkheimer & Emery, 2018; Bou Malham & Saucier, 2014; Fokkema, Smits, Kelderman, & Cuijpers, 2013; Sommer, et al., 2019). Given that previous research has shown that rFIML outperformed FIML for ordinal data (e.g., Chen, Wu, Garnier-Villarreal, Kite & Jia, 2019), we only consider rFIML and WLSMV_PD in the current study.

In theory, rFIML and WLSMV_PD have their own advantages and limitations. rFIML is capable of utilizing all available information for model estimation, but it assumed ordinal data to be continuous. WLSMV_PD, on the other hand, correctly accounts for the ordinal nature of the data, but it uses pairwise deletion (PD) to handle missing data, which is less optimal than a full information method (Savalei & Bentler, 2005).

Given that neither of the methods is ideal theoretically, it would be interesting to know how they perform in SBSS and which one would be better. The goal of the current study is thus to examine the relative performance of the two methods in SBSS with ordinal missing data. The rest of the article is organized as follows. We first review necessary background information for the study, including a typical ME/I test process and specification search methods. We then explain how rFIML and WLSMV_PD work and present designs for the simulation study. Finally, we report the result from the simulation study and provide practical recommendations to empirical researchers based on our results. We conclude the article by discussing limitations of the current study and potential directions for future research.

A typical ME/I testing procedure with MG-CFA

There are four basic levels of ME/I, namely, configural, metric, scalar, and strict invariance, representing the least restricted invariance model to the most restricted invariance model, respectively. When the configural invariance model is examined, the same CFA model is fit to groups, with model parameters allowed to vary across groups. The metric invariance model is simply the configural invariance model with equality constraints on all factor loadings across groups. The scalar invariance model adds equality constraints on intercepts/thresholds across groups. Finally, equality constraints on corresponding residual variances are further included into the scalar invariance model to create a strict invariance model (Gregorich, 2006).

These four basic ME/I models are nested and thus are usually evaluated in sequence to decide which level of ME/I is achieved, starting from the configural invariance model. If the configural invariance fits the data, then a Chi-square difference ($\Delta\chi^2$) test comparing it to the metric invariance model will be conducted. A non-significant $\Delta\chi^2$ test would indicate that the metric invariance model passes. Researchers can further test scalar invariance by comparing it to the metric invariance model through another $\Delta\chi^2$ test, and so on (Byrne, 2013). On the other hand, a significant $\Delta\chi^2$ test would suggest that imposing equality constraints on all target parameters are implausible, and at least one target parameter is not equal (non-invariant) across groups.

Partial invariance models

The four types of ME/I models described above are considered "full" invariance models, given that equality constraints are placed on all parameters of the same type (e.g., all loadings) across groups (Jung & Yoon, 2016). When a full invariance model is rejected, one could establish a partial invariance model by releasing equality constraints on some but not all of the target parameters (Byrne, Shavelson, & Muthén, 1989; Millsap, 2012; Putnick, & Bornstein, 2016; Schmitt & Kuljanin, 2008). For instance, a partial scalar invariance model will be a model with some of the intercepts/thresholds varying across groups.

Past research has demonstrated the importance of correctly identifying non-invariant items in partial invariance models. On one hand, failing to release wrong equality constraints could distort the estimated latent mean differences across groups (Shi, Song, & Lewis, 2017a) and group comparisons (e.g., French and Finch, 2016). On the other hand, imposing equality constraints on invariant parameters can improve the statistical power of detecting latent mean differences across groups (Xu & Green, 2016). Shi, Song, and Lewis (2017a) also showed that in comparison to models with only one correctly specified invariant item, partial invariance models with multiple correctly specified equality constraints could generate more accurate and efficient estimates (e.g., factor means and factor loadings).

Sequential backward specification search for partial invariance models

Establishing a correct partial invariance model involves an iterative process of specification search for non-invariant parameters. Ideally, this search should be guided by both theory and statistical criteria. In practice, however, it often relies solely on statistical criteria (Yoon & Kim, 2014). Multiple specification search methods have been proposed (e.g., Huang, 2018; Jung & Yoon, 2016; Yoon & Millsap, 2007; Shi, Song, Liao, Terry, & Snyder 2017b). These methods basically fall into two categories: forward search and backward search methods. Forward search methods start the search

with a model that does not contain any equality constraints on target parameters, while backward search methods start the search with a model with equality constraints on all target parameters. In a partial scalar invariance model, for example, a forward search method will start the search from the full metric invariance model, while a backward search method will start with the full scalar invariance model, which is the model that has been rejected. Note that the definitions above are provided by previous research that focuses on the search for partial invariance models (Jung & Yoon, 2016); while if one adopts the definitions that the backward search means to initiate the search from a more general model (e.g., Chou & Bentler, 2002), the search starts with a full invariance model with equality constraints on all target parameters will then become a "forward" search method instead.

In addition, different statistical criteria are used in the two types of search methods. Forward specification search usually relies or statistics such as confidence intervals (CI). Backward search methods, in comparison, often use modification indices (MFI) to identify implausible equality constraints (e.g., Byrne, 2013; Yoon & Millsap, 2007; Byrne et al. 1989; Oort, 1998; Yoon & Millsap, 2007). MFI measures the decrease in the Chi-square test statistic (i.e., improvement in goodness of fit) when a fixed parameter or an equality constraint is freed (Sörbom, 1989). MFI is assumed to follow a Chi-square distribution with $df = 1$. The search starts with releasing the equality constraint which has the largest MFI among all significant constraints (i.e., constraints with MFI > 3.841, the critical value of a χ^2 statistic with $df = 1$ and $\alpha = 0.05$), and then refits the model to identify the second equality constraint to release. The search process continues until none of the MFIs for the remaining equality constraints is significant, or until there is no need for model modification because of a good overall model fit indicated by global fit indices such as a non-significant χ^2 test ($p < 0.05$). (Kim & Yoon, 2014; Yoon & Millsap, 2007). This sequential search procedure is referred to as the sequential backward specification search method with the largest MFI (SBSS_LMFI).

Among the available search methods, SBSS_LMFI is not only one of the most widely used methods (Kim & Yoon, 2014) but also one of the few search methods that have been validated across continuous and ordinal indicators (e.g., Jung & Yoon, 2016; Kim & Yoon, 2014; Whittaker & Khojasteh, 2013; Yoon & Millsap, 2007). Thus, we believe it will be a good starting point for us to study the missing data issues in specification search. The other specification search methods will be briefly discussed at the end of the article.

Methods to handle ordinal missing data in SBSS_LMFI

Even though SBSS_LMFI has been validated with complete data (e.g., Kim & Yoon, 2014; Yoon & Millsap, 2007; Jung & Yoon, 2016; Shi, Song & Lewis, 2017a), to the best of our knowledge, no study on SBSS_LMFI so far has considered the issue of missing data. As mentioned above, two methods can be used to deal with ordinal missing data in SBSS_LMFI. These methods are described in detail below along with their strengths and limitations.

Robust full information maximum likelihood (rFIML) rFIML is an extension of FIML to account for continuous data with non-normal distributions. To explain how rFIML works, we start with the log-likelihood function for FIML. FIML accounts for missing data by creating case-wise log-likelihood functions according to missing data patterns, allowing it to efficiently use all available information in the dataset to estimate model parameters.

The log-likelihood function used in FIML for case i is written as

$$l_i(\theta) = K_i - \frac{1}{2}\log\left|\Sigma(\theta)_i\right| - \frac{1}{2}(x_i - \mu_i)'\Sigma(\theta)_i^{-1}(x_i - \mu_i), \quad (1)$$

where K_i is a constant. $\boldsymbol{\Sigma}_i$, $\boldsymbol{\mu}_i$, and $x_i$are the model implied covariance matrix, mean vector, and the observed data for case i, respectively. The individual likelihoods are summed to form the sample log-likelihood function (Arbuckle, 1996, p248; Yuan & Bentler, 2000, p.167–168):

$$l(\theta) = \sum_{i=1}^{N} l_i(\theta). \quad (2)$$

The test statistic of the model can be then calculated as follows, and assumed to follow a Chi-square distribution.

$$T_{FIML} = -2\left(l\left(\widehat{\theta}\right) - l\left(\widehat{\beta}\right)\right), \quad (3)$$

where $l\left(\widehat{\theta}\right)$ and $l\left(\widehat{\beta}\right)$ are maximized log likelihoods under the tested and saturated models, respectively (Yuan & Bentler, 2000).

Past research has shown that FIML produces more accurate parameter estimates and χ^2 test statistics than traditional missing data methods such as listwise or pairwise deletion (PD) (Enders & Bandalos, 2001). Effectively recovering missing information often requires integrating auxiliary variables (the variables that predict missingness but are not part of the tested model) into the missing data handling process for CFA models. This can be done with FIML using Graham's saturated model method by allowing the correlations between auxiliary variables and residuals of manifest indicators to be freely estimated (Graham, 2003).

However, FIML assumes multivariate normality, which is often violated in practice and lead to biased test statistics. To solve the problem, rFIML corrects the test statistic by multiplying the test statistic in Eq. (3) by a correction factor c as follows.

$$T_{rFIML} = c \times T_{FIML}. \quad (4)$$

Detailed information on how c is calculated can be found in Yuan and Bentler (2000). Since MFI is also assumed to follow a Chi-square test distribution, it can be corrected in a similar fashion when rFIML is used (see Muthén, 2011, Jul, 26).

As a direct extension of FIML, rFIML inherits the capability of FIML to handle missing data in the estimation process while accounting for nonnormality. However, rFIML still treated ordinal data as continuous. Past research has shown that this could result in distorted test statistics, although the bias may be ignorable when the number of categories within an item is no less than five (e.g., Rhemtulla, Brosseau-Liard, & Savalei, 2012).

WLSMV with pairwise deletion (WLSMV_PD) Unlike rFIML, WLSMV is an ordinal estimator extended from the weighted least squares (WLS) estimation method. WLS assumes that for each ordinal indicator y_j with C categories, there is a normal distributed latent response variable (y_j^*) underlying it. This latent response variable is categorized into C categories using C-1 thresholds ($\tau_{j,\,1}, \tau_{j,\,2}, \ldots \tau_{j,\,c-1}$) such as

$$y_j = \begin{Bmatrix} 1 & if & y_j^* \le \tau_{j,1} \\ 2 & if & \tau_{j,1} \le y_j^* \le \tau_{j,2} \\ \vdots & \vdots & \vdots \\ C-1 & if & \tau_{j,c-2} < y_j^* \le \tau_{j,c-1} \\ C & if & \tau_{j,c-1} < y_j^* \end{Bmatrix}. \quad (5)$$

The thresholds and other parameters in the model (e.g., loadings) are typically estimated via a three-stage process (e.g., Muthén, 1984; Muthén, De Toit & Spisic, 1997; Wirth & Edwards, 2007). First, the sample thresholds for each indicator are estimated using univariate information. Second, the polychoric correlation between each pair of ordinal indicators is estimated by treating the sample thresholds estimated in the previous step as fixed (Olsson, 1979 and Bollen, 1989, p.439–443). Third, the sample thresholds and polychoric correlations obtained in steps 1 and 2 are used to form a discrepancy function, which is minimized to obtain the estimates for the model parameters ($\widehat{\theta}$). The discrepancy function WLS can be written as

$$\mathrm{F_{WLS}} = (s - \sigma(\theta))' W^{-1} (s - \sigma(\theta)), \quad (6)$$

where s is a vector of sample thresholds and polychoric correlations (unique elements in the sample polychoric correlation matrix), $\theta$is a vector of model-implied thresholds and polychoric correlations; W is the weight matrix, which is usually a consistent estimate of the covariance matrix of **s**. The test statistic can be then calculated as

$$T_{WLS} = (N-1) \times F_{WLS}\left(\widehat{\theta}\right), df = p^* - q, \quad (7)$$

where $F_{WLS}\left(\widehat{\theta}\right)$ is the minimized discrepancy function, N is the sample size, p^* is the number of unique elements in **s**, and q is the number of parameters in the model.

T_{WLS} asymptotically follows a χ^2 distribution, though simulation studies show that it requires a very large sample size, which is often impractical for most research in social and behavioral sciences (e.g., Flora & Curran, 2004). A solution to this problem is to use only the diagonal elements of the weight matrix in Eq. (6) in the discrepancy function as shown below

$$\mathrm{F_{DWLS}} = (s - \sigma(\theta))' W_D^{-1} (s - \sigma(\theta)), \quad (8)$$

where$\mathbf{W}_D$ is a matrix with all the off-diagonal elements in **W** fixed to 0s. This approach is known as diagonally weighted least squares estimation (DWLS) (Muthén, et al., 1997; Wirth & Edwards, 2007).

The information loss caused by diagonalizing the weight matrix could distort the test statistic (Savalei, 2014). WLSMV provides a correction to DWLS for the information loss so that the mean and variance of the test statistic will approximate a χ^2 distribution (DiStefano & Morgan, 2014; Muthén, et al., 1997). Previous research found that WLMSV outperformed other correction methods (DiStefano & Morgan, 2014). Thus, we considered only WLMSV in the current study.

For complete ordinal data, WLSMV appears to outperform rFIML. It provides more accurate test statistics (Li, 2016) and valid MFI (Yoon & Kim, 2014). However, WLMSV has its limitations when missing data present (Muthén, Muthén & Asparouhov, 2015). As mentioned earlier, WLSMV is a multi-stage process that uses only univariate and bivariate information in the first two stages. This makes WLSMV a limited information method instead of a full information estimator. Consequently, it cannot incorporate missing data patterns into its discrepancy function and relies on other methods to deal with missing data. In practice, PD has often been coupled with WLSMV for dealing with missing data (e.g., Chan, Gerhardt & Feng, 2019; Erreygers, Vandebosch, Vranjes, Baillien, & De Witte, 2018; Hakkarainen, Holopainen, & Savolainen, 2016; Kim, Wang & Sellbom, 2018; Willoughby, Pek, Greenberg & Family Life Project Investigators, 2012). In the current study, we refer to this combination as WLSMV_PD.

PD is not an ideal missing data technique. Past research found that PD could substantially inflate the type I error rates of χ^2 tests in SEM with continuous data (e.g., Enders & Bandalos, 2001; Savalei & Bentler, 2005). A better method

for dealing with missing data with WLSMV is multiple imputation (Asparouhov, & Muthén, 2010; Teman, 2012). For example, Asparouhov & Muthén, (2010) found that WLSMV combined with multiple imputation would provide more accurate point estimates than WLSMV_PD. Nevertheless, we did not consider multiple imputation in the current study because so far, there is no good way to pool MFI and the χ^2 test statistic across imputations with WLSMV (Liu et al., 2017; Muthén, 2017).

Purpose of the current research

As described above, rFIML and WLSMV_PD both have their own strengths and limitations. rFIML is good at handling missing data, but it assumes that ordinal data are continuous. In contrast, WLSMV_PD is good at handling ordinal data, but it uses a suboptimal method to deal with missing data. To the best of our knowledge, it is still not clear which one will perform better or under what conditions one would be preferred over the other in SBSS. Without such information, researchers have often chosen one of the methods based on personal preference, without solid justifications (e.g., Fokkema, et al., 2013). To fill in this gap in the literature, the current study aims to examine the relative performance of rFIML and WLSMV_PD in SBSS_LMFI with ordinal missing data using Monte Carlo simulation.

Simulation design

Population generation model

Following Jung and Yoon (2016), we used a two-group (groups A and B) single-factor CFA model as the population model. There were six or 12 Likert-type indicators per group, representing smaller or larger models. A population model with six indicators per group is shown in Fig. 1. Group A was used as the reference group, for which all the model parameters were fixed across all conditions. Group B was the focus group, where the model parameters for some indicators were varied across conditions.

We created invariant and non-invariant conditions, allowing us to examine both type I and type II error rates associated with the methods. For invariant conditions, all items in groups A and B had factor loadings fixed at 0.7 (Jung & Yoon, 2016; Sass et al., 2014). To create non-invariant conditions, certain values (depend on the pattern of non-invariance) were subtracted from either the loadings or the thresholds of items 2 and 4 in group B for the six indicators per group model. For the 12 indicators per group model, non-invariance occurred on items 2, 4, 6, and 10. The number of non-invariant items is doubled here to maintain the proportion of non-invariant items constant. The residual term of each indicator follows a normal distribution with mean at 0 and variance as 1 – the square of the corresponding loading.

In addition, for both invariant and non-invariant conditions, there were complete data and missing data conditions. For missing data conditions, substantial amounts of missing data were imposed on all non-invariant items (i.e., items 2 and 4 in group B for the smaller model and items 2, 4, 8, and 10 in group B for the larger model). The missingness was determined by an auxiliary continuous variable correlated with the latent factor in group B by $r = 0.5$. Thus, the generated missing data mechanism was missing at random (MAR). Note that in the missing data conditions, in addition to the MAR data mentioned above, we also imposed 5% missing completely at random (MCAR) data on all items in the model to increase the external validity of our simulations, given that missing data could occur on all items and different missing data mechanisms could coexist in reality. The details of the missing data conditions and missing data generation process are explained below.

Design factors

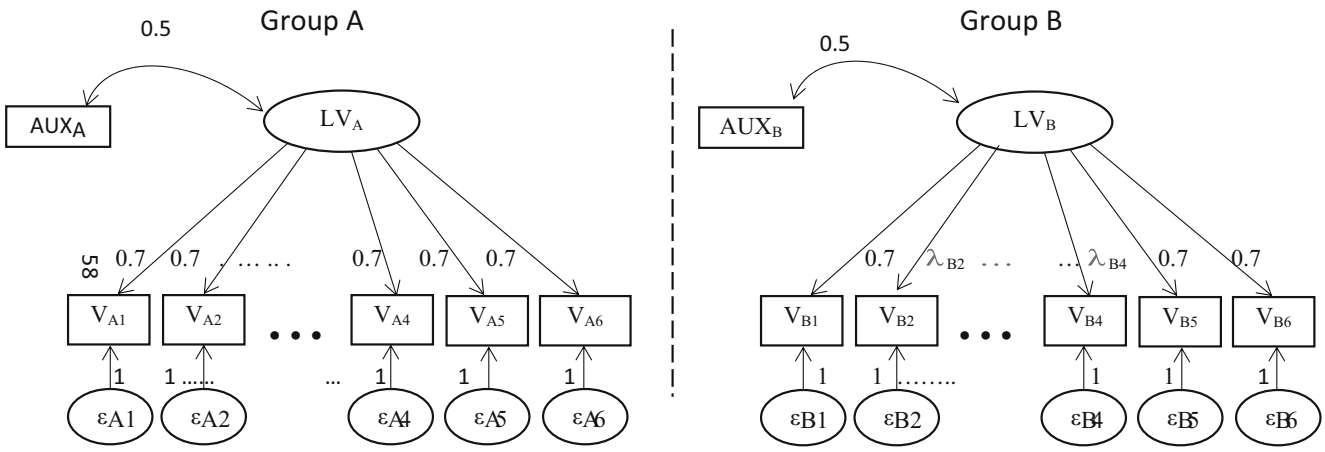
The design factors we manipulated included (1) Sample size, (2) Model size, (3) Number of categories within each indicator, (4) Distribution of thresholds, (5) Location of non-invariance, (6) Pattern of non-invariance, and (7) Missing data proportion. The first six factors were between-replication factors. The last factor was a within-replication factor.

Sample size The sample size was varied at three levels: 500 (250 per group), 1000 (500 per group), or 2000 (1000 per group), representing small, medium, or large sample sizes. These settings are identical to those adopted in Jung & Yoon (2016).

Model size The number of items per group was varied at two levels: six items (a smaller model) per group and 12 items per group (a larger model), referring to previous research on missing data and specification search problems in the framework of MF-CFA (Enders & Gottschall, 2011; Yoon & Millsap, 2007)

Number of categories per item We varied this factor at two levels: three or five. As mentioned above, past research has recommended using rFIML for ordinal data with ≥ 5 categories; however, it would still be interesting to see how rFIML would perform in relative to WLSMV_PD for ordinal data with less than five categories.

Distribution of thresholds The distribution of thresholds is varied to be either symmetric or asymmetric. For five-point indicators, (– 1.30, – 0.47, 0.47, 1.30) and (–0.253, 0.385, 0.842, 1.282) are used to represent symmetric and asymmetric



Note: Aux: auxiliary variable. V_A: observed indicators ingroup A. V_B: observed indicators ingroup A

Fig. 1. The population model with six indicators. Note: Aux: auxiliary variable. V_A: observed indicators ingroup. A. V_B: observed indicators ingroup A

thresholds, respectively, following Sass et al. (2014). For three-point indicators, (– 0.83, 0.83) and (– 0.50, 0.76) are used to represent symmetric and asymmetric thresholds, respectively, following Rhemtulla et al. (2012).

Location of non-invariance Non-invariance was placed on either the loadings or thresholds of non-invariant items (i.e., items 2 and 4 for the smaller model and items 2, 4, 8, and 10 for the larger model). For convenience, we refer to the two types of conditions as loading and threshold non-invariant conditions, respectively.

Pattern of non-invariance In the non-invariant conditions, the loadings or thresholds of non-invariant items were subtracted by certain values (depending on the pattern of non-invariance). We created four patterns of non-invariance: uniform (small, large, and mixed) and non-uniform (Jung & Yoon, 2016; Meade & Bauer, 2007). For uniform patterns, the directions of non-invariance were consistent across non-invariant items. Specifically, the loadings or thresholds of non-invariant items in group B were subtracted by (0.2, 0.2), (0.4, 0.4), or (0.3, 0.5) in conditions for the smaller model and by (0.2, 0.2, 0.2, 0.2), (0.4, 0.4, 0.4, 0.4), or (0.3, 0.5, 0.3, 0.5) in conditions for the larger model to create small, large, or mixed non-invariant conditions (see also Jung & Yoon, 2016).

For non-uniform non-invariance, the directions of non-invariance were different across the items. Specifically, the loadings or thresholds of non-invariant items in group B were subtracted by (0.2, – 0.2) in conditions for the smaller model and by (0.2, – 0.2, 0.2, – 0.2) in conditions for the larger model. Model parameters for these patterns of non-invariance for five-point indicators are presented in Table 1. Those for three-point indicators can be found in the supplementary material. We also quantified the magnitude (i.e., effect size) of non-invariance for each of these patterns using the d_{macs} statistics proposed by Nye & Drasgow (2011). Given the space limit, these effect sizes along with their interpretation guidelines provided by Nye, Bradburn, Olenick, Bialko & Drasgow (2019) can be also found in supplementary materials.

Missing data proportion We imposed MAR missing data on non-invariant items in group B. We varied the MAR data rate at three levels: 0% (no missing data), 30% MAR (medium) and 50% MAR (large). Similar to Wu, Jia and Enders (2015), the missingness on these items was determined by the auxiliary variable (see Fig. 1), such as that participants with higher scores on the auxiliary variable in focal group were more likely to have missing data. Specifically, three steps are involved. First, auxiliary variable values in focal group were rank ordered. Second, the probabilities of having missing data on items 2 and 4 for the smaller model and on items 2, 4, 8, and 10 for the larger model for each individual were then calculated as $1 - rank_i/n_b$. Here $rank_i$ is the rank order of the auxiliary variable score for individual i and$n_b$is the sample size of group B. Third, for each participant and incomplete item, a random number u was generated from a uniform distribution. If u was less than the probability, then the data point was missing. This procedure was repeated for non-invariant items until the desired MAR data rate (30% or 50%) was reached for each item. As mentioned above, in addition to the MAR data, we added 5% MCAR data to all items in both groups in the missing data conditions.

In sum, there are 240 between-replication conditions in total. Among them, there are 192 non-invariant conditions: sample sizes (3) × model sizes (2) × number of categories

Table 1. Model parameters of different patterns of non-invariance with five-point indicators

		Group B				
Parameter loadings	Group A	Baseline (item 1,3,5,6 in group B)	Small difference	Large difference	Mixed-size difference	Non-uniform difference
Loading in item 1&7[a]	.7	.7	.7	.7	.7	.7
Loading in item 2&8	.7	.7	**.5**	**.3**	**.4**	**.5**
Loading in item 3&9	.7	.7	.7	.7	.7	.7
Loading in item 4&10	.7	.7	**.5**	**.3**	**.2**	**.9**
Loading in item 5&11	.7	.7	.7	.7	.7	.7
Loading in item 6&12	.7	.7	.7	.7	.7	.7
Symmetric threshold						
item 1 & 7	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)
item 2 & 8	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(**– 1.5, – 0.67, 0.27, 1.1**)	(**– 1.7, – 0.87, 0.07, 0.9**)	(**– 1.6, – 0.77, 0.17, 1.0**)	(**– 1.6, – 0.77, 0.17, 1.0**)
item 3& 9	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)
item 4 & 10	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(**– 1.5, – 0.67, 0.27, 1.1**)	(**– 1.7, – 0.87, 0.07, 0.9**)	(**– 1.8, – 0.97, – .03, 0.8**)	(**– 1, – 0.17, 0.77, 1.60**)
item 5 & 11	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)
item 6 & 12	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)	(– 1.3, – 0.47, 0.47, 1.3)
Asymmetric threshold						
item 1 & 7	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)
item 2 & 8	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– **0.453, 0.185, 0.642, 1.082**)	(– **0.653, – 0.015, 0.442, 0.882**)	(– **0.553, 0.085, 0.542, 0.982**)	(– **0.453, 0.185, 0.642, 1.082**)
item 3& 9	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)
item 4 & 10	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– **0.453, 0.185, 0.642, 1.082**)	(– **0.653, – 0.015, 0.442, 0.882**)	(– **0.753, – 0.115, 0.342, 0.782**)	(**– 0.053, 0.585, 1.042, 1.482**)
item 5 & 11	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)	(– 0.253, 0.385, 0.842, 1.282)
item 6 & 12	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)	(–0.253, 0.385, 0.842, 1.282)

Note: [a] the parameter settings for item 7 – item 12 only apply to conditions with the larger model. Non-invariance occurred either on loadings or thresholds in items 2 and 4 in group B. The residual variance of each item is equal to 1- squared loading.

per item (2) × distribution of thresholds (2) × locations of non-invariance (2) × patterns of the non-invariance (4). There are 48 invariant conditions: sample size (3) × model sizes (2) × number of categories per item (2) × distribution of thresholds (2) × loading or threshold invariance (2). Note that the terms "loading" or "thresholds" in the invariant conditions only represent the target parameters of the search. All parameters (loadings and thresholds) are invariant in the invariant conditions. There are 500 replications for each condition. All datasets were generated using R 3.3.1 (R core team, 2016).

Implementation of the methods

For each replication, we conducted SBSS_LMFI using rFIML and WLSMV_PD (Muthén & Muthén, 1998–2017). The auxiliary variable was included in both groups using the saturated correlation model proposed by Graham (2003). In SBSS_LMFI, we used 3.841 as the cutoff for a significant MFI. Following Yoon & Kim (2014), the search procedure continues until there is no significant MFI in the model on target parameters or until the global χ^2 test became non-

Table 2 Model level type I error rates in invariance conditions with five-point indicators

				Symmetric thresholds		Asymmetric thresholds	
Model Size	DIF model	Sample size	MAR rates	rFIML	WLSMV_PD	rFIML	WLSMV_PD
6Items	Metric	500	0%	0.026	0.060	0.030	0.034
			30%	0.014	0.048	0.032	0.048
			50%	0.018	0.082	0.032	**0.145**
		1000	0%	0.028	0.048	0.020	0.034
			30%	0.022	0.074	0.022	0.056
			50%	0.032	**0.188**	0.024	**0.347**
		2000	0%	0.018	0.064	0.016	0.030
			30%	0.026	0.094	0.028	**0.146**
			50%	0.036	**0.406**	0.020	**0.692**
	Scalar	500	0%	0.028	0.062	0.036	0.056
			30%	0.028	**0.106**	0.030	**0.104**
			50%	0.038	**0.316**	0.016	**0.226**
		1000	0%	0.024	0.056	0.032	0.054
			30%	0.036	**0.148**	0.024	**0.124**
			50%	0.020	**0.610**	0.026	**0.498**
		2000	0%	0.016	0.040	0.022	0.058
			30%	0.020	**0.304**	0.024	**0.272**
			50%	0.020	**0.940**	0.014	**0.856**
12 items	Metric	500	0%	0.046	0.038	0.022	0.028
			30%	0.042	0.054	0.036	0.058
			50%	0.046	**0.114**	0.050	**0.126**
		1000	0%	0.026	0.042	0.020	0.030
			30%	0.034	0.052	0.020	0.070
			50%	0.030	**0.240**	0.030	**0.358**
		2000	0%	0.016	0.050	0.032	0.050
			30%	0.034	0.096	0.024	**0.140**
			50%	0.032	**0.540**	0.034	**0.756**
	Scalar	500	0%	0.058	0.048	0.044	0.060
			30%	0.064	**0.104**	0.048	0.092
			50%	0.056	**0.322**	0.054	**0.276**
		1000	0%	0.028	0.038	0.028	0.038
			30%	0.032	**0.112**	0.016	**0.144**
			50%	0.038	**0.674**	0.032	**0.586**
		2000	0%	0.032	0.054	0.018	0.056
			30%	0.032	**0.286**	0.024	**0.292**
			50%	0.022	**0.984**	0.026	**0.948**

Note: the type I error rates above 0.1 are highlighted in bold

significant ($p > 0.05$). We used the loading and intercept (first threshold) of item 1 as the anchor parameters; they were always constrained to be equal across groups thus were not included in the search. For WLSMV_PD, the model was specified using delta parameterization. Details about this parameterization can be found in Muthén and Asparouhov (2002). Given that WLSMV_PD estimates thresholds and there are multiple thresholds in an ordinal indicator, the specification search of WLSMV_PD involves more steps than rFIML n the thresholds non-invariant conditions (note: rFIML only search one intercept per item in the threshold non-invariant conditions).

Outcomes

We evaluated the two methods using three outcome variables that have been considered in past research (Jung & Yoon, 2016; Yoon & Kim, 2014; Yoon & Millsap, 2007): model

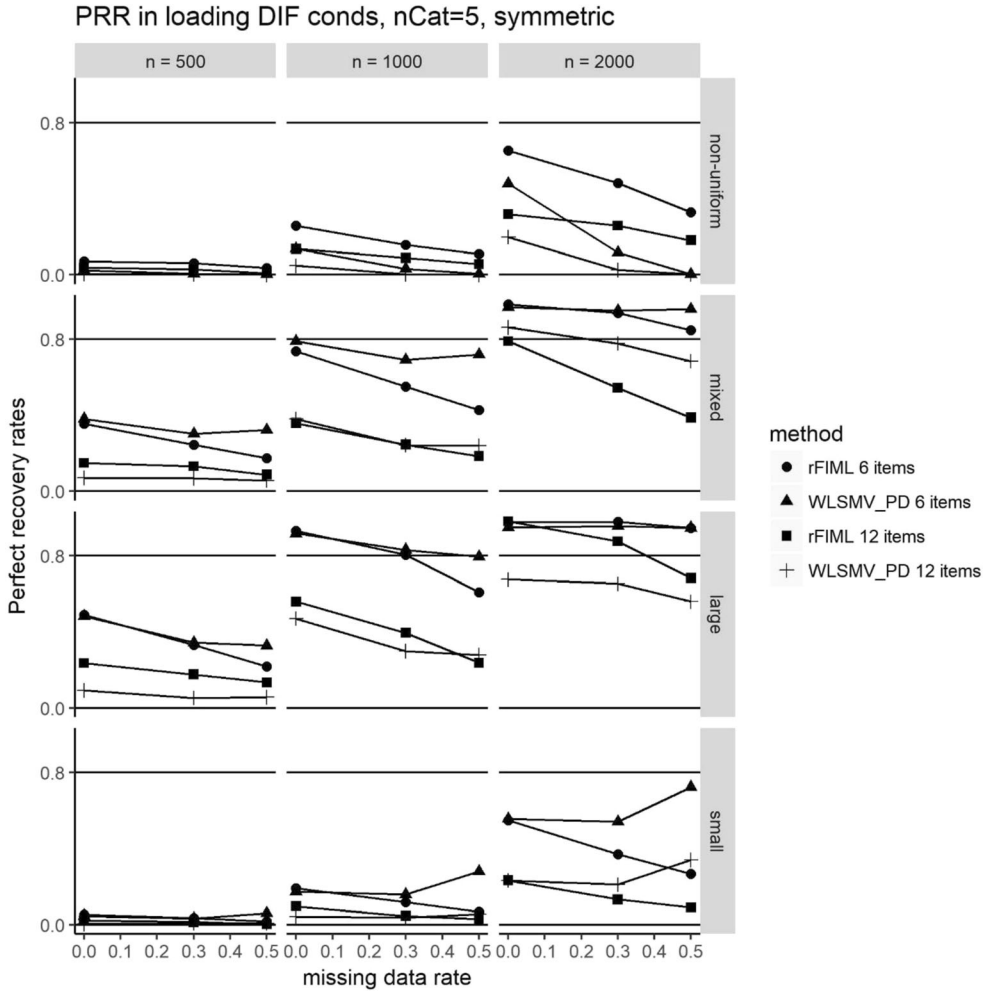


Fig. 2 Perfect recovery rates of the methods in loading non-invariance conditions with five-point symmetric indicators

level type I and type II error rates, and perfect recovery rates. The *model level type I error rate* is the probability of misidentifying any invariant parameters as non-invariant across all replications of a condition; the *model level type II error rate* is the probability of misidentifying any non-invariant parameters as invariant across all replications of a condition. The *perfect recovery rate* is defined as the probability of correctly identifying all non-invariant and invariant parameters in all items across all replications of a condition.

Note that with the definitions, model level type I errors and type II errors could appear in a partial invariance model in non-invariance conditions simultaneously. In contrast, the perfect recovery rate is considered as a criterion that is similar to but more rigorous than power (Jung & Yoon, 2016), as it is a function of both type I and type II error rates. A perfect recovery can be achieved only when neither type I nor type II errors occurred. For invariant conditions, only model level type I errors could occur and the perfect recovery rates would be equal to 1 – model level type I error rates. Thus, we only calculated and reported model level type I errors for invariant conditions.

Results

All specification searches conducted using rFIML converged. About 0.3% of the replications had convergence problems with WLSMV_PD. Most of these non-convergent replications occurred with the larger models, especially in loading non-invariance conditions where at least one loading in the model was lower than 0.4 with small sample sizes. These replications were excluded from the rest of the analyses.

Because five-point Likert scales are much more popular than three-point Likert scales, we present the results for five-point indicators first and then show how the results for three-point indicators were similar or different. In addition, given relative performances between methods are similar and consistent between conditions with six item per group model and

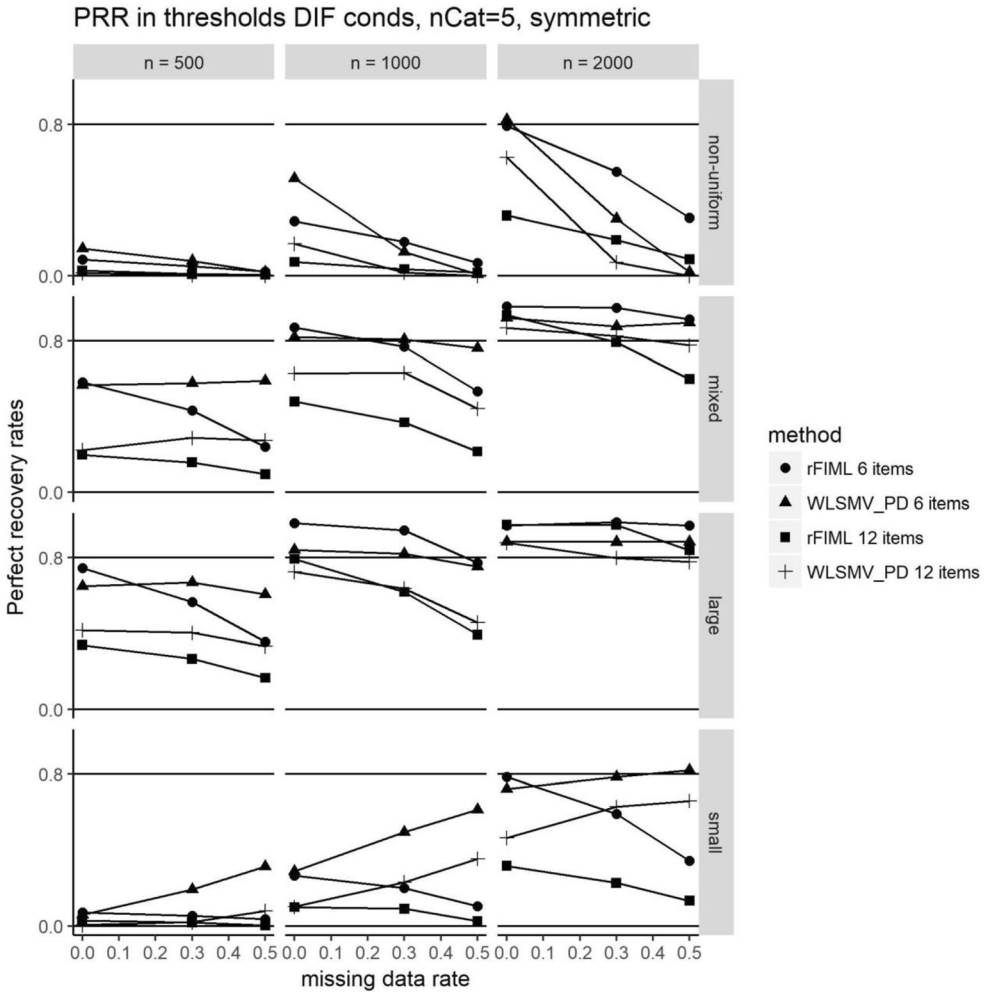


Fig. 3 Perfect recovery rates of the methods in thresholds non-invariance conditions with five-point symmetric indicators

the conditions with the 12 item per group models, we thus focused on describing the results for the smaller model.

The results for five-point indicators

Model level type I error rates and perfect recovery rates for invariant conditions The result for the model level type I error rates is shown in Table 2. The type I error rates above 0.1 are highlighted. When data were complete, both methods maintained the type I error rate around or below the nominal level (0.05). However, when data were missing, rFIML substantially outperformed the WLSMV_PD in controlling type I error rates. The type I error rates from rFIML were below or around .05 across all missing data conditions. In contrast, WLSMV_PD yielded inflated type I error rates, particularly in the conditions with large sample sizes and high missing data rates. For example, in the conditions where the sample size was 2000 and the missing data rate was 50%, the type I error rates from WLSMV_PD were inflated up to 0.40 and 0.98 for the metric and scalar invariance models, respectively. As mentioned above, for non-invariant conditions, perfect recovery rates are equal to 1 minus type I error rates, thus lower type I error rates can be directly translated to higher perfect recovery rates.

Perfect recovery rates for non-invariant conditions The perfect recovery rates of conditions with five-point indicators are presented in Figs. 2, 3, 4, and 5. As shown in these figures, the results from both methods shared the following patterns: (a) As sample size increased, the perfect recovery rate also increased; (b) As missing data rate increased, the perfect recovery rate decreased in most conditions; (c) As model size increased, the perfect recovery rate decreased; (d) Neither of the methods had sufficient perfect recovery rates (> 0.8) when sample size was smaller than 1000, regardless of the patterns of non-invariance. In addition, as the thresholds changed from symmetric to asymmetric, the perfect recovery rate tended to slightly decrease. The detailed values of the perfect recovery rates in the conditions can be found in supplementary materials.

The perfect recovery rates were more comparable between rFIML and WLSMV_PD under complete data conditions.

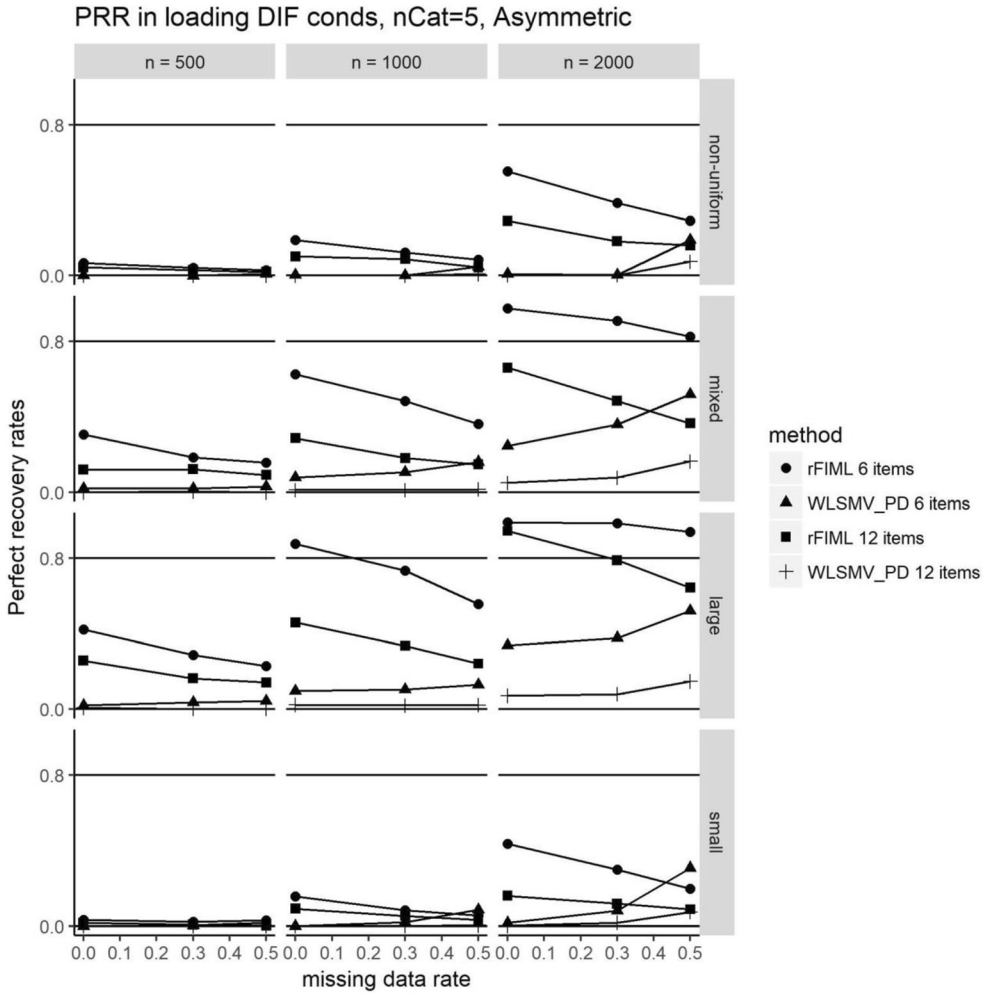


Fig. 4 Perfect recovery rates of the methods in loading non-invariance conditions with five-point asymmetric indicators

The performance of the two methods diverged with the presence of missing data. Specifically, perfect recovery rates of rFIML were more consistent across design factors and tended to decrease as missing data rates increased across conditions. Factors such as patterns of non-invariance appeared to have more profound and complex impacts on WLSMV_PD when missing data present. In most conditions with uniform non-invariance (i.e., the non-invariance of items followed the same direction), the perfect recovery rates from WLSMV_PD only slightly decreased as the missing data rate increased and in general were higher than those from rFIML. In contrast, for non-uniform non-invariance conditions, the perfect recovery rates from WLSMV_PD substantively decreased as the missing data rate increased and were lower than those from rFIML.

In addition to the pattern of non-invariance, the relative performance between rFIML and WLSMV_PD was also affected by the location of non-invariance and threshold asymmetry. Specifically, when non-invariance occurred in loadings of indicators with asymmetric thresholds (see Fig. 4), rFIML outperformed WLSMV_PD in perfect recovery rate even under uniform non-invariance conditions.

Model level type I and type II error rates for non-invariant conditions The type I error rates for the loading non-invariant conditions for the smaller model are shown in Table 3. As shown in the table, rFIML provided a good control of the type I error rate across all conditions. In contrast, WLSMV_PD yielded inflated model level type I error rates and the type I error rates increased as the missing data rate increased.

Tables 4 and 5 display the type II error rates for the smaller model . The type II error rates higher than 0.2 are highlighted in these tables. The type II error rates of rFIML generally increased as the missing data rate increased with only a few exceptions. Both methods had type II error rates lower than 0.2 when thresholds were symmetric, magnitude of non-invariance and sample size were large.

Comparing the two methods, rFIML tended to have lower model level type II error rates than WLSMV_PD under non-uniform non-invariance conditions, particularly with the presence of missing data. In contrast, for uniform non-invariance, WLSMV_PD outperformed rFIML, except when non-

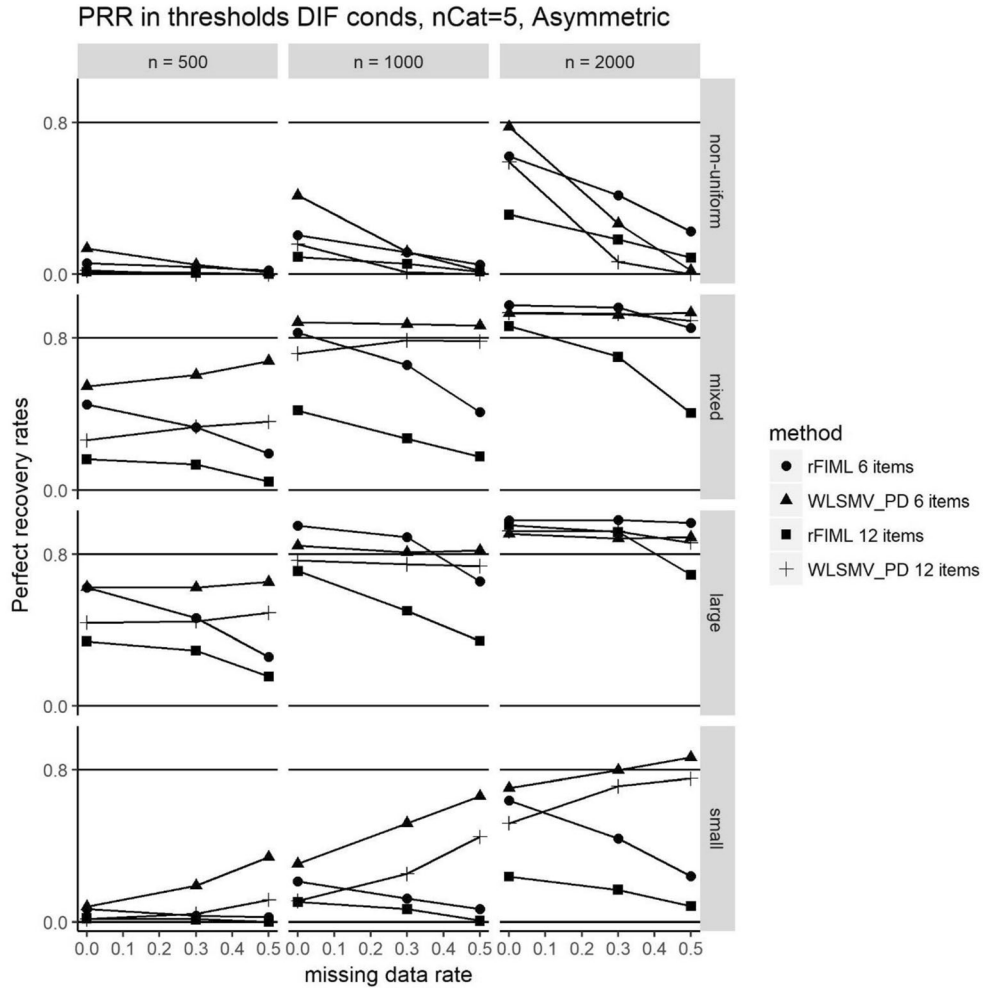


Fig. 5 Perfect recovery rates of the methods in thresholds non-invariance conditions with five-point asymmetric indicators

invariance occurred in loadings and thresholds were asymmetric.

The results for three-point indicators

The results from the three-point indicators were mostly consistent with those from the five-point indicators. For instance, rFIML continued outperforming WLSMV_PD in controlling the model level type I error rates (see Table 6). For perfect recovery rates, similar patterns were observed between the results for the three-point and five-point indicators (see Fig. 6), although the perfect recovery rates of rFIML under the three-point indicators were lower than those from the five-point indicators by comparing Fig. 6 to Fig. 2. In general, rFIML outperformed WLSMV_PD under non-uniform non-invariance or when non-invariance occurred in loadings and thresholds were asymmetric. WLMSV_PD, on the other hand, outperformed rFIML under uniform non-invariance.

There is one difference, however, between the two types of indicators. This difference is shown in the upper panel of Fig. 7. It appeared that with three-point indicators, WLSMV_PD could outperform rFIML under non-uniform non-invariance when the non-invariance occurred in the thresholds and there was a larger number of asymmetric items (i.e., 12 items per group) in the model.

Conclusions and discussion

Specification search is an important follow-up procedure in ME/I when full invariance models cannot be achieved (Putnick, & Bornstein, 2016). Focusing on backward specification search, we evaluated the performance of two commonly used methods, rFIML and WLSMV_PD, by examining their model-level type I rates, type II error rates, and perfect recovery rates in the current study. The major findings are summarized and discussed below.

Major findings

First, in both invariant and non-invariant conditions, we consistently found that rFIML provided a better control of model-

Table 3 Model level type I error rates in loading non-invariance conditions with six five-point indicators per group

			Symmetric thresholds		Asymmetric thresholds	
DIF type	Sample size	MAR rates	rFIML	WLSMV_PD	rFIML	WLSMV_PD
Small	500	0%	0.048	0.076	0.034	0.018
		30%	0.046	**0.112**	0.034	0.030
		50%	0.046	**0.144**	0.036	0.059
	1000	0%	0.044	**0.110**	0.044	0.044
		30%	0.042	**0.142**	0.058	0.070
		50%	0.056	**0.146**	0.036	**0.100**
	2000	0%	0.028	0.084	0.012	0.056
		30%	0.020	0.098	0.018	0.076
		50%	0.034	0.098	0.028	**0.106**
large	500	0%	0.024	**0.131**	0.026	0.030
		30%	0.030	**0.132**	0.028	0.032
		50%	0.054	**0.174**	0.040	0.047
	1000	0%	0.012	0.046	0.012	0.044
		30%	0.010	0.074	0.014	0.062
		50%	0.012	0.074	0.016	0.067
	2000	0%	0.022	0.050	0.006	0.042
		30%	0.018	0.042	0.008	0.064
		50%	0.022	0.052	0.012	0.052
Mixed	500	0%	0.020	0.088	0.026	0.030
		30%	0.036	**0.125**	0.020	0.027
		50%	0.036	**0.149**	0.030	0.079
	1000	0%	0.010	0.034	0.008	0.034
		30%	0.010	0.042	0.012	0.050
		50%	0.024	0.058	0.018	0.077
	2000	0%	0.010	0.028	0.006	0.034
		30%	0.016	0.038	0.006	0.062
		50%	0.012	0.030	0.008	0.074
Non-uniform	500	0%	0.026	0.052	0.036	0.014
		30%	0.032	0.066	0.024	0.022
		50%	0.034	0.064	0.038	0.040
	1000	0%	0.032	0.068	0.014	0.038
		30%	0.024	0.056	0.020	0.050
		50%	0.040	0.066	0.012	0.097
	2000	0%	0.012	0.066	0.018	0.018
		30%	0.028	0.076	0.014	0.062
		50%	0.018	0.048	0.008	**0.140**

Note: Model level type I error rates above 0.1 are highlighted in bold

level type I error rates than WLSMV_PD across all the conditions when missing data present, including those with three-point indicators. This suggests that the advantage of WLSMV_ PD in dealing with the ordinal nature of the data is comprised by its use of PD. At least two aspects of WLSMV_PD could contribute to the inflated type I error rates. First, it treats the sample thresholds and polychoric correlations as if they were obtained from complete data, thus fails to account for the uncertainty due to missing data in estimating these parameters. Second, depending on missing data patterns, the thresholds and polychoric correlations are likely calculated using different sample sizes, which could distort the globalχ^2 statistics (Bollen, 1989). Since MFI is a χ^2 statistic, this could also bias MFI.

Second, in non-invariance conditions, we found that decrease in sample size, increase in missing data rate, and

Table 4 Model level type II error rates in loading non-invariance conditions with six five-point indicators per group

			Symmetric thresholds		Asymmetric thresholds	
DIF type	Sample size	MAR rates	rFIML	WLSMV_PD	rFIML	WLSMV_PD
Small	500	0%	**0.946**	**0.946**	**0.964**	**1.000**
		30%	**0.964**	**0.966**	**0.978**	**0.996**
		50%	**0.980**	**0.926**	**0.968**	**0.982**
	1000	0%	**0.806**	**0.804**	**0.836**	**1.000**
		30%	**0.874**	**0.822**	**0.912**	**0.980**
		50%	**0.930**	**0.688**	**0.940**	**0.908**
	2000	0%	**0.438**	**0.415**	**0.558**	**0.980**
		30%	**0.620**	**0.422**	**0.694**	**0.908**
		50%	**0.726**	**0.218**	**0.796**	**0.670**
large	500	0%	**0.496**	**0.457**	**0.568**	**0.980**
		30%	**0.654**	**0.621**	**0.704**	**0.962**
		50%	**0.766**	**0.619**	**0.760**	**0.955**
	1000	0%	**0.056**	0.036	**0.110**	**0.896**
		30%	**0.184**	0.110	**0.250**	**0.888**
		50%	**0.382**	0.151	**0.428**	**0.859**
	2000	0%	0.000	0.000	0.002	**0.648**
		30%	0.002	0.000	0.006	**0.596**
		50%	0.030	0.000	0.048	**0.462**
Mixed	500	0%	**0.628**	**0.578**	**0.684**	**0.976**
		30%	**0.746**	**0.675**	**0.808**	**0.980**
		50%	**0.822**	**0.641**	**0.836**	**0.967**
	1000	0%	**0.254**	**0.182**	**0.366**	**0.918**
		30%	**0.440**	**0.275**	**0.506**	**0.890**
		50%	**0.556**	**0.238**	**0.630**	**0.829**
	2000	0%	0.006	0.004	0.020	**0.744**
		30%	0.044	0.012	0.086	**0.620**
		50%	0.140	0.014	0.166	**0.448**
Non-uniform	500	0%	**0.928**	**0.964**	**0.930**	**0.998**
		30%	**0.936**	**0.994**	**0.958**	**1.000**
		50%	**0.964**	**0.998**	**0.968**	**0.992**
	1000	0%	**0.728**	**0.848**	**0.804**	**0.996**
		30%	**0.836**	**0.962**	**0.872**	**0.998**
		50%	**0.884**	**0.990**	**0.914**	**0.950**
	2000	0%	**0.338**	**0.492**	**0.438**	**0.990**
		30%	**0.494**	**0.876**	**0.610**	**0.996**
		50%	**0.660**	**0.994**	**0.708**	**0.798**

Note: Model level type II error rates above 0.2 are highlighted in bold

decrease in the amount of non-invariance were generally associated with lower perfect recovery rates for both methods. When sample size was small, neither method produced > .8 perfect recovery rates in most conditions. In addition, increase in model size and asymmetry of thresholds could also decrease perfect recovery rates. These results are consistent with the findings from previous research based on complete data. For example, Yoon & Millsap (2007) found that increase the model size decreased the perfect recovery rates of SBSS_LMF. Sass et al. (2014) found that asymmetric thresholds lowered the power of $\Delta\chi^2$ tests between invariance models for robust maximum likelihood (ML) estimator in all conditions and WLSMV in most conditions.

Third, WLSMV_PD produced higher perfect recovery rates than rFIML in *uniform* non-invariance conditions. However, this may be due to the fact that WLSMV_PD

Table 5 Model level type II error rates in threshold non-invariance conditions with six five-point indicators per group

			Symmetric thresholds		Asymmetric thresholds	
DIF type	Sample size	MAR rates	rFIML	WLSMV_PD	rFIML	WLSMV_PD
Small	500	0%	**0.926**	**0.914**	**0.932**	**0.892**
		30%	**0.946**	**0.752**	**0.966**	**0.762**
		50%	**0.962**	**0.490**	**0.976**	**0.516**
	1000	0%	**0.726**	**0.616**	**0.778**	**0.580**
		30%	**0.792**	**0.282**	**0.868**	**0.282**
		50%	**0.892**	**0.046**	**0.930**	0.070
	2000	0%	**0.202**	0.074	**0.330**	0.106
		30%	**0.402**	0.012	**0.548**	0.012
		50%	**0.644**	0.000	**0.758**	0.002
Large	500	0%	**0.234**	0.128	**0.366**	0.156
		30%	**0.412**	0.082	**0.522**	0.106
		50%	**0.638**	0.036	**0.734**	0.064
	1000	0%	0.002	0.000	0.022	0.004
		30%	0.042	0.000	0.090	0.004
		50%	**0.208**	0.000	**0.322**	0.004
	2000	0%	0.000	0.000	0.000	0.000
		30%	0.000	0.000	0.000	0.000
		50%	0.002	0.000	0.012	0.000
Mixed	500	0%	**0.402**	**0.284**	**0.534**	**0.346**
		30%	**0.550**	0.186	**0.660**	**0.216**
		50%	**0.750**	0.102	**0.798**	**0.120**
	1000	0%	0.086	0.026	**0.146**	0.040
		30%	**0.208**	0.006	**0.316**	0.014
		50%	**0.430**	0.000	**0.574**	0.002
	2000	0%	0.000	0.000	0.002	0.000
		30%	0.006	0.000	0.016	0.000
		50%	0.064	0.000	0.134	0.000
Non- uniform	500	0%	**0.912**	**0.828**	**0.942**	**0.844**
		30%	**0.948**	**0.896**	**0.960**	**0.930**
		50%	**0.974**	**0.972**	**0.980**	**0.988**
	1000	0%	**0.704**	**0.398**	**0.768**	**0.528**
		30%	**0.818**	**0.838**	**0.872**	**0.848**
		50%	**0.928**	**0.986**	**0.944**	**0.978**
	2000	0%	0.194	0.042	**0.364**	0.098
		30%	**0.442**	**0.654**	**0.572**	**0.692**
		50%	**0.682**	**0.974**	**0.768**	**0.976**

Note: Model level type II error rates above 0.2 are highlighted in bold

had inflated power to detect uniform non-invariance given its inflated type I error rates. Thus, researchers need to be cautious about using WLSMV_PD even under uniform non-invariance. rFIML had lower power to detect uniform non-invariance. One possible explanation is that continuous estimators are in general less sensitive to item level features of ordinal indicators (Oort, 1998). Therefore rFIML requires a larger sample to precisely detect differences between ordinal items. This could also explain why the type I error rates from rFIML were conservative. Similar findings have been reported in previous research (e.g., Chen et al., 2019).

Fourth, rFIML showed higher perfect recovery rates and power than WLSMV_PD in detecting *non-uniform* non-invariance (i.e., directions of non-invariance are different across non-invariant items) when missing data present.

Table 6 Model level type I error rates in invariance conditions with three-point indicators

Model Size	DIF model	Sample size	MAR rates	Symmetric thresholds		Asymmetric thresholds	
				rFIML	WLSMV_PD	rFIML	WLSMV_PD
6 Items	Metric	500	0%	0.080	**0.140**	0.016	0.030
			30%	0.080	**0.160**	0.012	0.046
			50%	0.060	**0.160**	0.026	**0.140**
		1000	0%	0.020	0.020	0.026	0.032
			30%	0.000	0.020	0.022	0.060
			50%	0.000	**0.260**	0.030	**0.284**
		2000	0%	0.020	0.040	0.022	0.038
			30%	0.020	0.040	0.024	**0.118**
			50%	0.020	**0.480**	0.026	**0.656**
	Scalar	500	0%	0.020	0.040	0.024	0.048
			30%	0.000	0.040	0.024	0.074
			50%	0.020	**0.240**	0.022	**0.226**
		1000	0%	0.040	0.060	0.018	0.044
			30%	0.040	0.080	0.018	**0.122**
			50%	0.020	**0.420**	0.024	**0.466**
		2000	0%	0.040	**0.120**	0.016	0.032
			30%	0.040	**0.180**	0.018	**0.192**
			50%	0.040	**0.760**	0.016	**0.816**
12 Items	Metric	500	0%	0.040	0.030	0.032	0.030
			30%	0.050	0.040	0.030	0.044
			50%	0.052	**0.122**	0.046	**0.122**
		1000	0%	0.026	0.032	0.012	0.024
			30%	0.030	0.064	0.020	0.064
			50%	0.022	**0.254**	0.020	**0.324**
		2000	0%	0.034	0.040	0.018	0.044
			30%	0.042	**0.118**	0.020	**0.118**
			50%	0.032	**0.618**	0.018	**0.760**
	Scalar	500	0%	0.034	0.040	0.044	0.020
			30%	0.062	0.060	0.064	0.064
			50%	0.052	**0.144**	0.048	**0.210**
		1000	0%	0.034	0.036	0.048	0.048
			30%	0.042	0.074	0.036	0.094
			50%	0.038	**0.410**	0.038	**0.498**
		2000	0%	0.028	0.044	0.044	0.058
			30%	0.034	**0.140**	0.032	**0.178**
			50%	0.038	**0.850**	0.030	**0.912**

Note: the type I error rates that are above 0.1 are highlighted in bold

Similar findings have been obtained in previous research using complete continuous data. For example, Meade and Lautenschlager (2004) and Whittaker and Khojasteh (2013) found that the ML estimator for complete data had higher power to detect non-uniform non-invariance than uniform non-invariance. They believed that this had to do with communalities of non-invariance items (Whittaker & Khojasteh, 2013). Given how non-invariance was simulated in those studies (e.g., higher loadings), non-uniform non-invariance may have created a scenario where the non-invariant item(s) had a higher communality for the focal group, increasing the influence of these non-invariant items and consequently the power of rFIML. Higher power combined with lower type I error rates resulted in higher perfect recovery rates of rFIML.

These non-invariant items with high communalities, however, did not benefit but actually harm the performance of

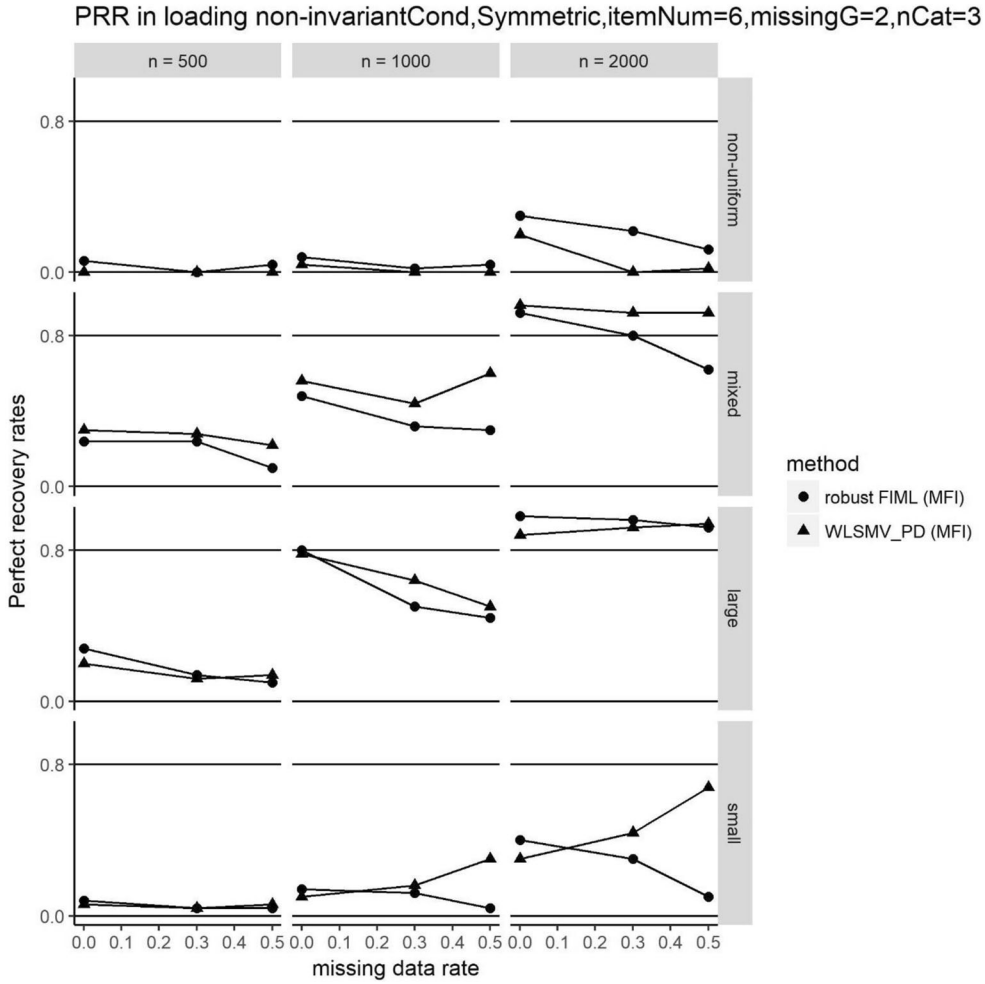


Fig. 6 Perfect recovery rates of the methods in loading non-invariance conditions with six symmetric three-point indicators per group in the model

WLSMV_PD, because impacts of missing data were magnified by PD in this scenario. A similar result was reported in Enders and Bandalos (2001) which showed that when missing data in SEM were handled using PD, indicators with higher communalities demonstrated more biased loading estimates. This could explain why perfect recovery rates of WLSMV_PD were much lower than rFIML under non-uniform non-invariance.

Other findings and issues

In addition to the general patterns described above, we found two interesting exceptions. First, in the threshold non-uniform non-invariance conditions, rFIML produced lower perfect recovery rates when there were only three categories per indicator, the model size was large, and thresholds were asymmetric. This implies that even with non-uniform non-invariance, there may be a tipping point where the advantage of rFIML in dealing with missing data would be outweighed by its disadvantage of dealing with the discrete nature of the ordinal data if the data are highly discrete and asymmetric.

Second, despite the general advantages of WLSMV_PD over rFIML in uniform non-invariance conditions, we found that if non-invariance appeared in loadings of items with asymmetric thresholds, WLSMV_PD was inferior to rFIML regardless of the non-invariance pattern. Similar findings were reported in Yoon & Kim (2014). This implies that WLSMV_PD may have more difficulty than rFIML in separating the source of non-invariance from the asymmetry of thresholds.

Practical recommendations

As we expected, neither of the methods is perfect. Thus, there is no one-size-fits-all recommendations, and preference for a method should be determined based on which criterion is deemed most important as well as characteristics of the data such as the pattern of non-invariance. Because type I errors are generally viewed as more harmful than type II errors, it is often more important to control type I error rates than minimize type II error rates. Thus, if controlling type I error rates is the priority, rFIML is generally preferred to WLSMV_PD.

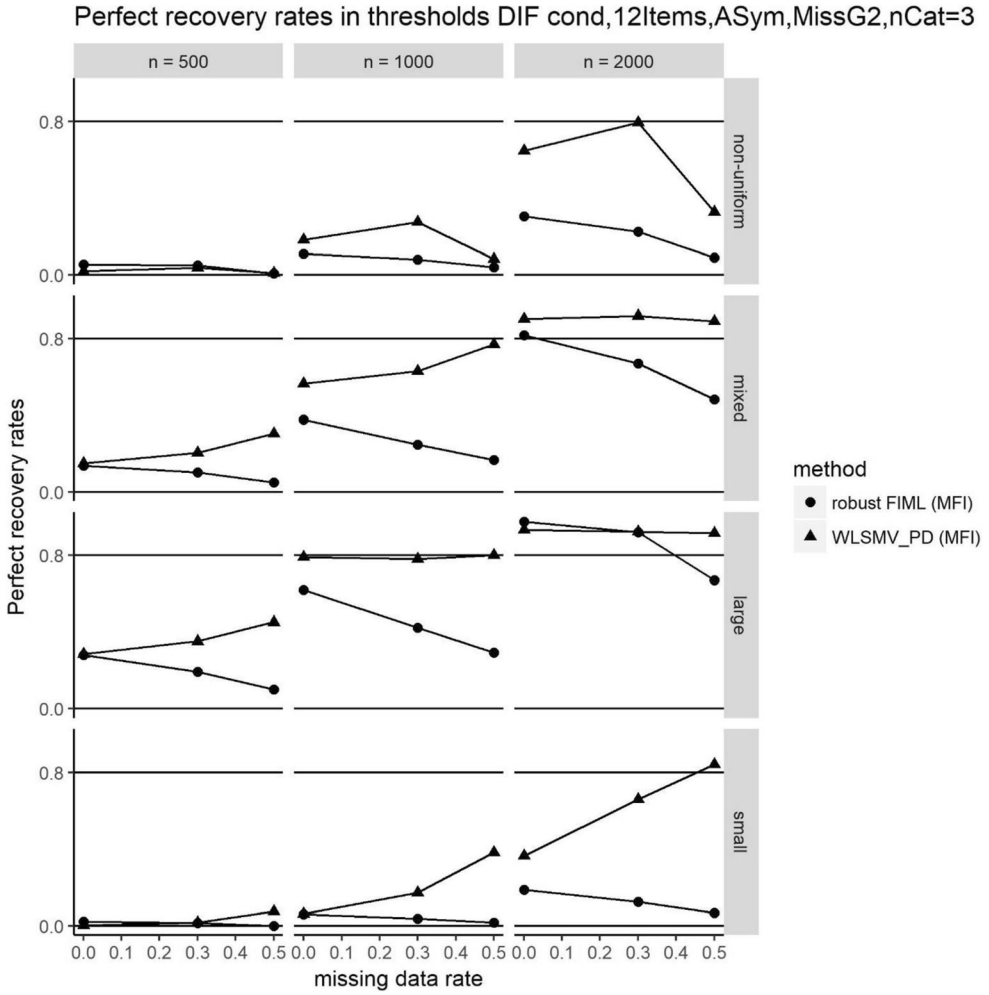


Fig. 7 Perfect recovery rates of the methods in thresholds non-invariance conditions with 12 asymmetric three-point indicators per group in the model

Using rFIML will not only ensure a good control of model level type I error rates but also provide higher power to detect non-invariant items if the pattern of non-invariance is non-uniform, unless there are less than five categories, the model size is large, and thresholds are asymmetric. On the other hand, if maximizing the perfect recovery rate is the goal, then WLSMV_PD appears to be more favorable when non-invariance is uniform. Otherwise, rFIML is still preferred. One downside of using rFIML is that it will have lower power to detect uniform non-invariant items than WLSMV_PD when thresholds are symmetric. According to our study, rFIML seemed to require at least 1000 and 2000 participants to demonstrate sufficient perfect recovery rates in conditions with five-point indicators and three-point indicators, respectively. These sample sizes could be unrealistic for some studies.

In addition to the above recommendations, we want to emphasize the importance of theory when using MFI to build partial invariance models. Many researchers have cautioned that using MFI as a purely data-driven, post hoc model modification index could lead to misleading results (e.g., MacCallum, Roznowski, & Necowitz, 1992; Yoon & Millsap, 2007). Thus, theoretical justification and cross-validation for invariance model modifications are important issues that should be considered during the search process (Byrne et al., 1989; Yoon & Millsap, 2007).

Limitation and future directions

As in many simulation studies, we could not include all conditions and methods of interest in the current study. First, we assume that the anchor item is correctly specified, which may not always be the case in practice. Yoon & Millsap (2007) demonstrated that incorrect anchor variables could make invariance items appear to be non-invariant. Johnson, Meade & DuVernet (2009) also found that using wrong anchor items could substantively affect item level ME/I tests and make group comparisons unreliable. Although several methods to select anchor items have been proposed (e.g., Jung & Yoon, 2017), they have not been examined with missing data. Thus, this is a potential avenue for future research.

Second, as in most specification search research, we assumed that the latent factors in both groups were normally distributed. However, researchers have found that non-normal latent variables could affect the performance of WLSMV and the robust ML estimator for complete data (Savalei & Folk, 2014; Suh, 2015). It would be interesting to examine the joint effect of non-normal latent factors and ordinal missing data on the performance of the two approaches considered in our study.

Third, we only examined one of the search processes, SBSS_LMFI, because of its popularity in both methodological and empirical literature. However, SBSS_LMFI has its limitations. One main limitation is that it may not perform well when the percentage of non-invariance items was high (e.g., 2/3 of the items are non-invariant) (Yoon & Millsap, 2007). In this case, the restricted model used in the beginning of a backward search process (e.g., full-scale invariance model) would be severely misspecified (i.e., too many inappropriate equality constraints), leading to biased results (Yoon & Millsap, 2007).

This problem can be mitigated by using forward specification search methods because they start with a less restricted invariance model. Recently, Jung and Yoon (2016) proposed a forward specification search process that utilizes the CIs of differences between corresponding parameters across groups (e.g., loading differences of corresponding items in a configural invariance model). Using complete continuous data, they demonstrated that their method was at least as good as or even slightly better than SBSS_LMFI. We did not include this approach in our study for two reasons: (1) this approach has not been widely adopted by empirical researchers, and (2) It had very similar performance to SBSS_LMFI in simulations (see Jung & Yoon, 2016).

Several other advances in specification search would also deserve researchers' attention. There are specification methods developed based on penalized ML estimators (e.g., Belzak, & Bauer, 2020; Huang, 2018; Jacobucci, Grimm, & McArdle, 2016; Liang, & Jacobucci, 2019). Emerging evidence showed that these methods could provide a better control on type I error rates than traditional methods such as item response theory (e.g., Belzak, & Bauer, 2020). In addition, with some recently developed software, researchers could use two-stage FIML to handle missing data problems with penalized ML estimators in a framework of multiple group analyses (e.g., the lslx package in R , Huang, 2020). There are also specification search methods developed using Bayesian approaches. For instance, Shi, Song, Liao, et al., (2017b) showed the possibility of using Bayesian SEM to preform forward specification search with complete continuous data. Given that Bayesian estimators are capable of handling ordinal missing data and advantageous in dealing with small sample sizes (e.g., Muthén, et al., 2015; McNeish, 2016), we believe that Bayesian SEM is a promising method to solve the limitations shared by the frequentist approaches. These new methods certainly warrant further research under ordinal missing data.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling: Issues and techniques*,*243*, 277.

Asparouhov, T., & Muthén, B. O. (2010). Multiple imputation with Mplus.*MPlus Web Notes*. Retrieved from https://www.statmodel.com/download/Imputations7.pdf

Beam, C. R., Marcus, K., Turkheimer, E., & Emery, R. E. (2018). Gender differences in the structure of marital quality. *Behavior genetics*, *48*(3), 209–223. https://doi.org/10.1007/s10519-018-9892-4

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*. https://doi.org/10.1037/met0000253

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

Bou Malham, P., & Saucier, G. (2014). Measurement invariance of social axioms in 23 countries. *Journal of Cross-Cultural Psychology*, *45*(7), 1046–1060. https://doi.org/10.1177/0022022114534771

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456. https://doi.org/10.1037/0033-2909.105.3.456

Chan, M. H. M., Gerhardt, M., & Feng, X. (2019). Measurement invariance across age groups and over 20 years' time of the Negative and Positive Affect Scale (NAPAS). *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000529

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, *95*(5), 1005. https://doi.org/10.1037/e514412014-064

Chen, P. Y., 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*, 1–15. https://doi.org/10.1080/00273171.2019.1608799

Chou, C. P., & Bentler, P. M. (2002). Model modification in structural equation modeling by imposing constraints. *Computational statistics & data analysis*, *41*(2), 271–287

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Erreygers, S., Vandebosch, H., Vranjes, I., Baillien, E., & De Witte, H. (2018). Development of a measure of adolescents' online prosocial behavior. *Journal of Children and Media*, *12*(4), 448–464. https://doi.org/10.1080/17482798.2018.1431558

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457. https://doi.org/10.1207/s15328007sem0803_5

Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, *18*(1), 35–54.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466. https://doi.org/10.1037/1082-989x.9.4.466

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520. https://doi.org/10.1037/a0031669

French, B. F., & Finch, H. (2016). Factorial invariance testing under different levels of partial loading invariance within a multiple group confirmatory factor analysis model. *Journal of Modern Applied Statistical Methods*, *15*(1), 26. https://doi.org/10.22237/jmasm/1462076700.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical care*, *44*(11 Suppl 3), S78. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100. https://doi.org/10.1207/s15328007sem1001_4

Hakkarainen, A. M., Holopainen, L. K., & Savolainen, H. K. (2016). The impact of learning difficulties and socioemotional and behavioural problems on transition to postsecondary education or work life in Finland: a five-year follow-up study. *European Journal of Special Needs Education*, *31*(2), 171–186

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499–522. https://doi.org/10.1111/bmsp.12130

Huang, P. H. (2020) lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v093.i07

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, *16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567–584. https://doi.org/10.1080/10705511.2015.1138092

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 65–79. https://doi.org/10.1080/10705511.2016.1251845

Kim, G., Wang, S. Y., & Sellbom, M. (2018). Measurement Equivalence of the Subjective Well-Being Scale Among Racially/Ethnically Diverse Older Adults. *The Journals of Gerontology: Series B*. https://doi.org/10.1093/geronb/gby110

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093

Liang, X., & Jacobucci, R. (2019). Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2019.1693273

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing Measurement Invariance in Longitudinal Data With Ordered-Categorical Measures. *Psychological Methods*. *22*(3), 486–506. https://doi.org/10.1037/met0000075

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490. https://doi.org/10.1037//0033-2909.111.3.490

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. https://doi.org/10.1080/10705510701575461

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*(1), 60–72. https://doi.org/10.1207/s15328007sem1101_5

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological methods*, *9*(1), 93. https://doi.org/10.1037/1082-989x.9.1.93

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/bf02294210

Muthén, B. O. (2017, Feb 26). Multiple imputation [Online forum comment]. Message posted www.statmodel.com/discussion/ messages/22 /381.html?1488130113

Muthén, L.K (2011, Jul, 26) Modification indices. [Online forum comment]. Message posted http://www.statmodel.com/discussion/messages/9/153.html?1457176945

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, *4*(5), 1–22. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B.O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes Unpublished Technical Report. Retrieved from https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K. & Muthén, B. O. (1998–2017). Mplus user's guide. Eighth Edition. Los Angeles, CA: Author.

Muthén, B.O, Muthén, L.K & Asparouhov, T. (2015). Estimator choices with categorical outcomes. Mplus Web Notes: March 2015. Retrieved from https://www.statmodel.com/download/EstimatorChoices.pdf

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966. https://doi.org/10.1037/a0022955

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/bf02296207

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124. https://doi.org/10.1080/10705519809540095

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*(1), 149–160. https://doi.org/10.1080/10705511.2013.824793

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. https://doi.org/10.1207/s15328007sem1202_1

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling, 21*(2), 280–302. https://doi.org/10.1080/10705511.2014.882692

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*(3), 371–384. https://doi.org/10.1007/BF02294623

Sommer, K., et al.. (2019). Consistency matters: measurement invariance of the EORTC QLQ-C30 questionnaire in patients with hematologic malignancies. *Quality of Life Research*, 1–9. https://doi.org/10.1007/s11136-019-02369-5

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Shi, D., Song, H., & Lewis, M. D. (2017a). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, https://doi.org/10.1177/1073191117711020

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017b). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, 52(4), 430–444. https://doi.org/10.1080/00273171.2017.1306432

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology, 9,* 1–12. https://doi.org/10.1027/1614-2241/a000049

Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling, 22*(4), 568–580. https://doi.org/10.1080/10705511.2014.937669

Teman, E.D. (2012). *The performance of multiple imputation and full information maximum likelihood for missing ordinal data in structural equation models*. Ann Arbor, MI: ProQuest.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70. https://doi.org/10.1177/109442810031002

Willoughby, M. T., Pek, J., Greenberg, M. T., & Family Life Project Investigators. (2012). Parent-reported attention deficit/hyperactivity symptomatology in preschool-aged children: Factor structure, developmental change, and early risk factors. *Journal of abnormal child psychology*, *40*(8), 1301–1312. https://doi.org/10.1007/s10802-012-9641-8

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58. https://doi.org/10.1037/1082-989x.12.1.58

Whittaker, T. A., & Khojasteh, J. (2013). A comparison of methods to detect invariant reference indicators in structural equation modelling. *International Journal of Quantitative Research in Education*, *1*(4), 426–443. https://doi.org/10.1504/ijqre.2013.058310

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484–503. https://doi.org/10.1080/00273171.2015.1022644

Xu, Y, Green, S.B. (2016). The impact of varying the number of measurement invariance constrains on the assessment of between group differences of latent means. *Structural equation modeling, 23*(2), 290–301. https://doi.org/10.1080/10705511.2015.1047932

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior research methods*, *46*(4), 1199–1206. https://doi.org/10.3758/s13428-013-0430-2

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463. https://doi.org/10.1080/10705510701301677