

Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares

Cheng-Hsien Li¹

Published online: 15 July 2015
© Psychonomic Society, Inc. 2015

Abstract In confirmatory factor analysis (CFA), the use of maximum likelihood (ML) assumes that the observed indicators follow a continuous and multivariate normal distribution, which is not appropriate for ordinal observed variables. Robust ML (MLR) has been introduced into CFA models when this normality assumption is slightly or moderately violated. Diagonally weighted least squares (WLSMV), on the other hand, is specifically designed for ordinal data. Although WLSMV makes no distributional assumptions about the *observed* variables, a normal *latent* distribution underlying each observed categorical variable is instead assumed. A Monte Carlo simulation was carried out to compare the effects of different configurations of latent response distributions, numbers of categories, and sample sizes on model parameter estimates, standard errors, and chi-square test statistics in a correlated two-factor model. The results showed that WLSMV was less biased and more accurate than MLR in estimating the factor loadings across nearly every condition. However, WLSMV yielded moderate overestimation of the interfactor correlations when the sample size was small or/and when the latent distributions were moderately nonnormal. With respect to standard error estimates of the factor loadings and the interfactor correlations, MLR outperformed WLSMV when the latent distributions were nonnormal with a small sample size of $N = 200$. Finally, the proposed model tended to be over-rejected by chi-square test statistics under both MLR and WLSMV in the condition of small sample size $N = 200$.

Keywords Robust estimation · Ordinal data · Monte Carlo Simulation · Confirmatory factor analysis

In the social and behavioral sciences, researchers often employ Likert-type scale items to operationalize unobserved constructs (e.g., optimism or motivation) by using more manageably observed variables. Confirmatory factor analysis (CFA) has been widely used as evidence of construct validity in theory-based instrument construction. A confirmatory factor-analytic model takes into account the differences between the true and observed scores by including pertinent error variances as model parameters in a structural equation modeling framework. The most common method used to estimate parameters in CFA models is maximum likelihood (ML), because of its attractive statistical properties (i.e., asymptotic unbiasedness, normality, consistency, and maximal efficiency). The use of ML requires the assumption that the observed variables follow a continuous and multivariate normal distribution given the covariates in the population (Bollen, 1989; Jöreskog, 1969; Satorra, 1990). As a consequence, each pair of observed variables is bivariate normally distributed in each subpopulation with the same values on the covariates (Raykov & Marcoulides, 2006). When this assumption is considered tenable, ML maximizes the likelihood of the observed data to obtain parameter estimates. This is equivalent to minimizing the discrepancy function F_{ML} (Bollen, 1989; Jöreskog, 1969):

$$F_{ML} = \ln \left| \sum (\theta) \right| + \text{trace} [S \Sigma^{-1}(\theta)] - \ln |S| - p, \quad (1)$$

where θ is the vector of model parameters, $\Sigma(\theta)$ is the model-implied covariance matrix, S is the sample covariance matrix, and p is the total number of observed variables in the model.

However, ML is not, strictly speaking, appropriate for ordinal variables. The normality assumption about observed

✉ Cheng-Hsien Li
Cheng.Hsien.Li@uth.tmc.edu

¹ Department of Pediatrics, Children's Learning Institute, University of Texas Health Science Center at Houston, Houston, TX, USA

variables is severely violated when the analyzed data have only a few response categories (Lubke & Muthén, 2004). When the normality assumption is not deemed empirically tenable, the use of ML may not only reduce the precision and accuracy of the model parameter estimates, but may also result in misleading conclusions drawn from empirical data. In previous simulation studies that have applied ML in cases of severe nonnormality due to categorization, researchers have found that chi-square statistics may be inflated, factor loadings may be downward-biased, and standard errors may be biased to some degree, as well (Beauducel & Herzberg, 2006; Kaplan, 2009; Muthén & Kaplan, 1985, 1992).

The existing estimators with statistical corrections to standard errors and chi-square statistics, such as robust maximum likelihood (robust ML: MLR in *Mplus*) and diagonally weighted least squares (DWLS in LISREL; WLSMV or robust WLS in *Mplus*), have been suggested to be superior to ML when ordinal data are analyzed. Robust ML has been widely introduced into CFA models when continuous observed variables slightly or moderately deviate from normality. WLSMV, on the other hand, is specifically designed for categorical observed data (e.g., binary or ordinal) in which neither the normality assumption nor the continuity property is considered plausible. Although WLSMV makes no distributional assumptions about *observed* variables, a normal *latent* distribution underlying each observed categorical variable is instead assumed. Application of the standard WLS estimator is not investigated in this article, mainly because (i) empirical research has shown that the performance of standard WLS is inferior to that of WLSMV when the sample size is small and the number of observed variables is relatively large, and (ii) standard WLS encounters more computational problems in the process of estimating model parameters than does WLSMV (Flora & Curran, 2004; Muthén, du Toit, & Spisic, 1997; Oranje, 2003).

Robust maximum likelihood

As compared to ML estimation, a robust ML approach is less dependent on the assumption of multivariate normal distribution. When the normality assumption about observed variables does not hold, and robust ML is implemented, parameter estimates are still obtained using the asymptotically unbiased ML estimator, but standard errors and chi-square test statistics are statistically corrected to enhance the robustness of ML against departures from normality (in the forms of skewness, kurtosis, or both). Specifically, the corrected standard error estimates are obtained by a sandwich-type estimator, rather than the inverse Fisher information matrix (Kaplan, 2009; Muthén & Muthén, 2007). The sandwich estimator implemented in MLR incorporates an observed Fisher information matrix $\hat{\Delta}'\hat{\mathbf{I}}_{\text{ob}}\hat{\Delta}$ (approximation to the Fisher information

matrix) into the asymptotic covariance matrix of the estimated parameter vector $\hat{\theta}$ (Muthén & Muthén, 2007; Satorra & Bentler, 1994; Savalei, 2010):

$$\text{aCov}(\hat{\theta})_{\text{MLR}} = N^{-1}(\hat{\Delta}'\hat{\mathbf{I}}_{\text{ob}}\hat{\Delta})^{-1}\hat{\Delta}'\hat{\mathbf{I}}_{\text{ob}}\hat{\Gamma}\hat{\mathbf{I}}_{\text{ob}}\hat{\Delta}(\hat{\Delta}'\hat{\mathbf{I}}_{\text{ob}}\hat{\Delta})^{-1} \quad (2.1)$$

and

$$\hat{\mathbf{I}}_{\text{ob}} = \mathbf{D}'\left\{\Sigma^{-1}(\hat{\theta}) \otimes \left[\left(\Sigma^{-1}(\hat{\theta})\right)\Sigma\Sigma^{-1}(\hat{\theta})^{-1}/_2\Sigma^{-1}(\hat{\theta})\right]\right\}\mathbf{D}, \quad (2.2)$$

where $\hat{\Gamma}$ is the estimated asymptotic covariance matrix of \mathbf{S} , $\hat{\Delta}$ is defined as $\partial\Sigma(\hat{\theta})/\partial\hat{\theta}$ [i.e., model first derivatives evaluated at parameter estimates $\hat{\theta}$], the “duplication” matrix \mathbf{D} is of order $r^2 \times \frac{1}{2}r(r+1)$ [r = the number of observed variables in $\Sigma(\Theta)$; see Magnus & Neudecker, 1986, p. 172], and \otimes denotes a Kronecker product. The corrected standard error estimates are calculated by taking the square roots of the diagonal elements of the above estimated asymptotic covariance matrix. The *upward* corrected standard error estimates have been found to be more favorable than those estimated by the inverse Fisher information matrix when the observed data are nonnormal (Satorra & Bentler, 1994).

The robust corrections applied to the chi-square statistic vary slightly across different current software programs. The Satorra–Bentler scaled chi-square statistic given by the “ML, Robust” estimator in EQS is equivalent to the mean-adjusted chi-square statistic obtained by MLM in *Mplus*. Another corrected chi-square statistic T_2^* , proposed by Yuan and Bentler (1997, 1998) using the generalized least squares approach, is asymptotically equivalent to the chi-square test statistic obtained by MLR (Muthén & Muthén, 2007):

$$T_{\text{MLR}} \xrightarrow{a} T_2^*, \quad (3)$$

where \xrightarrow{a} denotes asymptotic equivalence. A simulation study by Yuan and Bentler (1998) has shown that both the Satorra–Bentler scaled χ^2 and the Yuan–Bentler T_2^* are robust against nonnormal distributions of observed data. Note that a mean- and variance-adjusted chi-square statistic (i.e., MLMV in *Mplus*, also known as the Satorra–Bentler adjusted chi-square statistic) is also available in software programs but is outside the scope of this study. MLR is much more frequently used than MLM and MLMV in research practice.

Diagonally weighted least squares

Weighted least squares is generally referred to as the asymptotically distribution-free estimator when data are continuous but nonnormal and a consistent estimate of the asymptotic covariance matrix of sample-based variances and covariances is used (Browne, 1984). However, neither the assumption of normality nor the continuity property is clearly met by

observed variables that are measured on an ordinal scale. Muthén (1984) made a substantial breakthrough in analyzing ordinal observed data in CFA models by using a weighted least squares approach. In this approach, WLS assumes that a continuous, normal, latent response distribution x^* underlies an observed ordinal variable x in the population:

$$x = m, \text{ if } \tau_{m-1} < x^* < \tau_m, \quad (4)$$

where m ($=1, 2, \dots, c$) defines the observed value of an ordinal observed variable x , τ is the threshold ($-\infty = \tau_0 < \tau_1 < \tau_2 \dots < \tau_{c-1} < \tau_c = +\infty$), and c is the number of categories. The thresholds and polychoric correlations are first estimated using two-step ML estimation through bivariate contingency tables (Bollen, 1989; Jöreskog, 2005; Olsson, 1979). An estimated polychoric correlation captures the linear relationship between two normal, latent response variables. Parameter estimates and the associated standard errors are then obtained using the estimated asymptotic covariance matrix of the polychoric correlation and threshold estimates (denoted $\tilde{\mathbf{V}}$) in a weight matrix \mathbf{W} to minimize the weighted least squares fit function F_{WLS} (Muthén 1984):

$$F_{\text{WLS}} = [s - \sigma(\theta)]' \mathbf{W}^{-1} [s - \sigma(\theta)], \quad (5)$$

where θ is the vector of model parameters, \mathbf{W} ($= \tilde{\mathbf{V}}$) is the weight matrix, $\sigma(\theta)$ is the model-implied vector containing the nonredundant, vectorized elements of $\Sigma(\theta)$, and s is the vector containing the unique elements of sample statistics (i.e., threshold and polychoric correlation estimates). When the weight matrix \mathbf{W} is replaced with the identity matrix \mathbf{I} , WLS reduces to unweighted least squares (ULS). In order to address heteroscedastic disturbances in CFA models, a full weight matrix $\mathbf{W} = \tilde{\mathbf{V}}$ (i.e., the estimated asymptotic covariance matrix of the polychoric correlation and threshold estimates) is implemented in the WLS fit function above to account for distributional variability in and interrelationships among the observed variables (Kaplan, 2009). However, as the number of observed variables and response categories increases, the weight matrix grows rapidly in size.

Weighted least squares with means and variances adjusted in *Mplus* (WLSMV; Muthén & Muthén, 2007), a mathematically simple form of the WLS estimator, only incorporates diagonal elements of the full weight matrix in the fit function. The diagonal weight matrix $\mathbf{W}_D = \text{diag}(\tilde{\mathbf{V}})$ is more flexible (i.e., need not be positive-definite) than the full weight matrix $\mathbf{W} = \tilde{\mathbf{V}}$ (Kaplan, 2009; Kline, 2011; Muthén et al., 1997). The diagonal weight matrix prevents software programs from engaging in extensive computations and encountering numerical problems in model estimation. The results of simulation studies have indicated the relative superiority of WLSMV over WLS in the analysis of measurement models with ordinal indicators (Flora & Curran, 2004; Kaplan, 2009; Muthén, 1993; Muthén et al., 1997). The WLSMV estimation proceeds

by first estimating thresholds and polychoric correlations using ML. The parameter estimates are then obtained from the estimated asymptotic variances of the polychoric correlation and threshold estimates used in a diagonal weight matrix (Muthén et al., 1997; Muthén & Muthén, 2007):

$$F_{\text{WLSMV}} = [s - \sigma(\theta)]' \mathbf{W}_D^{-1} [s - \sigma(\theta)], \quad (6)$$

where $\mathbf{W}_D = \text{diag}(\tilde{\mathbf{V}})$ is the diagonal weight matrix. Using the same sandwich-type matrix form as for MLR, the obtained standard error estimates are given by the square roots of the diagonals of the estimated asymptotic covariance matrix of the estimated parameter vector of $\hat{\theta}$ (Muthén et al., 1997):

$$\text{aCov}(\hat{\theta})_{\text{WLSMV}} = N^{-1} (\tilde{\Delta}' \mathbf{W}_D^{-1} \tilde{\Delta})^{-1} \tilde{\Delta}' \mathbf{W}_D^{-1} \tilde{\mathbf{V}} \mathbf{W}_D^{-1} \tilde{\Delta} (\tilde{\Delta}' \mathbf{W}_D^{-1} \tilde{\Delta})^{-1}, \quad (7)$$

where $\tilde{\mathbf{V}}$ is a consistent estimator of the asymptotic covariance matrix of s , \mathbf{W}_D [$= \text{diag}(\tilde{\mathbf{V}})$] contains only diagonal elements of the estimated asymptotic covariance matrix, and $\tilde{\Delta}$ is defined as $\partial \sigma(\hat{\theta}) / \partial \hat{\theta}$. In the meantime, $\tilde{\mathbf{V}}$ need not be inverted in computations of standard error estimates (Muthén, 1993; Muthén et al., 1997). A mean- and variance-adjusted chi-square test statistic with the degrees of freedom computed on the basis of a given model specification is defined as (Muthén et al., 1997):

$$T_{\text{WLSMV}} = [df' / \text{trace}(\tilde{\mathbf{U}} \tilde{\mathbf{V}})] T_{\text{WLS}}, \quad (8)$$

where df' is computed as the integer closest to $df^* = \{\text{trace}(\tilde{\mathbf{U}} \tilde{\mathbf{V}}) / \text{trace}(\tilde{\mathbf{U}} \tilde{\mathbf{V}})^2\}$, T_{WLS} is the standard WLS chi-square test statistic, $\tilde{\mathbf{V}}$ is the estimated asymptotic covariance matrix of s , and $\tilde{\mathbf{U}} = \mathbf{W}_D^{-1} - \mathbf{W}_D^{-1} \tilde{\Delta} (\tilde{\Delta}' \mathbf{W}_D^{-1} \tilde{\Delta})^{-1} \tilde{\Delta}' \mathbf{W}_D^{-1}$. It is worth noting that the aim of statistical corrections to standard errors in WLSMV is to compensate for the loss of efficiency when the full weight matrix is not calculated, and the mean and variance adjustments for test statistics in WLSMV are targeted to make the shapes of the test statistics be approximately close to the reference chi-square distribution with the associated degrees of freedom.

Previous simulation studies

Simulation studies have investigated the properties of different estimation methods, typically reporting on the relative performance (e.g., precision and accuracy) of parameter estimates, standard error estimates, and Type I error rates associated with chi-square statistics. A literature review of Monte Carlo simulation studies carrying out ordinal confirmatory factor-analytic models was conducted across several high-impact journals (e.g., *Psychological Methods*, *Structural Equation Modeling*, *Educational and Psychological Measurement*, and *Multivariate Behavioral Research*) over 20 years (1994–2013). The empirical findings, using ML

and WLS and their statistical corrections, can be briefly summarized below. Generally, the overall performance of WLS was inferior to that of WLSMV in the analysis of CFA models using ordinal variables across almost every condition investigated by Flora and Curran (2004). Factor loading estimates were less biased by WLS and WLSMV than by ML, but interfactor correlations were found to be less overestimated by ML than by WLS and WLSMV (Beauducel & Herzberg, 2006; DiStefano, 2002). In contrast, Yang-Wallentin, Jöreskog, and Luo (2010) gave empirical evidence that the parameter estimates (both factor loadings and interfactor correlations) obtained by WLS were substantially biased, whereas those obtained by WLSMV and ML were essentially unbiased, regardless of the number of categories (two, five, or seven) and the shape of the observed distributions (symmetrical vs. asymmetrical). In addition, Lei (2009) found that the relative bias in parameter estimates was generally negligible for both ML and WLSMV across different levels of asymmetric observed distributions (symmetric, mildly skewed, and moderately skewed). Oranje (2003) also concluded that both ML and WLSMV produced equally good parameter estimates across the numbers of categories (two, three, and five). Note that ML and MLR yield the same parameter estimates, but different standard error estimates and chi-square statistics. In addition, it is worth noting that the simulation studies of Lei (2009), Oranje (2003), and Yang-Wallentin et al. (2010) used a polychoric correlation matrix, instead of a sample-based covariance matrix, in ML with robust corrections to standard errors and chi-square statistics.

In terms of standard error estimates, ML has been found to produce much smaller standard errors of factor loadings than does mean-adjusted ML in LISREL, statistically equivalent to MLM in *Mplus*, (Yang-Wallentin et al., 2010) and WLSMV (Beauducel & Herzberg, 2006), indicating that uncorrected ML standard errors are generally underestimated. On the other hand, simulation studies have shown that standard errors in WLSMV were generally less biased than those obtained by mean-adjusted ML, irrespective of the number of categories (Yang-Wallentin et al., 2010) and the level of asymmetric observed distributions (Lei, 2009). As for chi-square statistics, Beauducel and Herzberg (2006) revealed that the unadjusted chi-square statistics produced by ML were more likely to over-reject the proposed models than were the mean- and variance-adjusted chi-square statistics obtained by WLSMV. Additionally, Lei found that WLSMV was slightly more powerful than mean-adjusted ML in the evaluation of the overall model fit across different levels of asymmetric observed distributions, whereas Oranje (2003) concluded that the mean-adjusted ML provided the most correct rejection rate than WLSMV across the numbers of response categories.

Present study

The present study was designed to advance scholarly understanding of the impact of ordinal observed variables on parameter estimates, the associated standard errors, and chi-square statistics for ordinal CFA models. Jackson, Gillaspay, and Purc-Stephenson (2009) reviewed 194 studies from 1998 to 2006 and found that the most commonly tested CFA models were correlated-factor models (50.5 %), followed by orthogonal (12.0 %), hierarchical (10.6 %), single-factor (9.5 %), and so on. A correlated two-factor model was chosen as the representative of the common/simple CFA model specification that is often-times examined in simulation studies and is frequently encountered in practice. The literature in ordinal CFA is abundant for the joint performance of robust estimators on both factor loading and interfactor correlation estimates (e.g., Lei, 2009; Oranje, 2003; Yang-Wallentin et al., 2010) or the performance of non-robust estimators (e.g., DiStefano, 2002; Forero & Maydeu-Olivares, 2009). However, little is currently known about the performance of different estimators with statistical corrections when examining factor loadings and interfactor correlations separately, along with corrected standard errors and chi-square test statistics, as well.

Second, MLR is not designed specifically for ordinal data, but one may assume that observed data are “approximately” continuous if the number of categories is sufficiently large. Johnson and Creech (1983) have noted that variability in parameter estimates is quite small with five or more response categories in the model. In practice, empirical researchers have suggested using MLR in ordinal CFA or CFA-based models (e.g., multiple-indicator multiple-cause models, or measurement invariance) when the number of response categories for each item was equal to or greater than five (e.g., Raykov, 2012; Rigdon, 1998, and the references therein). Yet, unlike other robust corrections implemented in ML estimation, MLR implemented in *Mplus* has not been systematically evaluated by means of a Monte Carlo study in previous studies, although its robust correction is similar, but not always equivalent, to other robust ML corrections (e.g., MLM or MLMV in *Mplus*; ML, ROBUST in EQS). On the other hand, WLSMV has been specifically proposed to deal with ordinal data (the default setting in *Mplus*), mainly because it makes no distributional assumptions about the observed variables. When it comes to a CFA model with ordinal data, applied researchers tend to choose one or another estimator to perform data analysis in *Mplus*. Some researchers prefer treating ordinal variables with more than five alternatives as if they were approximately continuous variables, and in turn they perform MLR in data analysis to adjust the violation of nonnormality, whereas others highly recommend using WLSMV as long as the observed variables are ordinally scaled. An examination of

the two estimators under varying empirical conditions is needed, since they are frequently used in research practice.

Third, several simulation studies have compared WLSMV to other estimators (e.g., ML, WLS, and ULS) across different numbers of categories: When the number of categories was five, six, or seven, (1) the performance of WLSMV estimation was slightly superior to the performance of ML (Beauducel & Herzberg, 2006); (2) WLSMV outperformed WLS across most conditions (Flora & Curran, 2004); and (3) ULS was found to produce more accurate and precise factor loadings than did WLSMV, but it encountered a higher rate of model nonconvergence (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). However, the impact of the number of categories on parameter variability and adjusted standard error estimates has not yet been examined using MLR and WLSMV with larger numbers of categories (e.g., eight or ten). The number of categories affects not only the distribution of the observed ordinal variables, but also the possibility of treating ordinal variables as approximately continuous. This extension can help validate whether MLR is equally as good as WLSMV in a CFA model when ordinal observed variables have more than five response categories, and/or it can explore whether the superiority of MLR over WLSMV can be reached with a larger number of categories.

Fourth, MLR was developed to permit the parameter estimation from nonnormality of continuous observed variables, whereas WLSMV has been implemented in CFA models with nonnormal observed data due to the categorical nature of measurement (i.e., ordinal data), conditional on the assumption of a continuous, normal underlying distribution in the population. Although polychoric correlation estimates have shown robustness against violation of the latent-normality assumption (Coenders, Satorra, & Saris, 1997; Flora & Curran, 2004; Micceri, 1989; Quiroga, 1992), what is not clearly known from the current literature is the extent to which the precision of WLSMV in estimating factor loadings, standard errors, and chi-square statistics would be sustained. Therefore, the latent distribution was manipulated by varying skewness and kurtosis in the study. It could be expected that the effect of the latent normality distribution violation would more likely be salient for the performance of WLSMV than for that of MLR, holding other experimental conditions equal. Moreover, nonnormality, in the form of asymmetry observed in psychometric measurements, has not been uncommon in applied studies. Micceri (1989) found that only about 3 % of the 125 observed distributions that he investigated were close to normal or near symmetric, and over 80 % displayed at least slight or moderate asymmetry. Therefore, in order to be more realistic from an applied standpoint, this study also included asymmetric observed distributions of ordinal variables in the simulation design.

Finally, this study was designed to examine the effect of sample size on the parameter estimates produced while

utilizing the two estimators, because researchers have noted that a desirable sample size is known to be an important factor in CFA models. A small sample may cause inaccurate parameter estimates and unstable standard errors, and may result in nonconvergence and improper solutions, as well.

Method

Model specification

A Monte Carlo simulation study was carried out to compare the effects of different configurations of latent response distributions, numbers of categories, and sample sizes on model parameter estimates, standard errors, and chi-square test statistics in a correlated two-factor model. Marsh, Hau, Balla, and Grayson (1998) concluded that the accuracy of parameter estimates appeared to be optimal when the number of observed variables per factor was four, and marginally improved as the number of observed variables increased. Therefore, each factor was measured by five ordinal observed indicators in the study. Two estimation procedures that are given by MLR and WLSMV in *Mplus* were used. For the first estimation procedure, ordinal observed indicators were treated as if they were approximately continuous variables in the data analysis. The parameter estimates, standard errors, and chi-square statistics were obtained using MLR. The data analysis for MLR was based on a sample-based covariance matrix. Regarding the second estimation procedure, ordinal observed indicators were specified as categorical variables in the data analysis. A polychoric correlation matrix and the asymptotic covariance matrix of the polychoric correlation and threshold estimates were used in WLSMV to obtain the parameter estimates, standard errors, and chi-square statistics.

Population model

Reported standardized factor loadings range from .4 to .9 in the majority of empirical research and simulation studies (DiStefano, 2002; Hoogland & Boomsma, 1998; Li, 2012; Paxton, Curran, Bollen, Kirby, & Chen, 2001). For the sake of facilitating interpretation, each factor loading was therefore held constant at .7, with its corresponding uniqueness automatically set to .51 under a standardized solution in the population model. The interfactor correlation was set to .3 in the population, reflecting a reasonable and empirical interfactor correlation value that has ranged from .2 to .4 in the applied literature and in simulation studies. The factor variances were all set equal to 1 in the population.

Latent response distributions

Two *latent* distributions that varied in skewness and kurtosis were used: (1) a slightly nonnormal latent distribution with skewness = 0.5 and kurtosis = 1.5, and (2) a moderately nonnormal latent distribution with skewness = 1.5 and kurtosis = 3.0. The desired levels of skewness and kurtosis for the two latent distributions were simply specified in the EQS program. For the slightly nonnormal latent distribution, the asymmetric distributions of observed variables with skewness ranged from .38 to .64, and for the moderately nonnormal latent distribution, the asymmetric distributions of observed variables with skewness ranged from 1.01 to 1.31. The response probabilities of the ordinal observed indicators used in the study are displayed in Fig. 1. Note that Fig. 1a to d represent slightly asymmetric observed distributions generated from a slightly nonnormal latent distribution; and Fig. 1e to h represent moderately asymmetric observed distributions generated from a moderately nonnormal latent distribution. In the slight-asymmetry condition, the response probabilities were 3 %, 58 %, 33 %, and 6 % for four categories; 4 %, 17 %, 47 %, 17 %, 8 %, and 7 % for six categories; 3 %, 6 %, 14 %, 36 %, 21 %, 10 %, 5 %, and 5 % for eight categories; and 3 %, 3 %, 7 %, 18 %, 30 %, 18 %, 7 %, 6 %, 4 %, and 4 % for ten categories. In the moderate-asymmetry condition, the response probabilities were 0 %, 67 %, 26 %, and 7 % for four categories; 0 %, 26 %, 46 %, 12 %, 7 %, and 9 % for six categories; 0 %, 0 %, 29 %, 36 %, 15 %, 8 %, 4 %, and 8 % for eight categories; and 0 %, 0 %, 9 %, 30 %, 28 %, 13 %, 5 %, 4 %, 3 %, and 8 % for ten categories.

Number of categories

In order to explore the impact of categorization, this study extended previous research by introducing two larger numbers

of categories (i.e., eight and ten). Four, six, eight, and ten categories were generated for each ordinal indicator within both the slightly and moderately nonnormal latent distributions. The chief goal here was to examine whether the general recommendation “more than five categories” is empirically tenable when using MLR and/or whether MLR can outperform WLSMV when a larger number of categories is used.

Sample size

A sufficient sample size is highly associated with the amount of model complexity (e.g., the number of observed variables). In order to estimate asymptotic covariance matrices, Jöreskog and Sörbom (1996, p. 171) recommended a minimum sample size requirement of $(p + 1)(p + 2)/2$, where p is the number of observed variables. A correlated two-factor CFA model with ten observed indicators in this study required a minimum sample size of 66. However, some researchers have suggested a large sample size of 150 for a simple CFA model with normal observed variables, and an even larger sample size of 265 for a CFA model with nonnormal observed variables (Muthén & Muthén, 2002). Jackson, Gillaspay, and Purc-Stephenson (2009) systematically reviewed 101 studies in CFA from 1998 to 2006 and reported that the median sample size was 389. Three different sample sizes commonly encountered in empirical investigations were employed in this study: $N = 200$, 500, and 1,000. In the case of a correlated two-factor model with ten observed indicators, a sample size of $N = 200$ is considered small (typically a benchmark in structural equation models), a sample size of $N = 500$ as medium, and a sample size of $N = 1,000$ as large.

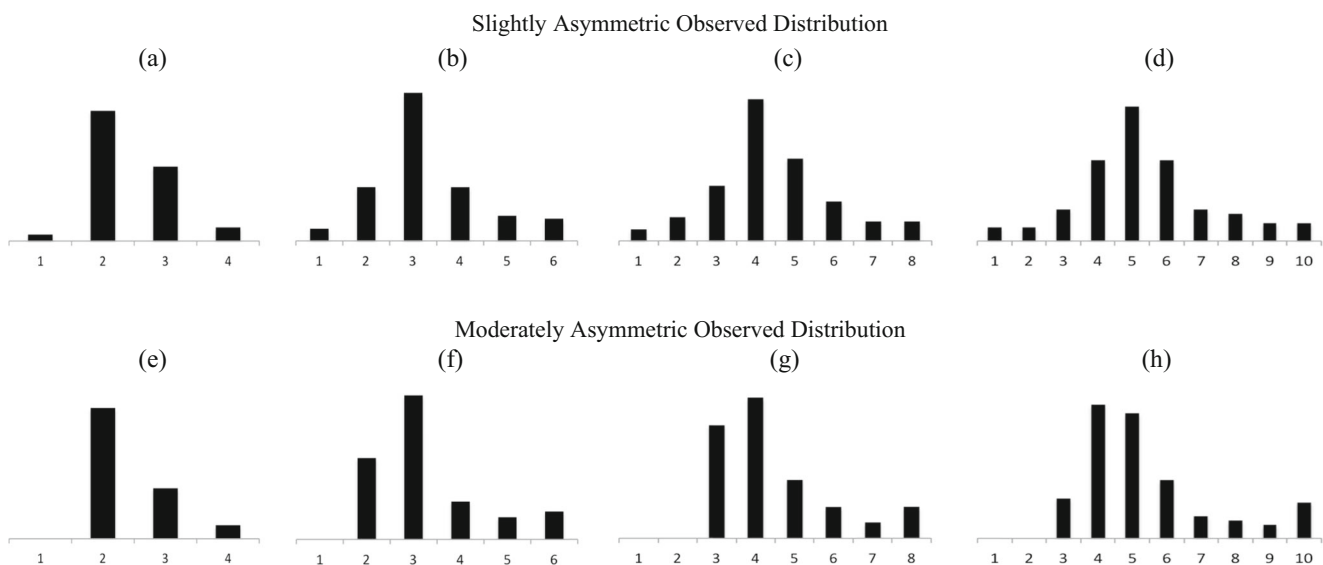


Fig. 1 Response probabilities of ordinal observed indicators

Data generation and analysis

A total of 2 (latent distributions) $\times 4$ (number of categories) $\times 3$ (sample size) = 24 experimental conditions were created in the study. Five hundred data sets were generated per experimental condition using EQS 6 (Bentler, 2006), resulting in a total of 12,000 data sets. The choice of 500 replications was made with consideration to sampling variance reduction, adequate power, and practical manageability (Muthén, 2002). Model parameters, standard errors, and the chi-square statistics were estimated for each replication using both MLR and WLSMV in *Mplus* 5.2 (Muthén & Muthén, 2007).

Outcome variables

Four outcome variables were examined in this study: rates of improper solutions or nonconvergence, parameter estimates (i.e., factor loadings and interfactor correlations, respectively), standard errors, and chi-square test statistics. The rate of improper solutions or nonconvergence for each experimental condition was defined as the proportions of replications for which the proposed model had a nonconverged solution or a solution that converged but had estimated interfactor correlations greater than 1 or negative residual variances under the *Mplus* default setting. For parameter estimates, the average relative bias (ARB), the average root mean squared error (ARMSE), and the coverage of 95 % confidence intervals were studied across experimental conditions. Standard error estimates obtained by the two estimators were compared using ARB and ARMSE. The model rejection rates associated with the chi-square test statistic were calculated at an alpha level of .05.

The difference between the estimated and true values of each parameter (i.e., the bias) was used to evaluate the performance of the two different estimators. Since bias is highly dependent on the magnitude of the true parameter value, and a great number of parameter estimates and standard errors were involved in each experimental condition, ARB and ARMSE were calculated. An ARB value less than 5 % was interpreted as a *trivial* bias, between 5 % and 10 % as a *moderate* bias, and greater than 10 % as a *substantial* bias (Curran, West, & Finch, 1996). Note that ARB was interpreted with caution, since it is used to describe an “overall” picture of average bias—that is, summing up bias in a positive and a negative direction together. A smaller absolute value of ARB indicates more accurate parameter estimates “on average.”

The root mean squared error can be regarded as a measure of the overall estimation quality, since it accounts for both the amount of bias and the sampling variability of estimates; this value was also averaged across 500 replications (i.e., ARME). A smaller ARMSE value is suggested as being favorable, reflecting better overall quality of the estimates. The 95 %

confidence intervals were formed for each replication using parameter estimates and the associated standard errors.

Confidence interval coverage was determined as the percentage of confidence intervals containing the true parameter. A lower rate of coverage (i.e., below 95 %) would indicate poor recovery of the true parameters, mainly due to a higher degree of bias in parameter estimates, the underestimation of standard errors, or a combination of both. The rate of rejection of the proposed model should approximate 5 %, because a nominal alpha level of .05 was specified in the population model. A higher rate of rejection (i.e., over 5 %) suggests inflated Type I error rates, reflecting that the chi-square test statistics may have been underestimated; a lower rate of rejection indicates that the chi-square statistics may have been overadjusted.

Results

Rates of improper solutions or nonconvergence

The rates of improper solutions and nonconvergence across the 24 experimental conditions were 0 % for both MLR and WLSMV, irrespective of the number of categories (four, six, eight, and ten), level of latent distribution violations (slightly and moderately nonnormal), and sample sizes ($N = 200, 500,$ and $1,000$). In summary, the problems of improper solutions or nonconvergence did not occur for MLR or WLSMV, even when ordinal asymmetric data generated from a moderately nonnormal latent distribution and a small sample size of $N = 200$ were analyzed in a correlated two-factor model.

Parameter estimates

Factor loadings The ARB and ARMSE values for factor loadings are presented in Table 1. Inspecting Table 1, the factor loadings were, on average, underestimated by MLR. In particular, they were *substantially* underestimated when the ordinal data had only four response categories. The negative bias was an inverse function of the number of categories. Conversely, the factor loadings were overestimated, although *trivially*, on average by WLSMV. Thus, the factor loadings in WLSMV can be considered essentially unbiased, especially when the latent distribution is only slightly nonnormal. The positive bias did not vary as a function of the number of categories. It is worth noting that the ARB for factor loadings in WLSMV increased as the degree of latent normality violation increased. However, the shape of the observed/latent distribution did not seem to have a remarkable effect on the ARB for factor loadings in MLR. Most importantly, regardless of the number of categories, WLSMV was consistently superior to MLR for factor loading estimates. Generally, WLSMV

Table 1 Average relative bias (ARB) and average root mean squared error (ARMSE) for factor loadings

N	Categories	MLR				WLSMV			
		Slightly Nonnormal		Moderately Nonnormal		Slightly Nonnormal		Moderately Nonnormal	
		ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE
200	4	-10.61	.1333	-10.73	.1426	0.96	.0887	3.82	.1036
	6	-5.47	.0915	-4.49	.0895	0.63	.0757	3.06	.0801
	8	-2.80	.0737	-3.29	.0843	0.62	.0705	3.24	.0786
	10	-2.46	.0724	-2.83	.0820	0.58	.0687	1.82	.0709
500	4	-10.73	.1185	-10.73	.1217	0.55	.0554	3.56	.0691
	6	-4.86	.0659	-4.48	.0657	0.57	.0406	2.95	.0543
	8	-2.84	.0508	-3.29	.0578	0.48	.0433	3.10	.0535
	10	-2.48	.0488	-2.85	.0552	0.47	.0425	1.70	.0455
1,000	4	-10.56	.1113	-10.58	.1133	0.63	.0393	3.64	.0553
	6	-4.79	.0574	-4.35	.0552	0.52	.0329	2.95	.0438
	8	-2.70	.0401	-3.14	.0462	0.56	.0311	3.15	.0444
	10	-2.34	.0381	-2.66	.0429	0.54	.0306	1.80	.0347

yielded more accurate factor loading estimates than MLR, as evidenced by its relatively small amount of ARB.

Regarding the overall quality of the estimated factor loadings, ARMSE varied as an inverse function of the sample size and the number of categories for both estimation methods. ARMSE was most pronounced in the conditions in which ARB was appreciable. ARMSE was, for instance, noticeably larger for MLR than for WLSMV when the observed indicators had only four response categories. Moreover, this discrepancy in overall performance between WLSMV and MLR became larger as the sample size increased. It is of particular interest that WLSMV was better than MLR in the overall quality of factor loading estimates from four to ten categories across different sample sizes, even when ordinal observed data were generated from a moderately nonnormal latent distribution. Uniformly, WLSMV can be considered better than MLR on the performance of factor loading estimates.

Interfactor correlations The ARB and ARMSE for interfactor correlations are provided in Table 2. Interfactor correlations were, on average, trivially biased (either positively or negatively) for both estimators. However, WLSMV introduced a marked bias into the estimates of interfactor correlations when the observed data were generated from a moderately nonnormal latent distribution, particularly in the sample of $N = 200$. Roughly speaking, the ARB for interfactor correlations was comparably smaller for MLR than for WLSMV across almost all conditions, indicating that WLSMV is inferior to MLR due to its higher degree of bias in interfactor correlation estimates.

With respect to the overall quality of the estimated interfactor correlations, ARMSE varied as an inverse function of the sample size and number of categories for both

estimation methods. The ARMSEs were similar for both estimators with slightly nonnormal latent distributions; however, ARMSE appeared to be smaller in MLR than in WLSMV with moderately nonnormal latent distributions. This implies that for a moderately nonnormal latent distribution, MLR demonstrates better performance than WLSMV for estimating the interfactor correlations.

Standard errors

Standard errors of factor loadings Table 3 presents the ARB and ARMSE values for standard errors of the factor loadings. The standard errors exhibited, on average, a slight bias (either positive or negative) for both estimators. It is noteworthy that a moderately negative bias was produced by WLSMV when the sample size was small (i.e., $N = 200$), reflecting that standard errors seem to be underestimated by WLSMV in the case of sample size $N = 200$. However, ARB in WLSMV reduced with increasing sample size. Generally, the performance of MLR surpasses that of WLSMV for estimating standard errors when the sample size is small and latent distributions are nonnormal. Regarding the overall quality of the estimated standard errors of factor loadings, ARMSE was an inverse function of sample size for both estimation methods. The ARMSEs were not very different for MLR and WLSMV across the conditions investigated here.

Standard errors of interfactor correlations Table 4 displays the ARB and ARMSE values for standard errors of the interfactor correlations. The standard errors demonstrated, on average, a slight bias (in either a positive or a negative direction) with MLR, whereas they were moderately

Table 2 Average relative bias (ARB) and average root mean squared error (ARMSE) for interfactor correlations

N	Categories	MLR				WLSMV			
		Slightly Nonnormal		Moderately Nonnormal		Slightly Nonnormal		Moderately Nonnormal	
		ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE
200	4	-0.20	.2083	-3.81	.2216	3.19	.2097	8.37	.2396
	6	0.15	.1949	1.59	.2006	2.53	.1966	7.00	.2087
	8	0.82	.1935	1.09	.2001	2.36	.1961	7.17	.2044
	10	0.28	.1936	1.48	.1971	2.14	.1948	6.45	.1980
500	4	-1.79	.1397	-5.43	.1565	2.70	.1414	5.04	.1604
	6	-1.37	.1277	0.38	.1316	2.18	.1289	4.72	.1312
	8	-0.74	.1230	-0.53	.1320	2.25	.1305	4.91	.1323
	10	-0.59	.1225	-0.14	.1316	1.98	.1258	4.18	.1297
1,000	4	-0.98	.0928	-4.34	.1055	0.99	.0932	5.56	.1170
	6	-0.38	.0877	0.79	.0896	0.80	.0876	4.59	.0949
	8	0.08	.0849	0.18	.0883	0.66	.0854	4.85	.0960
	10	-0.02	.0845	0.58	.0882	0.66	.0852	4.22	.0917

underestimated by that WLSMV estimator on average for a sample smaller than 500. It is noteworthy that ARB became trivial for WLSMV with a sample size of $N = 1,000$. In general, the performance of WLSMV is only considered reliable for estimating the standard errors of interfactor correlations, in terms of bias, when $N = 1,000$, but otherwise MLR outperforms WLSMV across most conditions. With respect to the overall quality of the estimated standard errors of interfactor correlations, ARMSE was an inverse function of sample size for both estimation methods. In examining the ARMSEs at each sample size, it is of particular interest that MLR was consistently smaller than WLSMV when the sample size

was $N = 200$ or 500 ; in contrast, WLSMV was steadily smaller than MLR for a sample size of $N = 1,000$. This reveals that WLSMV is better than MLR for estimating the standard errors of interfactor correlations when the sample size is relatively large—that is, $N = 1,000$ —whereas MLR has some advantage with smaller sample sizes of $N = 200$ or 500 .

Coverage of confidence intervals

Factor loadings Table 5 shows the average coverage of 95 % confidence intervals for the parameter estimates. The average coverage for factor loadings with MLR was adversely affected

Table 3 Average relative bias (ARB) and average root mean squared error (ARMSE) for standard errors of the factor loadings

N	Categories	MLR				WLSMV			
		Slightly Nonnormal		Moderately Nonnormal		Slightly Nonnormal		Moderately Nonnormal	
		ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE
200	4	-0.52	.1220	-1.16	.1337	-5.80	.1349	-5.09	.1474
	6	-1.35	.1214	-1.07	.1238	-6.44	.1307	-3.83	.1215
	8	-1.46	.1209	-1.55	.1280	-7.92	.1298	-4.76	.1216
	10	-1.18	.1215	-1.65	.1269	-8.30	.1310	-4.83	.1183
500	4	0.35	.0824	1.32	.0893	-1.37	.0919	-0.72	.0945
	6	0.72	.0776	0.10	.0807	-1.22	.0800	-0.38	.0812
	8	0.73	.0797	0.85	.0841	-1.46	.0805	0.05	.0779
	10	0.87	.0816	0.52	.0826	-1.92	.0803	0.17	.0772
1,000	4	-0.42	.0598	-0.06	.0678	-0.76	.0660	-0.03	.0760
	6	-1.10	.0610	-1.13	.0601	-0.94	.0628	-0.49	.0628
	8	-0.68	.0632	-0.83	.0624	-1.10	.0624	-0.98	.0632
	10	-1.53	.0616	-0.57	.0612	-1.82	.0613	1.08	.0656

Table 4 Average relative bias (ARB) and average root mean squared error (ARMSE) for standard errors of the interfactor correlations

N	Categories	MLR				WLSMV			
		Slightly Nonnormal		Moderately Nonnormal		Slightly Nonnormal		Moderately Nonnormal	
		ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE	ARB (%)	ARMSE
200	4	3.45	.0804	4.12	.0866	-7.76	.0873	-4.48	.0701
	6	3.13	.0768	1.99	.0668	-8.59	.0906	-8.70	.0920
	8	0.45	.0742	2.31	.0678	-13.68	.1374	-9.12	.0943
	10	0.59	.0750	3.66	.0724	-15.86	.1588	-10.71	.1086
500	4	-1.94	.0449	-2.53	.0466	-5.31	.0572	-6.18	.0652
	6	-1.23	.0451	-1.37	.0405	-4.55	.0514	-4.65	.0509
	8	-0.21	.0456	-1.64	.0414	-5.13	.0550	-5.84	.0600
	10	0.76	.0455	-1.36	.0415	-5.82	.0604	-6.40	.0651
1,000	4	0.83	.0304	0.90	.0285	-0.95	.0272	-1.13	.0268
	6	1.36	.0336	2.93	.0374	-1.14	.0251	0.40	.0227
	8	2.20	.0370	2.81	.0365	-0.84	.0241	-0.36	.0219
	10	2.98	.0408	3.45	.0408	-0.58	.0233	-0.02	.0227

by the size of ARB, in particular for those indicators with four response categories. It is noteworthy that an increase in sample size appeared to exacerbate the problem of lower average coverage, partly because of the comparably smaller standard error estimates in larger sample sizes. Furthermore, as the level of latent nonnormality increased, the average coverage decreased in the WLSMV estimation. This is thought to be due mainly to the relatively large bias with moderately nonnormal latent distributions. Moreover, the lower rate of coverage indicated lower power to capture the true factor loadings using MLR. In general, 95 % confidence intervals constructed using WLSMV estimates appear to be more reliable than those using

MLR estimates, in line with the true factor loadings when latent distributions are only slightly nonnormal. With moderately nonnormal latent distributions, WLSMV, however, is not superior to MLR in the recovery of true factor loadings, except for the conditions with four response categories.

Interfactor correlations As is shown in Table 5, the average coverage for interfactor correlations in MLR seemed to be stable and satisfactory across experimental conditions. On the other hand, the average coverage for interfactor correlations in WLSMV seemed, to some degree, to deviate from 95 % with moderately nonnormal latent distributions, mainly

Table 5 Average coverage of 95 % confidence intervals for the factor loadings and interfactor correlations

N	Categories	MLR				WLSMV			
		Slightly Nonnormal		Moderately Nonnormal		Slightly Nonnormal		Moderately Nonnormal	
		Loadings	Correlations	Loadings	Correlations	Loadings	Correlations	Loadings	Correlations
200	4	79.58 %	95.60 %	85.00 %	95.20 %	91.96 %	92.80 %	88.38 %	91.60 %
	6	91.60 %	94.80 %	93.28 %	95.20 %	91.76 %	91.60 %	89.60 %	89.80 %
	8	94.54 %	94.40 %	93.86 %	95.60 %	91.66 %	89.60 %	88.62 %	89.60 %
	10	94.56 %	94.20 %	94.06 %	94.60 %	91.34 %	88.80 %	90.78 %	91.00 %
500	4	48.50 %	94.80 %	60.98 %	93.40 %	94.20 %	93.00 %	87.96 %	90.80 %
	6	85.32 %	94.40 %	87.38 %	94.60 %	93.94 %	94.00 %	87.64 %	91.00 %
	8	91.78 %	95.00 %	91.26 %	93.20 %	93.40 %	95.00 %	86.92 %	90.40 %
	10	93.04 %	95.00 %	92.44 %	94.40 %	93.84 %	94.40 %	91.72 %	91.40 %
1,000	4	17.76 %	94.60 %	30.96 %	94.80 %	93.96 %	95.00 %	84.68 %	92.40 %
	6	71.28 %	96.00 %	78.48 %	95.80 %	93.92 %	95.40 %	83.40 %	93.00 %
	8	88.24 %	95.80 %	86.86 %	96.00 %	93.50 %	95.80 %	81.20 %	92.00 %
	10	89.80 %	96.00 %	89.74 %	96.20 %	93.54 %	95.20 %	90.06 %	92.80 %

because of a higher degree of bias and the underestimation of standard errors. Generally, it is evident that MLR is superior to WLSMV in the recovery of the true interfactor correlations across all conditions; however, WLSMV appears to be dependable when a latent distribution is slightly nonnormal in larger sample sizes of $N = 500$ or 1,000.

Chi-square test statistics

Table 6 gives the chi-square rejection rates for the two estimators. The boldface numbers in the table indicate unacceptable rejection rates, implying that acceptable difference rates in the table are within the range [2.5 %, 7.5 %] (Bradley, 1978). It was found that both MLR and WLSMV performed well, yielding an approximately 5 % rejection rate across most conditions, with some exceptions in the smallest sample size, $N = 200$, condition. In these exceptional conditions, the proposed model seemed to be over-rejected, producing slightly inflated Type I error rates. Generally speaking, both the corrected chi-square test statistics performed well in controlling for Type I error rates when the sample size was greater than 500.

Discussion

This study was designed to compare the performance of MLR and WLSMV with regard to parameter estimates, standard errors, and chi-square test statistics in a correlated two-factor model with ordinal observed indicators under different experimental configurations of latent response distributions, numbers of categories, and sample sizes. Several general findings

are discussed, as follows. First, both estimators were not subject to the problems of improper solutions or nonconvergence with a small sample ($N = 200$) in a correlated two-factor model, consistent with previous simulation studies (Flora & Curran, 2004; Herzog, Boomsma, & Reinecke, 2007). Prior scholarship, however, has observed nonconvergence or improper solutions, in particular, when data were analyzed in quite small samples $N = 100$ or 150 (Rhemtulla, Brosseau-Liard, & Savalei, 2012; Yang-Wallentin et al., 2010).

Second, this study replicated previous results that factor loadings are typically underestimated by MLR but are essentially unbiased with WLSMV (Beauducel & Herzberg, 2006; DiStefano, 2002; Flora & Curran, 2004). Interestingly, a clear superiority of WLSMV over MLR in factor loading estimates was found in this study, irrespective of the number of categories. This study also revealed that the factor loadings obtained by WLSMV were more precise and accurate than those obtained by MLR when the latent normality assumption was moderately violated. Generally speaking, WLSMV was preferable to MLR across most of the conditions observed in this study, given its properties of being less biased and having small sampling variation in estimating factor loadings.

As occurred in previous simulation studies, a mixture of positive and negative bias in interfactor correlations was found with MLR, and the interfactor correlation was essentially unbiased by WLSMV under the latent normality assumption (Beauducel & Herzberg, 2006). In this study, the increased bias of interfactor correlation estimates made WLSMV inferior to MLR in overall performance, particularly for a sample size of $N = 200$ or under a moderate violation of latent normality. That is, WLSMV may overestimate the

Table 6 Rejection rates for the null model at a probability of Type I error $\alpha = .05$

N	Categories	MLR		WLSMV	
		Slightly Nonnormal Rejection (%)	Moderately Nonnormal Rejection (%)	Slightly Nonnormal Rejection (%)	Moderately Nonnormal Rejection (%)
200	4	10.6 %	10.4 %	8.0 %	6.4 %
	6	8.0 %	8.4 %	7.8 %	6.4 %
	8	9.4 %	8.8 %	10.2 %	6.4 %
	10	7.0 %	9.0 %	11.6 %	8.2 %
500	4	7.4 %	5.0 %	8.2 %	6.6 %
	6	8.0 %	7.8 %	7.6 %	4.0 %
	8	6.8 %	7.2 %	7.0 %	5.8 %
	10	6.2 %	8.0 %	6.8 %	5.2 %
1,000	4	5.0 %	6.6 %	5.8 %	4.0 %
	6	4.8 %	5.6 %	5.8 %	4.8 %
	8	5.8 %	6.8 %	6.0 %	5.4 %
	10	7.2 %	5.6 %	5.8 %	6.0 %

The boldface numbers indicate unacceptable rejection rates, i.e., outside the range [2.5 %, 7.5 %]

association between factors when the sample size is relatively small and/or when a latent distribution is moderately nonnormal. These findings suggest that the quality of the factor loading estimates is better for WLSMV than for MLR, but that WLSMV may lead to more biased interfactor correlations than MLR because the latent normality assumption is moderately violated. This implies that the polychoric correlation estimates may demonstrate robustness against violations of the latent normality assumption in estimating factor loadings rather than in interfactor correlations. This observation is consistent with that of Coenders, Satorra, and Saris (1997), who concluded that Pearson product-moment correlations between ordinal observed indicators using ML perform badly in factor loading estimates due to the categorical nature of measurement. However, such lower measurement quality estimates of ordinal variables can lead to relatively unbiased point estimates of factor relationships. The overall quality of parameter estimates (i.e., ARMSE) varied positively as a function of the number of categories and sample size, suggesting that increasing sample size and the number of categories can advance the overall quality of factor loading and interfactor correlation estimates.

Third, with respect to the standard errors of factor loadings in the present study, it was observed that MLR outperformed WLSMV when the sample size was $N = 200$ or when the latent distributions were slightly nonnormal. In addition, less biased standard errors of interfactor correlations led to overall performance of MLR that was superior to that of WLSMV when the sample size was $N = 200$ or 500 ; however, this advantage was not observed when $N = 1,000$. It is interesting that the overall quality of the standard error estimates was quite sensitive to sample size, regardless of the number of categories and the level of the latent normality assumption violation.

Fourth, the substantially negative bias in parameter estimates, coupled with small standard errors, led to a lower rate of coverage for factor loadings with four categories using MLR, which resonates with previous research (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Thus, researchers have to pay attention to comparably poor MLR coverage rates with few categories, because bias in MLR's parameter estimates is highly pronounced. On the whole, the WLSMV coverage rates of factor loadings are higher than those of MLR when a latent distribution is slightly nonnormal, whereas MLR's coverage for interfactor correlations is uniformly better than that of WLSMV across all experimental conditions examined in this study.

Fifth, in a few conditions, models were rejected more often than expected using adjusted chi-square test statistics for both estimators when the sample size was $N = 200$. Researchers need to exercise caution in the evaluation of model fits under a small sample size, and they should take into account the supplemental fit indices (e.g., RMSEA) usually provided by statistical software programs. An assessment of supplemental fit indices was not included in the present study, mainly because (1) these fit indices, unlike chi-square statistics, do not follow

a known sampling distribution and (2) they do not have coherent cutoff values for fit indices in applications, to use to evaluate their performance. However, one can expect that RMSEA, for example, would exhibit adequate power in the model evaluation when a model had no specification error (like the CFA model in this study). Future research assessing the effects of various fit indices is still suggested, specifically addressing the question of which fit indices are reliable and robust to detect model misspecification.

Finally, as we may be aware, there are multitudinous combinations to manipulate in a simulation study, but researchers can only focus on some factors of particular interest to make the research design feasible. Therefore, this study shares the same limitation as all simulation studies, in that the results cannot be generalized beyond the experimental conditions investigated in the study. Although previous simulation studies have suggested that the estimation of ordinal CFA models is robust to slight model misspecification, a natural extension of this study would consider different levels of misspecified models (e.g., cross-factor loadings) using MLR and WLSMV. In addition, given that a simple/common two-factor CFA model was specified in this study, an interesting avenue of further investigation would consider complex/advanced models (e.g., multiple-group CFA models or structural equation models) in order to examine other scenarios in empirical research.

Summary and conclusions

In closing, the conclusions of this study can be summarized as follows: (1) regardless of the number of categories, the factor loading estimates under WLSMV are less biased than those under MLR; (2) WLSMV appears to yield moderate overestimation of the interfactor correlations when the sample size is relatively small and/or when a latent distribution is moderately nonnormal; (3) the estimates of standard errors under WLSMV demonstrate much larger sampling variability than those under MLR when the sample size is small and the latent distribution is nonnormal; (4) the substantial underestimation of factor loadings using MLR may result in a lower rate of confidence interval coverage for factor loadings; (5) the MLR coverage rates of interfactor correlations are uniformly better than the WLSMV coverage rates across all experimental conditions; and (6) the proposed model tends to be over-rejected by corrected chi-square test statistics under both MLR and WLSMV in the case of a small sample size of $N = 200$.

It is worthwhile to point out that each estimator has its advantages and disadvantages, as was discussed above. This study does provide conclusive evidence that WLSMV performs uniformly better than MLR in factor loading estimates across all experimental conditions (i.e., regardless of sample size, the number of categories, and the degree of latent

normality violation). However, WLSMV, for instance, also has its own weaknesses of interfactor correlations and standard errors in estimation when the sample size is small and/or when a latent distribution is moderately nonnormal. Likewise, MLR has its unique strengths—for instance, generally less biased standard error estimates and good recovery of the population interfactor correlations. Thus, further research will be needed to help applied researchers better understand the pros and cons of different estimators under certain circumstances in order to select an “appropriate” estimator for the factor-analytic model with ordinal data.

References

- 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*, 186–203. doi:10.1207/s15328007sem1302_2
- Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *58*, 430–450.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematics and Statistical Psychology*, *37*, 62–83.
- Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling*, *4*, 261–282. doi:10.1080/10705519709540077
- Curran, P. J., West, S. G., & Finch, G. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346.
- 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*, 466–491.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275–299.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicator: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling*, *16*, 625–641.
- Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling*, *14*, 361–390. doi:10.1080/10705510701301602
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research*, *26*, 329–367.
- Jackson, D. L., Gillaspay, J. A., & Pure-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6–23.
- Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48*, 398–407.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.
- Jöreskog, K. G. (2005). *Structural equation modeling with ordinal variables using LISREL*. Retrieved from www.ssicentral.com/lisrel/techdocs/ordinal.pdf
- Jöreskog, K. G., & Sörbom, D. (1996). *Prelis 2: User's reference guide: A program for multivariate data screening and data summarization*. Chicago, IL: Scientific Software.
- Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*, 495–507.
- Li, C.-H. (2012). Validation of the Chinese version of the Life Orientation Test with a robust weighted least squares approach. *Psychological Assessment*, *24*, 770–776. doi:10.1037/a0026612
- Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514–534.
- Magnus, J. R., & Neudecker, H. (1986). Symmetry, 0–1 matrices and Jacobians: A review. *Econometric Theory*, *2*, 157–190. doi:10.1017/S0266466600011476
- Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.
- Micceri, T. (1989). The unicorn, the normal curve, than other improbable creatures. *Psychological Bulletin*, *105*, 156–166.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.
- Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.
- Muthén, B. O. (2002). *Using Mplus Monte Carlo simulation in practice: A note on assessing estimation quality and power in latent variable models*. Retrieved from <https://www.statmodel.com/download/webnotes/mc1.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*. Retrieved from http://gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf
- Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–180.
- Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor-analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*, 19–30.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.
- Oranje, A. (2003, April). *Comparison of estimation methods in factor analysis with categorical variables: Applications to NAEP data*. Paper presented at the annual meeting of the American Education Research Association (AERA), Chicago, IL.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312.

- Quiroga, A. M. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables*. Unpublished Doctoral dissertation, Uppsala University, Uppsala, Sweden.
- Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: Guildford Press.
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Erlbaum.
- 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*, 354–373. doi:10.1037/a0029315
- Rigdon, E. E. (1998). Structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 251–294). Mahwah, NJ: Erlbaum.
- Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24*, 367–386.
- Satorra, C., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods, 15*, 352–367.
- Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*, 392–423. doi:10.1080/10705511.2010.489003
- Yuan, K. H., & Bentler, P. M. (1997). Improving parameter tests in covariance structure analysis. *Computational Statistics and Data Analysis, 26*, 177–198.
- Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289–309.