Behavior Research Methods (2021) 53:2191–2213
https://doi.org/10.3758/s13428-021-01547-z

Statistical estimation of structural equation models with a mixture of continuous and categorical observed variables

Cheng-Hsien Li[1]

Accepted: 15 January 2021 / Published online: 31 March 2021

Abstract

In the social and behavioral sciences, observed variables of mixed scale types (i.e., both continuous and categorical observed variables) have long been included in structural equation models. However, little is known about the impact of mixed continuous and categorical observed variables on the performance of existing estimation methods. This study compares two popular estimation methods with robust corrections, robust maximum likelihood (MLR) and diagonally weighted least squares (DWLS), when mixed continuous and categorical observed data are analyzed, evaluating the behavior of DWLS and MLR estimates in both measurement and full structural equation models. Monte Carlo simulation was carried out to examine the performance of DWLS and MLR in estimating model parameters, standard errors, and chi-square statistics. Two population models, a correlated three-factor measurement model and a five-factor structural equation model, were tested in combination with 36 other experimental conditions characterized by the number of observed variables' categories (2, 3, 4, 5, 6, and 7), categorical observed distribution shape (symmetry and slight asymmetry), and sample size (200, 500, and 1000). Data generation and analysis were performed with M*plus* 8. Results reveal that (1) DWLS yields more accurate factor loading estimates for categorical observed variables than MLR, whereas DWLS and MLR produce comparable factor loading estimates for continuous observed variables; (2) inter-factor correlations and structural paths are estimated equally well by DWLS and MLR in nearly all conditions; (3) robust standard errors of parameter estimates obtained by MLR are slightly more accurate than those produced by DWLS in almost every condition, but the superiority of MLR over DWLS is not clearly evident once a medium or large sample is used (i.e., $n = 500$ or 1000); and (4) DWLS is systematically superior to MLR in controlling Type I error rates, but this superiority is attenuated with increasing sample size. The article concludes with a general discussion of the findings and some recommendations for practice and future research.

Keywords Diagonally weighted least squares · Maximum likelihood · Robust statistics · Mixed scale types · Monte Carlo

Introduction

The use of observed variables of mixed scale types (e.g., a mixture of continuous and categorical observed variables within the same measurement model) to operationalize a latent construct is becoming increasing common in applied studies, meriting increased research attention to measurement models with mixed item response types. In an organizational setting, for instance, studies measuring "job embeddedness" can naturally involve both continuous and categorical observed variables (Hom et al., 2009; Mitchell et al., 2001). In such a case, examples of continuous observed variables may include the number of years a person has been in their present position or has worked for a company and the number of coworkers highly dependent on them, while categorical variables may include the frequency of communication with managers and key customers of the company, measured on a five-point Likert-type scale, or home ownership and current marital status (yes or no). In the field of political science, the construct of political-economic risk in question (Quinn, 2004) is typically linked to continuous (e.g., gross domestic product, black-market premium), ordinal observed variables using a six-point scale (e.g., lack of expropriation risk, lack of corruption), and dichotomous variables (e.g., judicial independence). Another general example in psychology comes from the study of substance abuse, in which measurements of alcohol use can be manifested by a set of continuous (e.g., age at first drinking,

✉ Cheng-Hsien Li
lichengh@mail.nsysu.edu.tw

[1] Institute of Human Resource Management, College of Management, National Sun Yat-sen University, 70 Lienhai Rd, Kaohsiung 80424, Taiwan

largest number of drinks consumed during the last 30 days, largest number of drinks one can hold), ordinal (e.g., weekly drinking frequency, perception about availability of alcohol), and binary observed variables (e.g., experience of drinking alcohol ever, a blackout experience). There have been many applications of measurement models with a mixture of continuous and categorical observed variables in medical, social, and behavioral research (see, e.g., Diemer & Li, 2012; Gueorguieva & Sanacora, 2006; Lee & Xu, 2003; Morisky et al., 1998; Sammel et al., 1997; Song et al., 2007; Zhou et al., 2014; and references therein). Computational challenges for the estimation process caused by mixed continuous and categorical data suggest that increased attention to these situations in the broader latent variable modeling framework is warranted.

Numerous simulation studies have compared the relative performance of different estimators in confirmatory factor analysis (CFA) and structural equation models (SEM) with observed data characterized as either continuous or categorical. For example, Olsson et al. (2000) found that maximum likelihood (ML) resulted in less biased structural parameters and more accurate overall model inference than weighted least squares (WLS) when observed variables were continuous. Among all ML-based estimators in M*plus*, Maydeu-Olivares (2017) concluded that robust maximum likelihood (MLR) yielded the most reliable standard error estimates, and mean- and variance-adjusted maximum likelihood (MLMV) performed the best on overall model evaluation using chi-square test statistics. For categorical data, previous simulation studies demonstrated that factor loading estimates were generally less biased by diagonally weighted least squares (DWLS) than ML-based estimators (Bandalos, 2014; Beauducel & Herzberg, 2006; Li, 2016a). Bandalos (2014), and Li (2016b) found that DWLS outperformed ML-based estimators at estimating structural coefficients in a correctly specified model. Previous simulation studies have observed mixed findings regrading standard error estimates of model parameters. For instance, Bandalos (2014) suggested that standard errors obtained from DWLS were generally less biased than those from ML. On the other hand, ML exhibited better performance than DWLS under some conditions (e.g., small sample size, asymmetric distribution of observed variables, or a combination of both). Li (2016a, 2016b) showed that MLR generally produced less bias in standard error estimates of factor loadings, inter-factor correlations, and structural paths than did DWLS across nearly all conditions. In terms of overall model evaluation, earlier research suggested that DWLS was superior to MLR but inferior to MLMV in controlling Type I error rates for testing an SEM model (Bandalos, 2014; Li, 2016b); under latent normality assumption violation conditions, DWLS performed as well as MLR at maintaining Type I error rates for a measurement model (Li, 2016a).

However, statistical estimation of CFA and SEM models measured by mixed continuous and categorical observed variables has not yet been fully studied. It is noted that there have been some simulation studies approaching this inquiry from a Bayesian perspective (for more detailed discussions see, e.g., Lee & Zhu, 2000; Fahrmeir & Raach, 2007; Quinn, 2004; Samani & Ganjali, 2011; Song & Lee, 2001), which has valuably contributed to the literature. In contrast, the properties of existing frequentist estimation methods for model parameters, standard errors, and chi-square statistics when observed variables have mixed scales are still unclear.

There appears to be no consensus in favor of certain estimation methods among researchers when observed variables have different types of scales in current research practice. A review of empirical studies from two high-impact journals (i.e., *Journal of Organizational Behavior* and *Journal of Applied Psychology*) was conducted to examine the frequency of each estimation method used in confirmatory factor analysis or structural equation modeling between the years 2018 and 2019. Note that multilevel CFA or SEM and path analysis were excluded from this review. A total of 84 studies were identified from the search, resulting in 161 CFA models and 10 SEM models. Of the 84 studies, 35 studies (41.67%) clearly reported the method of estimation. ML ($n = 6$) or robust ML ($n = 8$) was used for continuous observed variables; ML ($n = 7$), robust ML ($n = 5$), or DWLS ($n = 4$) was utilized for categorical observed variables; and ML ($n = 3$) or robust ML ($n = 2$) was employed in the model estimation with mixed continuous and categorical observed variables.

It seems that normal theory-based ML with the sample covariance matrix has been widely used across CFA and SEM applications primarily because ML is the most well-known estimator and is usually the default setting for most software programs (e.g., M*plus*, LISREL, Amos, EQS, or R). However, based on previous simulation studies on categorical observed variables (Bandalos, 2014; Beauducel & Herzberg, 2006; DiStefano, 2002; Li, 2016a, 2016b; Yang-Wallentin et al., 2010), the "conventional" ML estimation method without robust corrections is not advisable in research practice. In general, when employing categorical variables in a model, the classical estimation techniques (i.e., frequentist statistics) DWLS and ML (with various robust corrections) are the two most commonly employed or recommended estimation methods in the current SEM literature, including empirical applications and simulation studies.

Among robust ML-based estimators (e.g., MLR, MLM, or MLMV) in M*plus*, MLR has been set as the default estimator for various models with continuous or categorical dependent variables (Muthén & Muthén, 1998–2017). Also, MLR is the only estimation option to (1) adjust standard errors for the effects of non-normality or clustering, and (2) model random effects in the conditions of complex data (e.g., non-independence of observations, sampling weights). Instead,

MLMV or MLM requires continuous dependent variables and is limited to specific models in which missing data, mixture analyses, or random effects are not present (Muthén & Muthén, 1998–2017). However, implementing random effects in moderation analysis or dealing with missing data in an SEM model is not uncommon in the social and behavioral sciences. Therefore, this study mainly focused on the estimation performance of MLR not only because of its robust correction to standard errors using a sandwich estimator but also its flexibility in model specification. On the other hand, although Muthén and Muthén (1998–2017) recommended that DWLS should be implemented in a general model specification (e.g., general CFA or SEM models) when at least one binary or ordered categorical dependent variable is present, one interesting finding from the above review is that DWLS seemed not the first choice for applied researchers when a mixture of continuous and categorical observed variables was present. One possible argument could be due to the uncertainty of estimation performance of DWLS in applications.

Overall, MLR has been developed to permit modeling non-normal (approximately) continuous variables, whereas DWLS has been implemented to deal with categorical data. Although MLR with the sample covariance matrix is not, generally speaking, appropriate for categorical data, researchers have suggested that data can be considered "approximately continuous" if the number of observed variable response categories is sufficiently large. Moreover, results of simulation studies imply that MLR estimates exhibit some robustness, even when observed variable distributions depart from normality. In practice, MLR is often treated as a viable alternative to DWLS with a polychoric correlation matrix in categorical CFA and SEM models when the number of response categories for each observed variable is five or more (Raykov, 2012; Rhemtulla et al., 2012; Yang-Wallentin et al., 2010). However, until the present, limited information has been provided comparing MLR and DWLS in the context of mixed scale observed variables. Comparison of MLR and DWLS may shed some light on their statistical estimation performance under suboptimal conditions, when observed variables are not made up solely of continuous or categorical ones, but of a mixture of both. The importance of studying the impact of mixed continuous and categorical data, and consequently selecting an appropriate estimation method based on empirical conditions, cannot be overstated.

To address the aforementioned research gaps in the established literature, this study is motivated to advance scholarly understanding regarding the impact of mixed item scale types (continuous and categorical variables) on parameter estimates (factor loadings, inter-factor correlations, and structural paths), standard errors, and chi-square statistics through a Monte Carlo simulation study. The performance of the two different estimation methods (MLR and DWLS) are evaluated for two model specifications: one CFA model and one SEM model. The two frequentist estimation methods, MLR and DWLS, with their robust corrections to standard errors and chi-square statistics are briefly reviewed in the next section.

Estimation methods

Latent variable modeling, particularly applications of confirmatory factor analysis (CFA) and structural equation modeling (SEM), have enjoyed widespread popularity in the social and behavioral sciences for more than two decades. Confirmatory factor analysis has been extensively used to provide evidence of construct validity in theory-based instrument construction and development. Confirmatory factor analytic models have the practical advantage of accounting for measurement error in observed variables by explicitly modeling pertinent error variances as parameter estimates. Given a tenable confirmatory measurement model, a structural equation model simultaneously captures relational phenomena among latent constructs of interest, including, but not limited to, inter-factor correlations, direct effects, mediating/indirect effects, and moderated effects.

In practice, normal theory-based maximum likelihood (ML) with the sample covariance matrix remains the most well-known and most frequently used estimation method, because of its desirable estimation properties of asymptotic unbiasedness, consistency, normality, and maximal efficiency in an infinite sample (Bollen, 1989). ML has been developed with the assumption that observed data are continuous and multivariate normally distributed in the population (Bollen, 1989; Jöreskog, 1969). Under the assumption of normality, parameter estimates of ML are obtained by maximizing the likelihood of the observed data, which is equivalent to minimizing the maximum likelihood fit function (Bollen, 1989). If observed variables exhibit non-normality to some degree (due to skewness, leptokurtosis, or/and heavy tails), standard errors and chi-square statistics should be statistically corrected to enhance robustness against the presence of non-normality. More specifically, the asymptotic covariance matrix of the parameter estimates from ML is no longer consistently estimated, resulting in inaccurate standard error estimates (Yuan et al., 2005; Yuan & Hayashi, 2006). Instead, to compute standard error estimates by robust ML (MLR) estimation, a consistent estimator of the asymptotic covariance matrix of the parameter estimates can be obtained using the pseudo maximum likelihood (PML) approach (Asparouhov & Muthén, 2005; Savalei, 2010; Yuan & Schuster, 2013). The asymptotic covariance matrix of the parameter estimates contains the sample estimates of skewness and kurtosis of the observed variables in order to correct for possible violation of the normality assumption (Yuan et al., 2005).

At the same time, the chi-square statistic for overall model fit computed using a Wishart-based likelihood is very likely to be substantially overestimated. When non-normal observed

variables are present, a robust correction to the chi-square statistic using PML is therefore constructed to closely follow a chi-square distribution. A scale factor in the correction can accommodate the effects of skewness and kurtosis in the observed data to adjust for deviation from normality. These statistical adjustments for standard errors and chi-square statistics in MLR, which can in general enhance the precision of parameter estimates and reduce the inflation of chi-square statistics, have improved estimation when modeling non-normal data (Asparouhov & Muthén, 2005).

However, observed variables measured with a set of ordered categories (e.g., Likert scales) are commonly used as indicator variables for latent constructs in the social and behavioral sciences. By treating ordered categorical variables as if they were continuous in nature, as is required by MLR, the accuracy and precision of model parameter estimates could be compromised, resulting in erroneous conclusions drawn from empirical data. Strictly speaking, it is not advisable to use ML with the sample covariance matrix when observed variables are categorical, mainly because Pearson product-moment correlations cannot reflect proper relationships among categorical observed variables (Bollen, 1989; Muthén & Kaplan, 1992; Olsson, 1979). This problem has generally plagued applied researchers utilizing various latent variable modeling techniques.

Practically, the diagonally weighted least squares method (DWLS: Jöreskog & Sörbom, 1996; Muthén et al., 1997) with a polychoric correlation matrix has been proposed for use when categorical data are employed in statistical analysis. Although DWLS makes no assumptions about observed variable distribution shape, it assumes that a continuous, normal, latent response distribution gave rise to each categorical observed variable in the population from which samples are drawn. Parameter estimates of DWLS are then obtained by minimizing the diagonally weighted least squares fit function (Muthén et al., 1997). Because the weight matrix contains only reduced information (i.e., diagonal elements), the parameter estimates obtained by DWLS are not asymptotically efficient (i.e., smaller sampling error), leading to inaccurate standard error estimates (Savalei, 2014). Therefore, upward corrections applied to standard errors are suggested to compensate for the loss of efficiency (Muthén et al., 1997). A robust correction to standard errors is implemented in the estimated asymptotic covariance matrix of the parameter estimates for DWLS estimation (Muthén et al., 1997).

In addition, as chi-square statistics produced by DWLS are no longer asymptotically chi-square distributed, a robust correction implemented in DWLS estimation entails adjusting for both the chi-square statistic's mean and variance to make its shape approximate the reference chi-square distribution (Asparouhov & Muthén, 2010). Therefore, the mean- and variance-adjusted chi-square statistic can be implemented in the DWLS estimator (e.g., WLSMV in M*plus*). Importantly, the estimated asymptotic covariance matrix need not be inverted (i.e., a positive definite matrix) in the computation of adjusted chi-square test statistics. Chi-square statistics are downwardly adjusted to compensate for the effect of including only reduced information in the weight matrix. This correction can help control the probability of Type I error (i.e., rejecting a correctly specified model by chance). DWLS with robust corrections to standard errors and chi-square statistics has proved useful in the analysis of categorical CFA and SEM models under a variety of conditions (e.g., a varying number of response alternatives, different levels of distributional asymmetry in observed variables, sample sizes) investigated by several simulation studies (Bandalos, 2014; Li, 2016a; Li, 2016b; Rhemtulla et al., 2012; Yang-Wallentin et al., 2010). Next, statistical estimation of a structural equation model with mixed continuous and categorical observed variables is discussed below.

Structural equation modeling

A structural equation model, a synthesis of measurement models and structural (regression) models, is used to simultaneously examine hypothetical relationships among latent variables. A measurement model with a mixture of continuous and categorical observed variables, in general, is partitioned into two sub-models: (i) one measurement model manifested by a set of continuous observed variables, and (ii) one measurement model manifested by a set of continuous latent response variables underlying categorical observed variables. Each sub-model in the measurement model can be specified as

$$\mathbf{x} = \boldsymbol{\Lambda}_{\mathrm{x}}\boldsymbol{\xi} + \boldsymbol{\delta}_{\mathrm{x}}, \mathbf{s} = \boldsymbol{\Lambda}_{\mathrm{s}}\boldsymbol{\eta} + \boldsymbol{\varepsilon}_{\mathbf{s}}, \tag{1}$$

$$\mathbf{y}^{*} = \boldsymbol{\Lambda}_{\mathrm{y}*}\boldsymbol{\xi} + \boldsymbol{\delta}_{\mathbf{y}*}, \text{and } \mathbf{t}^{*} = \boldsymbol{\Lambda}_{\mathbf{t}*}\boldsymbol{\eta} + \boldsymbol{\varepsilon}_{\mathbf{t}*}, \tag{2}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_p]'$ and $\mathbf{s} = [s_1, s_2, \ldots, s_m]'$ are $p \times 1$ and $m \times 1$ vectors of continuous observed variables x and s, respectively, $\boldsymbol{\xi}$ is an $r \times 1$ vector of exogenous latent variables with $\mathrm{Cov}(\xi) = \boldsymbol{\Phi}$ (an $r \times r$ variance-covariance matrix of latent variables ξ), η is a $g \times 1$ vector of endogenous latent variables, $\boldsymbol{\Lambda}_{\mathrm{x}}$ and $\boldsymbol{\Lambda}_{\mathrm{s}}$ are $p \times r$ and $m \times g$ matrices of factor loadings linking ξ and x, and η and s, respectively, δ_{x} is a $p \times 1$ vector of measurement errors in x with a multivariate normal distribution $\mathrm{N}\sim(\mathbf{0}, \boldsymbol{\Theta}_{\delta_{\mathrm{x}}})$, and $\boldsymbol{\varepsilon}_{\mathrm{s}}$ is an $m \times 1$ vector of measurement errors in s with $\mathrm{N}\sim(\mathbf{0}, \boldsymbol{\Theta}_{\varepsilon_{\mathrm{s}}})$. $\boldsymbol{\Theta}_{\delta_{\mathrm{x}}}$ and $\boldsymbol{\Theta}_{\varepsilon_{\mathrm{s}}}$ are $p \times p$ and $m \times m$ diagonal matrices of error variances for $\mathbf{x}$ and $\mathbf{s}$, respectively, assuming $\boldsymbol{\delta}_{\mathbf{x}}$ are independent of one another and of latent variables $\boldsymbol{\xi}$, and $\boldsymbol{\varepsilon}_{\mathbf{s}}$ independent of all other measurement errors and latent variables $\boldsymbol{\eta}$. For categorical observed variables, $\mathbf{y}^{*} = [y_1^{*}, y_2^{*}, \ldots, y_q^{*}]'$ and $\mathbf{t}^{*} = [t_1^{*}, t_2^{*}, \ldots, t_n^{*}]'$ represent $q \times 1$ and $n \times 1$ vectors of continuous, normal, latent response

variables y^* underlying each categorical observed variable y, and latent response variables t^* underlying each categorical observed variable t, $\mathbf{\Lambda_{y^*}}$ and $\mathbf{\Lambda_{t^*}}$ are $q \times r$ and $n \times g$ matrices of factor loadings linking $\boldsymbol{\xi}$ and $\mathbf{y}^*$, and $\boldsymbol{\eta}$ and $\mathbf{t}^*$, respectively, $\boldsymbol{\delta}_{\mathbf{y}^*}$ is a $q \times 1$ vector of measurement errors in $\mathbf{y}^*$ with $N\sim\left(\mathbf{0}, \mathbf{\Theta}_{\delta_{y^*}}\right)$, and $\boldsymbol{\varepsilon}_{\mathbf{t}^*}$ is an $n \times 1$ vector of measurement errors in $\mathbf{t}^*$ with $N\sim\left(\mathbf{0}, \mathbf{\Theta}_{\varepsilon_{t^*}}\right)$. $\mathbf{\Theta}_{\delta_{y^*}}$ and $\mathbf{\Theta}_{\varepsilon_{t^*}}$ are $q \times q$ and $n \times n$ diagonal matrices of error variances for $\mathbf{y}^*$ and $\mathbf{t}^*$, respectively, assuming $\boldsymbol{\delta}_{\mathbf{y}^*}$ are independent of one another and of latent variables $\boldsymbol{\xi}$, and $\boldsymbol{\varepsilon}_{\mathbf{t}^*}$ independent of all other measurement errors and latent variables $\boldsymbol{\eta}$.

Further, the relationship between each continuous, normal, latent response variable y_i^* and its categorical observed variable y_i or between each latent response variable t_i^* and its categorical observed variable t_i is defined by a set of thresholds. Note that unlike continuous observed variables, the variances of measurement errors (i.e., the diagonal elements) are not identified in a measurement model with categorical observed variables. These variances can be identified by standardizing either the latent response variables $\mathbf{y}^*$ and $\mathbf{t}^*$ or their measurement error terms. In order to metricize the latent response variables, variances of the latent response variables $\mathbf{y}^*$ and $\mathbf{t}^*$ are assumed for mathematical convenience to be equal to unity here. However, if all categorical observed variables were intended to be treated as if they were approximately continuous observed variables in the model estimation (i.e., using ML with the sample-based covariance matrix), the above measurement model would reduce to the general model for confirmatory factor analysis with continuous observed variables simply (see, e.g., Bollen, 1989, p. 233).

Let $\mathbf{z} = [\mathbf{x}, \mathbf{y}^*]$' and $\mathbf{u} = [\mathbf{s}, \mathbf{t}^*]$'. A structural (regression) model with exogenous and endogenous latent variables measured by a mixture of continuous and categorical observed variables is then defined as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{3}$$

where $\mathbf{B}$ is a $g \times g$ matrix of structural regression coefficients with zero diagonal elements among $\boldsymbol{\eta}$ (assuming $|\mathbf{I} - \mathbf{B}| \neq 0$), $\mathbf{\Gamma}$ is a $g \times r$ matrix of structural regression coefficients between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ is a $g \times 1$ vector of disturbance terms in $\boldsymbol{\eta}$ with a multivariate normal distribution $N\sim(\mathbf{0}, \mathbf{\Psi})$, and $\mathbf{\Psi}$ is a $g \times g$ diagonal matrix of residual variances for $\boldsymbol{\eta}$, assuming disturbance terms $\boldsymbol{\zeta}$ are independent of all other disturbance terms and latent variables ($\boldsymbol{\xi}$ and $\boldsymbol{\eta}$). It follows that $\mathrm{Cov}(\boldsymbol{\eta}) = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi})(\mathbf{I} - \mathbf{B})^{-1}$'.

Let θ continuous observed variables x and s as well as continuous, normal, latent response variables y^* and t^* of a general SEM model implied by θ can be expressed as

$$\mathbf{\Sigma}(\theta) = \begin{matrix} \mathbf{\Sigma_u} & \\ & \mathbf{\Sigma_w} \end{matrix} \quad \begin{matrix} \\ \mathbf{\Sigma_z}, \end{matrix} \tag{4}$$

$$\Sigma_u(\theta) = \begin{matrix} \Sigma_{ss} & \\ \Sigma_{st^*} & \Sigma_{t^*t^*}, \end{matrix} \quad \Sigma_w(\theta) = \begin{matrix} \Sigma_{sx} & \Sigma_{t^*x} \\ \Sigma_{sy^*} & \Sigma_{t^*y^*}, \end{matrix} \tag{5}$$

$$\text{, and} \quad \Sigma_z(\theta) = \begin{matrix} \Sigma_{xx} & \\ \Sigma_{xy^*} & \Sigma_{y^*y^*}. \end{matrix}$$

By standardizing the latent response variables $\mathbf{y}^*$ and $\mathbf{t}^*$ to achieve model identification, variances of the latent response variables $\mathbf{y}^*$ and $\mathbf{t}^*$ are assumed to be equal to unity in the aforementioned section. As a consequence, three different sets of matrices are produced: (1) $\mathbf{\Sigma_{xx}}$ is a $p \times p$ variance-covariance matrix, $\mathbf{\Sigma_{ss}}$ an $m \times m$ variance-covariance matrix, and $\mathbf{\Sigma_{sx}}$ a $p \times m$ variance-covariance matrix; (2) $\mathbf{\Sigma_{xy^*}}$ is a $p \times q$ matrix representing polyserial correlations, $\mathbf{\Sigma_{st^*}}$ an $m \times n$ matrix of polyserial correlations, $\mathbf{\Sigma_{sy^*}}$ a $q \times m$ matrix of polyserial correlations, and $\mathbf{\Sigma_{t^*x}}$ is a $p \times n$ matrix of polyserial correlations; and (3) $\mathbf{\Sigma_{y^*y^*}}$ and $\mathbf{\Sigma_{t^*t^*}}$ have unit diagonal elements only and therefore reduce to a $q \times q$ polychoric correlation matrix and an $n \times n$ polychoric correlation matrix, respectively, and $\mathbf{\Sigma_{t^*y^*}}$ a $q \times n$ matrix of polychoric correlations. Likewise, if all categorical observed variables were purposefully treated as approximately continuous observed variables in the measurement models, it would simplify the above covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$ to the general covariance matrix for an SEM model with continuous observed variables (see, e.g., Bollen, 1989, p. 325). More specifically, $\mathbf{\Sigma}_u$ reduces to a $(m+n) \times (m+n)$ variance-covariance matrix, $\mathbf{\Sigma}_z$ to a $(p+q) \times (p+q)$ variance-covariance matrix, and $\mathbf{\Sigma}_w$ to a $(p+q) \times (m+n)$ variance-covariance matrix.

Present study

The present study was designed to advance scholarly understanding of the impact of mixed scale observed variables (both continuous and categorical) on parameter estimates, standard errors, and chi-square statistics for CFA and SEM models using MLR and DWLS. Previously, several simulation studies have examined the impact of categorical data on maximum likelihood estimation and on the family of least squares estimators in categorical CFA models. However, what is not yet known is the impact of mixed scale observed variables on the overall quality of parameter estimates, especially factor loadings and structural coefficients, on robust standard error estimates, and on the sensitivity of adjusted chi-square statistics using MLR and DWLS in CFA and SEM models. This study extends previous simulation studies (see, e.g., Bandalos,

2014; Beauducel & Herzberg, 2006; Li, 2016a, 2016b; Maydeu-Olivares, 2017) by pursuing the following overarching research question: Is either of the two estimation methods with its robust corrections consistently better or worse than the other for estimation of model parameters, standard errors, and chi-square statistics across the experimental conditions investigated (i.e., the number of observed variables' categories, the level of distributional asymmetry of the observed variables, and sample size)?

DWLS has been well developed to deal with categorical data, but a combination of continuous and categorical observed variables might potentially degrade its estimation performance. On the contrary, MLR is likely to exhibit superior estimation performance when observed variables have mixed scale types than when they are exclusively categorical. Therefore, the statistical estimation performance of DWLS and MLR is worth further exploring to inform applied researchers and methodologists. Comparing the performance of MLR and DWLS estimation of factor loadings and structural coefficients in the context of mixed scale observed variables remains an open research question. Given abundant findings in the simulation studies (see, e.g., Bandalos, 2014; Beauducel & Herzberg, 2006; Li, 2016a), it is expected that DWLS performs better than MLR at estimating factor loadings of categorical observed variables. However, it is postulated that DWLS and MLR can produce equally good factor loading estimates of continuous observed variables. The rationale behind this formulated expectation is that continuous observed variables are actually treated as continuous in the DWLS estimation where variances, covariances, or polyserial correlations are used for data analysis. In terms of structural coefficients, the expected outcome is that parameter estimates obtained from DWLS are slightly more accurate than those yielded from MLR mainly because previous simulation studies have shown that an undesired low quality of measurement estimates (i.e., underestimation of factor loadings) obtained from MLR could however correct covariance estimates for attenuation, leading to "approximately" unbiased point estimates for inter-factor correlations and structural paths (Beauducel & Herzberg, 2006; Coenders et al., 1997; Li, 2016b).

Relatedly, robust corrections to standard errors and chi-square statistics have recently received considerable attention in applied research, so findings about their performance with mixed continuous and categorical response variables would have practical utility in research practice. A general expectation regarding robust standard error estimates is that the performance of MLR is consistently better than that of DWLS across most conditions. In support of this proposition are findings that MLR yielded more accurate standard error estimates of model parameters than DWLS did in previous simulation studies, irrespective of latent constructs manifested by continuous or categorical observed variables (Li, 2016a, 2016b; Maydeu-Olivares, 2017). Finally, it is anticipated that DWLS outperforms MLR at controlling Type I error rates across CFA and SEM models. Prior research findings suggested that MLR is prone to yielding inflated chi-square statistics in the conditions of small samples with asymmetric ordinal data (Li, 2016b) or moderately non-normal continuous data (Maydeu-Olivares, 2017). That is, test statistics produced by MLR tend to over-reject the hypothesized model when the sample size is not large enough, and non-normal observed variables are continuous or categorical in nature. Moreover, DWLS and MLR may have equivalent performance in maintaining Type I error rates for a measurement model even when the underlying normality assumption was violated in the DWLS estimation (Li, 2016a).

Method

A Monte Carlo simulation study was carried out to determine any effects that different configurations of the number of observed variables' categories, level of distributional asymmetry of the observed variables, and sample size have on parameter estimates, standard errors, and chi-square statistics in one correlated three-factor measurement model and one five-factor structural equation model.

Population models

In latent variable modeling applications, the number of observed variables per factor typically falls within the range of two to five (Ding et al., 1995), and five or more observed variables per factor have rarely appeared in the literature (Gerbing & Anderson, 1985). One literature review across 194 CFA models reported that the median number of first-order factors for a CFA model was 3 in the area of scale development (Jackson et al., 2009). Li (2016b) noted that the median number of factors was 5, and the median number of total observed variables was 18 (with 15 and 24 representing the 25th and 75th percentiles, respectively) across 36 empirical studies using structural equation modeling. In the earlier search for CFA and SEM applications, across 84 empirical studies, the median numbers of latent variables and total observed variables for a hypothesized model were 4 and 20, respectively. More specifically, for those analytical models with a mixture of continuous and categorical observed variables across seven empirical studies, the median numbers of latent variables and total observed variables were 5 and 26, respectively. In addition, Marsh et al. (1998) gave empirical evidence that the accuracy of parameter estimates appeared to nearly reach a maximum when the number of observed variables per factor was 4, improving only trivially when the number increased further.

In line with empirical studies reviewed above, model specifications of the two population models are described next. To represent a "medium-sized" CFA model design from the standpoint of scale development applications, a correlated three-factor measurement model with the first factor having three continuous observed variables and one categorical observed variable (the ratio of continuous observed variables to categorical observed variables = 3:1), the second factor having two continuous observed variables and two categorical observed variables (the ratio = 2:2), and the last factor having one continuous observed variable and three categorical observed variables (the ratio = 1:3) was examined. A five-factor structural equation model with each factor having two continuous observed variables and two categorical observed variables (the ratio = 2:2) was examined as representative of a "medium-sized" SEM model specification frequently encountered in the applied literature. Results from this study are expected to address important generalizability limitations of previous simulation studies, which failed to examine the effect of a mixture of continuous and categorical observed variables in CFA and SEM models.

Population parameters

For the sake of simplicity, homogeneous factor loadings are sometimes used in simulation studies (see, e.g., Anderson, 1996; Flora & Curran, 2004; Forero & Maydeu-Olivares, 2009), but can hardly ever be expected under real-world conditions. In this study, the six factor loadings of continuous observed variables were held at .8, .7, .6, .8, .7, and .7, with corresponding error variances automatically set to .36, .51, .64, .36, .51, and .51, and the other six factor loadings of categorical observed variables were held at .7, .8, .7, .8, .7, and .6, with corresponding error variances automatically set to .51, .36, .51, .36, .51, and .64 under a standardized solution. The inter-factor correlations were all set to .3 in the population, reflecting a realistic and empirical inter-factor correlation coefficient based on the results of previous simulation studies and the applied literature. The correlated three-factor measurement model with a mixture of continuous and categorical observed variables is depicted in Fig. 1a.

Regarding the five-factor SEM model, four factor loadings were fixed at .8, .6, .8, and .6, with corresponding error variances automatically set to .36, .64, .36, and .64 under a standardized solution across all exogenous and endogenous latent variables. The variance-covariance matrix of the two exogenous latent variables (Φ) consisted of two components: (1) the inter-factor correlation was set to .3 in the population, and (2) the two exogenous factor variances were set equal to 1. Considering plausible structural regression coefficients for the population model, common coefficients in standardized solutions range from .1 to .7, and their associated residual variances (i.e., $1 - R^2$) from .2 to .8 in practice and simulation studies. Structural regression coefficients below .1 are, in general, not practically important or statistically significant in applied research (Bandalos, 2006; Ethington, 1987; Hoogland & Boomsma, 1998; Paxton et al., 2001). Therefore, two matrices of structural regression coefficients $\mathbf{B}$ and $\mathbf{\Gamma}$ were each set up as

$$\mathbf{B}=\begin{bmatrix} 0 & 0 & 0 \\ .3 & 0 & 0 \\ .2 & .5 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{\Gamma}=\begin{bmatrix} .4 & .6 \\ .4 & .2 \\ .1 & .1 \end{bmatrix}.$$

The residual variances of the three endogenous latent variables (Ψ) were designated at .336, .436, and .379, in order to obtain standardized structural regression coefficients. The five-factor structural equation model with a mixture of continuous and categorical variables is depicted in Fig. 1b.

Number of categories

Consistent with experimental conditions from previous simulation studies, this study examined the impact of the number of categories in observed variables on statistical estimation. Li (2016b) found that odd-numbered Likert scales with a middle category occur more frequently in empirical studies than even-numbered rating scales. Among 157 psychometric measures, the most common number of response categories was five (39.4%), followed by seven (29.9%), four (10.2%), and six (8.3%). Besides, a total of 647 instruments were examined for the number of response categories in the earlier review of 84 empirical studies. A five-category set (48.5%) was the most widely employed, followed by seven-category (40.6%), six-category (3.7%), and four-category (2.8%), two-category (0.9%), and three-category (0.5%). Specifically, for the models with a mixture of continuous and categorical observed variables, a total of 66 measures were identified. Similarly, the highest percentage of response category was five (62.1%), followed by seven (33.3%), two (3.0%), and six (1.5%). Use of MLR has been considered "legitimate" in published studies when categorical observed variables have five or more response categories without ceiling or floor effects. In order to explore the impact of mixed scale types in borderline usage situations without a clear preferred estimation method, two, three, four, five, six, and seven categories were generated for each categorical observed variable, in combination with two types of distribution shape, as discussed in the next section.

Categorical observed distributions

This study also compared the behavior of MLR and DWLS estimators under varying degrees of normality violation in the categorical observed variables. The presence of non-normality in the form of distribution asymmetry (due to categorization)

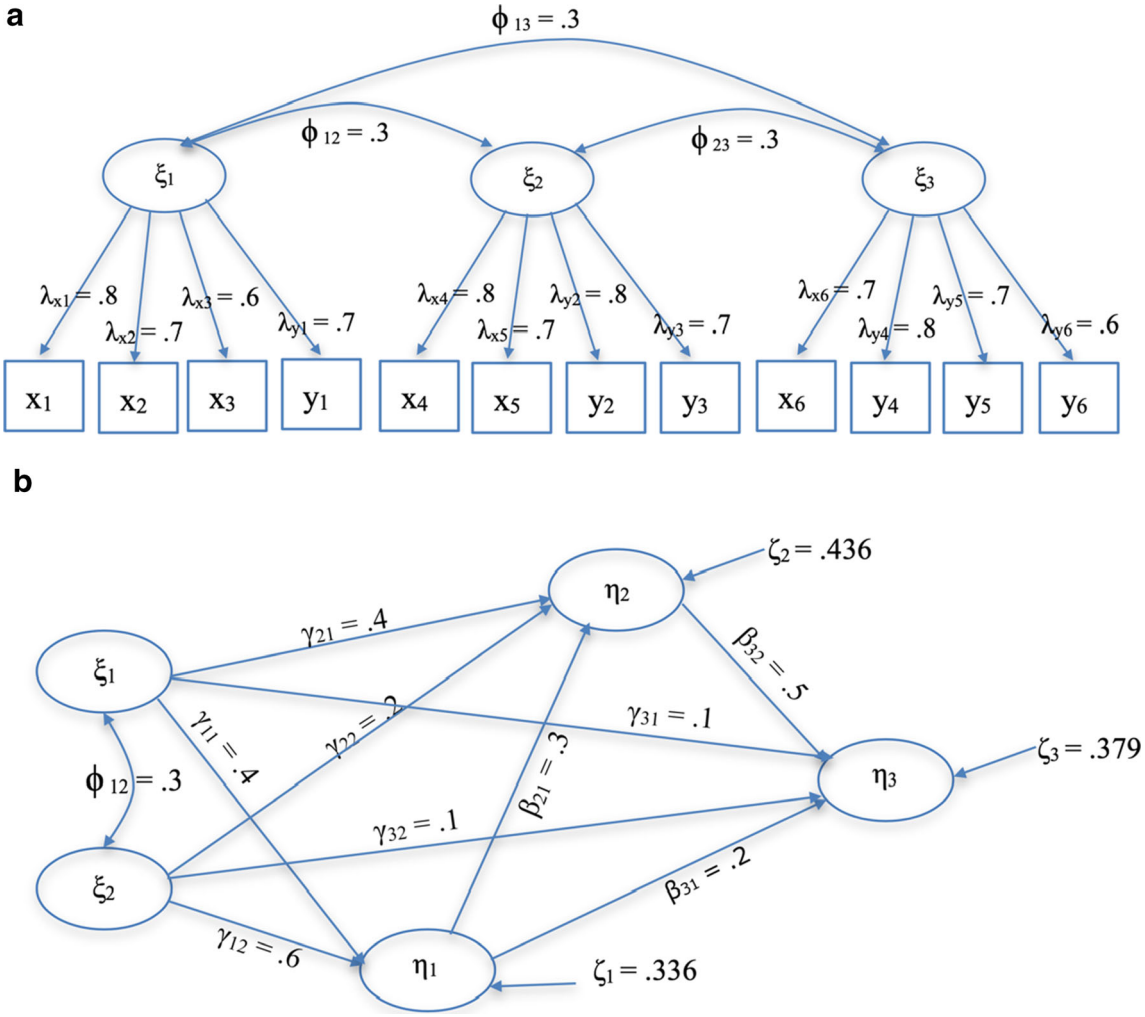

Fig. 1 **a** The postulated correlated three-factor measurement model with standardized coefficients. *Note*. xs are continuous observed variables, ys are categorical observed variables, λs are factor loadings, ξs are exogenous latent variables, and φs are inter-factor correlations **b** The postulated five-factor structural equation model with standardized coefficients. *Note*. Continuous and categorical observed variables of each factor are not depicted here for clarity. ξs are exogenous latent variables, ηs are endogenous latent variables, φ is the inter-factor correlation, βs are structural paths capturing the relationship among endogenous latent variables, γs are structural paths capturing the relationship between exogenous and endogenous latent variables, and ζs are residual variances of the endogenous latent variables

is typical in applied psychometric studies (Micceri, 1989). To this end, two categorical observed distributions were manipulated to vary in symmetry: (1) a symmetric distribution, and (2) a slightly asymmetric distribution. For the symmetric distribution, the middle categories had the highest probabilities; for the slightly asymmetric distributions, the probabilities increased from low to high categories to different degrees.

In the symmetry condition, the threshold value for the two-category observed variables was 0, corresponding to a response proportion of 50% for each category; the threshold values for the three-category observed variables were [−.84, .84], respectively, corresponding to response proportions 20%, 60%, and 20% falling into each category; [−1.282, 0, 1.282] for the four categories, corresponding to response proportions 10%, 40%, 40%, and 10%; [−1.282, −.524, .524, 1.282] for the five categories, corresponding to response proportions 10%, 20%, 40%, 20%, and 10%; [−1.645, −.806, 0, .806, 1.645] for the six categories, corresponding to response proportions 5%, 16%, 29%, 29%, 16%, and 5%; and [−1.645, −.954, −.385, .385, .954, 1.645] for the seven categories, corresponding to response proportions 5%, 12%, 18%, 30%, 18%, 12%, and 5%.

In the slight asymmetry condition, the threshold value for the two-category observed variables was −.553, corresponding to response proportions 29% and 71% falling into each category; the threshold values for the three-category observed variables were [−1.282, −.202], respectively, corresponding to response proportions 10%, 32%, and 58% falling into each category; [−1.645, −1.08, .412] for the four categories, corresponding to response proportions 5%, 9%, 52%, and 34%; [−1.751, −1.341, −.524, .706] for the five categories, corresponding to response proportions 4%, 5%, 21%, 46%, and

24%; [−1.751, −1.341, −1.08, 0, .878] for the six categories, corresponding to response proportions 4%, 5%, 5%, 36%, 31%, and 19%; and [−1.751, −1.341, −1.036, −.613, .496, 1.341] for the seven categories, corresponding to response proportions 4%, 5%, 6%, 12%, 42%, 22%, and 9%. Response proportions of categorical observed variables used in the study are displayed in Fig. 2.

Sample size

Finally, this study was also designed to examine the effect of sample size while utilizing these two estimation methods, because identifying a desired minimum sample size is an important practical consideration in latent variable modeling, which often relies on large-sample assumptions. Sample size is almost universally an experimental factor in Monte Carlo simulation studies (Paxton et al., 2001). A small sample may not only cause inaccurate parameter estimates and unreliable standard errors, but can also produce problems of non-convergence and improper or inadmissible solutions. In addition, when sample size is small, the test statistic for overall model fit is likely not asymptotically chi-square distributed. Applied researchers are therefore interested in determining the sufficient sample size at which parameter estimates will be sufficiently accurate, standard error estimates will be stable, and chi-square model fit statistics will be interpretable.

DiStefano and Hess (2005) reviewed 101 studies using CFA from 1990 to 2002, and reported that the median sample size was 377, and about 19% of the models were tested on samples smaller than 200. Jackson et al. (2009) systematically reviewed 194 CFA models from 1998 to 2006 and found that the median sample size was 389, and about 20% of the models were tested on samples smaller than 200. Li (2016b) reviewed 36 SEM studies and noted that the sample size ranged from 110 to 2512, with a mean of 518. The median sample size was 341, with the 25th and 75th percentiles of 245 and 603 respectively. The earlier review of 84 empirical studies revealed that the median sample size was 275 and the mean sample size was 479, with the 25th and 75th percentiles of 179 and 393, respectively. The highest percentage of sample size category was $n = 201$~500 (57.3%), followed by $n = 101$~200 (25.7%), $n = 501$~1000 (8.8%). For those analytical models with a mixture of continuous and categorical variables, the median sample size was 208, and the mean sample size was 275, with the 25th and 75th percentiles of 121 and 351, respectively. The highest percentage of sample size category was $n = 201$~500 (41.7%), followed by $n = 101$~200 (33.3%), $n < 100$ (16.7%), $n = 501$~1000 (8.3%). Therefore, three different levels were employed to represent small ($N = 200$), medium ($N = 500$), and large ($N = 1000$) sample sizes, bearing more resemblance to those commonly occurring in research practice and frequently manipulated in simulation studies as well (see, e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009).

Data generation and analysis

Two population models, a correlated three-factor measurement model and a five-factor structural equation model, were tested in combination with 36 other experimental conditions characterized by the number of observed variables' categories (2, 3, 4, 5, 6, and 7), categorical observed distribution shape (symmetry and slight asymmetry), and sample size (200, 500, and 1000). A random seed number was set across experimental conditions to initiate random draws from each distribution during data generation. A thousand data sets were generated per experimental condition. The choice of 1000 replications was made with consideration to sampling variance reduction, adequate power, and practical manageability. Model parameters, standard errors, and the chi-square statistic were estimated for each replication using MLR and DWLS. In implementing robust maximum likelihood estimation

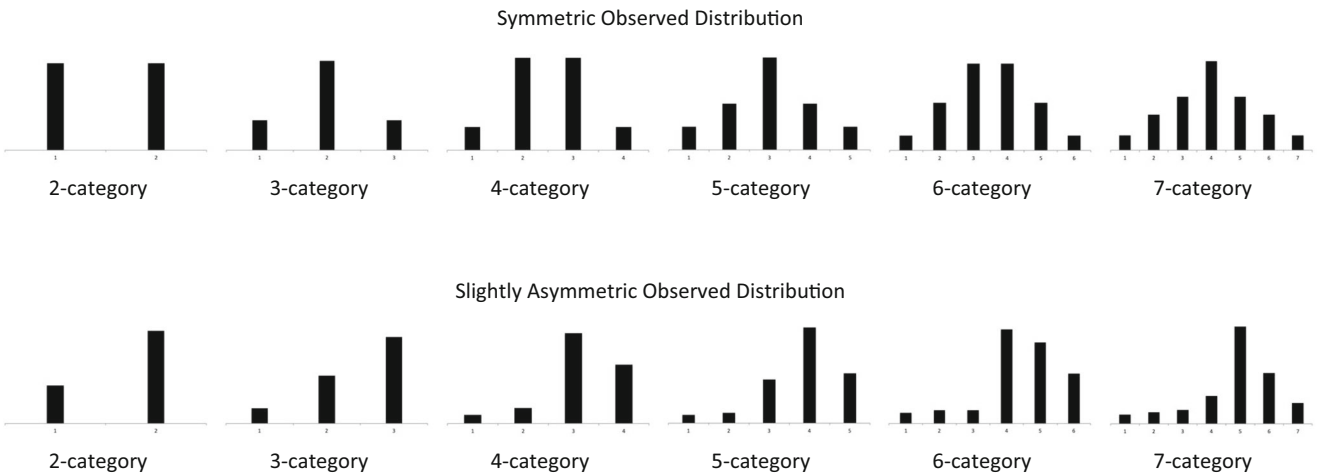

Fig. 2 Response probabilities of categorical observed variables across six different numbers of response categories (from 2 to 7) and the two types of categorical observed distributions (symmetry and slightly asymmetry)

(Estimator = MLR in M*plus*), categorical observed variables were treated as continuous variables (i.e., computing the sample covariance matrix for data analysis). For robust diagonally weighted least squares (Estimator = WLSMV in M*plus*), categorical observed variables were specified as categorical variables and continuous observed variables as continuous (i.e., computing variances, covariances, polychoric correlations, and polyserial correlations as appropriate). Data generation and analysis were performed with M*plus* 7.4 (Muthén & Muthén, 1998–2017). For the sake of simplicity, a standard normal distribution was selected for each latent response variable in the data generation phase (i.e., with zero mean and variance of one), leading to a zero mean structure implied by the model. The multivariate normally distributed data were first generated, then ordinally categorized using prespecified threshold values to induce the desired distributions and response proportions along a standard normal distribution (Muthén & Muthén, 1998–2017). M*plus* code used for data generation and data analysis are available in the supplementary materials.

Outcome variables

Estimation performance was judged according to the following five study outcome variables: (1) average relative bias of parameter estimates, (2) average relative mean squared error of parameter estimates, (3) average relative bias of standard error estimates, (4) relative bias of chi-square statistics, and (5) the model rejection rate associated with the chi-square statistic at an alpha level of .05.

The relative bias (RB) in estimates over the replications and average relative bias (RB_A) across parameter estimates (i.e., factor loadings, inter-factor correlations, or structural paths) were calculated, in tandem, by

$$\mathrm{RB} = \left(\widehat{\theta}_i\right) = \frac{1}{\mathrm{n_r}}\sum_j \left[\frac{\widehat{\theta}_{ij}-\theta_i}{\theta_i}\right] \times 100\%, \mathrm{i} = 1, 2, \ldots, \mathrm{np}; \mathrm{j} = 1, 2, \ldots, \mathrm{nr} \quad (6)$$

and

$$\mathrm{RBA}\left(\widehat{\theta}\right) = \frac{1}{\mathrm{n_p}}\sum_i \mathrm{RB}\left(\widehat{\theta}_i\right), \quad (7)$$

where $\mathrm{RB}(\widehat{\theta}_i)$ denotes the relative bias of the parameter estimate $\widehat{\theta}_i$ over the replications, $\widehat{\theta}_{ij}$ is the parameter estimate of the *i*th population parameter estimate θ_i in the *j*th replication, $\mathrm{n_r}$ is the number of replications in each experimental condition, and $\mathrm{n_p}$ is the number of parameter estimates. An averaged RB with a positive or negative sign indicates overestimation or underestimation of parameter estimates, respectively. An absolute value of RB_A less than 5% can be interpreted as a trivial bias, between 5% and 10% as a moderate bias, and greater than 10% as a substantial bias (Curran et al., 1996).

To quantify the overall quality of parameter estimates, the mean squared error is commonly used in simulation studies because it accounts for both the amount of bias and the sampling variability of parameter estimates (i.e., efficiency). The relative mean squared error (RMSE) and average relative mean squared error ($RMSE_A$) can be defined as

$$\mathrm{RMSE}\left(\widehat{\theta}_i\right)\frac{1}{\mathrm{n_r}} = \sum_j \left[\frac{\widehat{\theta}_{ij}-\theta_i}{\theta_i}\right]^2 \quad (8)$$

and

$$\mathrm{RMSEA}\left(\widehat{\theta}\right)\frac{1}{\mathrm{n_p}} = \sum_i \mathrm{RMSE}\left(\widehat{\theta}_i\right), \quad (9)$$

where $\mathrm{RMSE}(\widehat{\theta}_i)$ denotes the relative mean squared error of the parameter estimate $\widehat{\theta}_i$ over the replications; and other notations have been defined. A small $RMSE_A$ value is suggested as favorable because it indicates better overall quality of parameter estimates, that is, less biased and more precise.

Obtaining accurate standard error estimates is also a primary concern in applied studies. In a similar way, the bias formulations can be used for standard error estimates, which are compared to the standard deviation of the parameter estimates over the replications (also referred to as the empirical standard error) as a proxy for the population standard error. The RB and RB_A for standard error estimates are formulated as

$$\mathrm{RB}\left[\widehat{\mathrm{SE}\left(\widehat{\theta}_i\right)}\right] = \frac{1}{\mathrm{n_r}}\sum_j \left[\frac{\widehat{\mathrm{SE}\left(\widehat{\theta}_i\right)}_j - \mathrm{SD}\left(\widehat{\theta}_i\right)}{\mathrm{SD}\left(\widehat{\theta}_i\right)}\right] \times 100\% \quad (10)$$

and

$$\mathrm{RBA}\left[\widehat{\mathrm{SE}\left(\widehat{\theta}\right)}\right] = \frac{1}{\mathrm{n_p}}\sum_i \mathrm{RB}\left[\widehat{\mathrm{SE}\left(\widehat{\theta}_i\right)}\right], \quad (11)$$

where $\widehat{\mathrm{SE}\left(\widehat{\theta}_i\right)}_j$ is the estimated standard error of parameter $\widehat{\theta}_i$ in the *j*th replication, and $\mathrm{SD}\left(\widehat{\theta}_i\right)$ is the standard deviation of parameter $\widehat{\theta}_i$ over the replications.

Likewise, the performance of chi-square statistics can be assessed by their relative bias. Because the expected value of a chi-square distribution is equal to its degrees of freedom, the relative bias of chi-square statistics over the replications can be expressed as

$$\mathrm{RB}\left(\widehat{\chi^2}_j\right) = \left[\frac{\widehat{\chi^2}_j - \mathrm{df}}{\mathrm{df}}\right] \times 100\% \quad (12)$$

and

$$\mathrm{RB}\left(\widehat{\chi^2}\right) = \frac{\sum_j \mathrm{RB}\left(\widehat{\chi^2}_j\right)}{\mathrm{n_r}}, \mathrm{j} = 1, 2, \ldots, \mathrm{nr}, \quad (13)$$

where $\widehat{\chi^2}_j$ is the estimate of the chi-square statistic in the *j*th replication, *df* is the model degrees of freedom, and $\mathrm{n_r}$ is the number of replications in each experimental condition.

Alternatively, chi-square statistics' performance has often been examined through calculation of the rejection rate at a given nominal alpha level of .05. The rejection rate equals the number of replications for which the chi-square value was greater than the critical value, divided by the number of successfully analyzed replications. The rejection rate of the hypothetical model should, therefore, approximate 5% specified in the population model. Obtained rejection rates lying between 2.5% and 7.5% can be considered acceptable at a nominal alpha level of .05 (Bradley, 1978). A high rate of rejection implies an unintended increased likelihood of concluding against the null hypothesis, whereas a low rate of rejection may indicate a potential compromise of power for rejecting the hypothetical model.

Results

Result presentation begins with information about nonconvergence and inadmissible solutions. Next, in terms of average relative bias ($\mathrm{RB_A}$) and average relative mean squared error ($\mathrm{RMSE_A}$) of parameter estimates, figures and tables are collapsed across same-type parameters within each cell: factor loadings of continuous observed variables, factor loadings of categorical observed variables, inter-factor correlations, structural paths in the two matrices of **B** and **Γ**. Full results of factor loadings for the five-factor SEM model specification are not presented here, because the pattern of loading results was similar to that observed in the correlated three-factor measurement model. Presentations of standard error bias follow the same logic described above. Finally, relative bias of chi-square statistics and rejection rates are reported for each estimation method by sample size, number of observed variables' categories, categorical observed distribution shape, and model type. To free up space in the text, all cross-tabulations of (1) the $\mathrm{RB_A}$ of factor loadings, inter-factor correlations, and structural paths, (2) the $\mathrm{RB_A}$ for robust standard errors of factor loadings, inter-factor correlations, and structural paths, and (3) the relative bias of chi-square goodness of fit statistics and rejection rates associated with the LR test can be found in the supplemental materials.

Nonconvergence and inadmissible solutions

Nonconvergence occurs when the number of iterations exceeds the default maximum number in M*plus* or when the program experiences computational difficulties in optimizing the fit function before the maximum number of iterations has been reached (Muthén & Muthén, 1998–2017). An inadmissible solution (i.e., Heywood case) usually involves out-of-bounds estimates in a statistically converged solution, such as standardized coefficients or correlations larger than 1 in absolute value, or negative residual/error variances.

The rates of nonconvergence across the 36 experimental conditions in measurement models were 0% for both MLR and DWLS. That is, estimation that failed to converge did not occur for MLR or DWLS in CFA models, given varying suboptimal conditions investigated in the study. However, out of 1000 replications, MLR and DWLS both encountered only one case of nonconvergence in SEM models when the number of observed variables' categories was two or three in the smallest sample size $n = 200$. Regarding to inadmissible solutions, neither MLR nor DWLS suffered from improper solutions in the measurement estimation. On the other hand, in the SEM estimation, both MLR and DWLS typically experienced inadmissible solutions when sample size was small (i.e., $n = 200$). Besides, improper solutions most frequently occurred when categorical data were asymmetric and had a small number of observed variables' categories (e.g., 2 or 3) and considerably decreased in frequency when the number of categories increased. Across the smallest sample size ($n = 200$) conditions, MLR and DWLS produced a total of 29 and 47 inadmissible solutions (out of 12,000 replications), respectively. These findings generally suggest that DWLS has a slightly higher probability of producing inadmissible solutions than MLR in a small sample. However, no inadmissible solution was ever identified across all conditions as sample size increased to $n = 500$ or more. The results table is not presented here but available in the supplementary materials.

Any replications producing an inadmissible solution were classified as invalid empirical observations and were removed from subsequent analyses (cf. Boomsma, 2013; Chen et al., 2001; Flora & Curran, 2004; Ferero & Maydeu-Olivares, 2009); that is, all the outcome variables below were computed based on admissible solutions in converged replications for each estimation method. A sensitivity analysis that included the inadmissible solutions was also conducted. Although these analyses produced negligible changes in the reported result tables, the conclusions remained unchanged.

Parameter estimate bias

Factor loadings

Figure 3 presents the average relative bias (RB_A) of factor loadings of continuous and categorical observed variables for MLR and DWLS across all conditions. Factor loadings of categorical observed variables were, on average, underestimated by MLR. When observed variables had five or more categories, they were moderately downward-biased (ranging from −7.04% to −5.17%) in asymmetric categorical data and showed trivial bias (from −4.44% to −2.31%) in symmetric categorical data. In conditions of fewer than five categories (i.e., from 2 to 4 categories), factor loadings of categorical observed variables were substantially downward-biased (from −22.10% to −10.08%) in asymmetric categorical data and showed substantial to moderate bias (from −18.99% to −6.87%) in symmetric categorical data. This negative bias was significantly reduced with increasing number of observed variables' categories, irrespective of sample size. On the other hand, MLR estimation resulted in average relative bias of less than 1% in absolute value for factor loadings of continuous observed variables across all conditions. Conversely, DWLS factor loading estimates of both continuous and categorical observed variables appeared to be only negligibly biased on average (within±.5%), regardless of the number of observed variables' categories, categorical observed distribution shape, and sample size.

An additional evaluation was conducted to assess the effect of the proportion of categorical observed variables (i.e., 25%, 50%, and 75%) in a mixture of continuous and categorical observed variables on factor loading estimation while utilizing MLR and DWLS. Figure 4 presents the RB_A of factor loadings of the three continuous (x_2, x_5, and x_6) and three categorical (y_1, y_3, and y_5) observed variables for MLR and DWLS. The values of these six factor loadings were all .7 in the population model. The proportions of categorical observed variables for the three distinct latent variables ξ_1, ξ_2, and ξ_3 were 25%, 50%, and 75% in the correlated three-factor measurement model, respectively. As shown in Fig. 4, DWLS factor loading estimates of the three continuous and three categorical observed variables appeared to be consistently biased within .5% in absolute value across all conditions, regardless of the proportion of categorical observed variables in a mixture of continuous and categorical data. Similarly, the proportion of categorical observed variables seemed not to exert a major influence toward MLR factor loading estimates across the three categorical observed variables (y_1, y_3, and y_5). However, in terms of continuous observed variable factor loadings, when the proportion of categorical observed variables was 75%, MLR estimation did result in average relative bias of −2% in the conditions of symmetric categorical observed variables with two categories and displayed average relative biases between −1.06% and −4.67% in the conditions of asymmetric categorical data, regardless of sample size, indicating that a high proportion of categorical observed variables in a mixture of continuous and categorical data within the same latent construct can deteriorate the accuracy of MLR loading estimates for continuous observed variables.

As expected, DWLS was consistently superior to MLR for factor loading estimation of categorical observed variables, particularly in the conditions of fewer than five response categories. Also, DWLS generally performed as well as MLR at estimating factor loadings of continuous observed variables and even outperformed MLR when the proportion of categorical observed variables was high in a mixture of continuous and categorical observed variables. These findings indicate that DWLS is better than MLR at estimating factor loadings of mixed scale observed variables in the measurement model.

Tables 1, 2 and 3 display the average relative mean squared error ($RMSE_A$) of factor loadings, inter-factor correlations, and structural paths by the number of observed variables' categories, categorical observed distribution shape, and model type for the two estimation methods. The average relative mean squared error was most pronounced in the conditions where RB_A was also appreciable, and it decreased as sample size and the

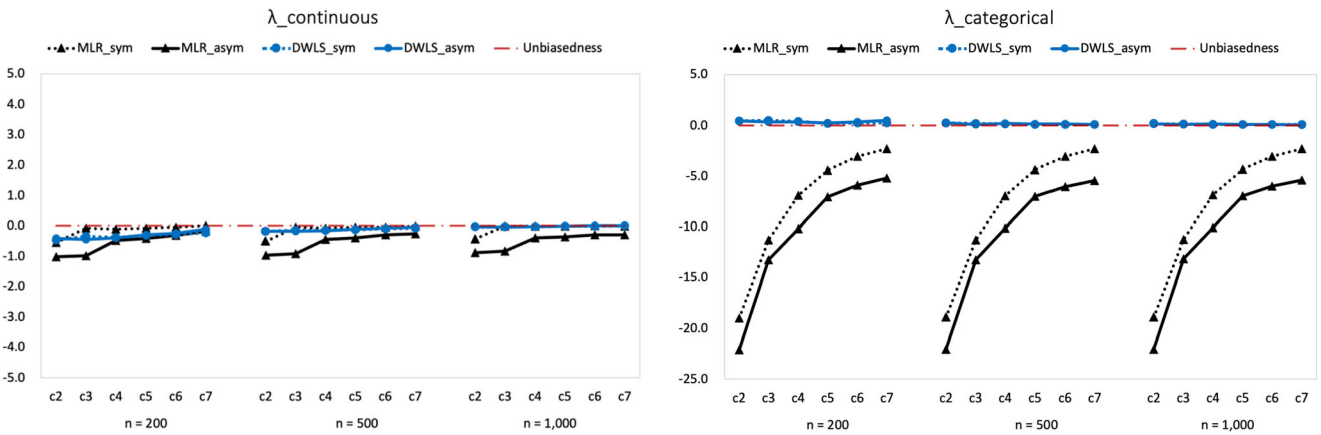

Fig. 3 The average relative bias (RB_A) of factor loadings of continuous and categorical observed variables

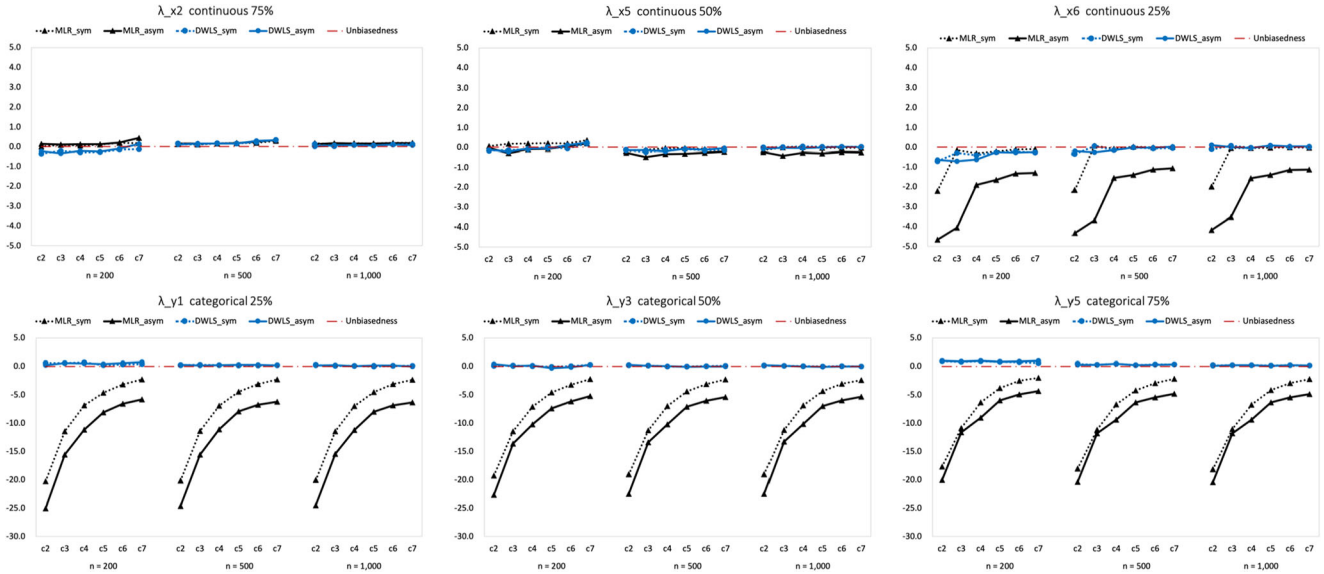

Fig. 4 The average relative bias (RB_A) of factor loadings of the three continuous observed variables (x_2, x_5, and x_6) and three categorical observed variables (y_1, y_3, and y_5)

number of observed variables' categories increased for both estimation methods. The $RMSE_A$ was particularly large for asymmetric categorical variables when MLR was used. Overall, the $RMSE_A$ obtained with DWLS was much smaller than MLR across nearly all categorical data conditions, indicating that DWLS produced less biased and more precise factor loading estimates of categorical observed variables, in comparison to MLR. The two estimation methods had comparable performance for continuous observed variables, marginally favoring MLR, however. The discrepancies between the two estimators decreased as sample size increased.

Inter-factor correlations

The upper panel of Fig. 5 presents the average relative bias (RB_A) of inter-factor correlations for the two model types (i.e., the three-factor correlated measurement model and the five-factor SEM model). Inter-factor correlation estimates of the two model types were, on average, trivially biased for MLR across all conditions (from −1.41% to 1.74%), with high bias typically occurring in the SEM model with a small sample size of 200. Compared to MLR, DWLS inter-factor correlation estimates were generally overestimated to a great extent in the smallest sample size of 200 conditions (i.e., .72%~2.77%

Table 1 The average relative mean squared error ($RMSE_A$) for model parameters (n = 200)

		Measurement model parameters						Structural model parameters					
		MLR			DWLS			MLR			DWLS		
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
sym	2	.0056	.0437	.0889	.0069	.0123	.0931	.0856	.8693	.4930	.0905	.9801	.5624
	3	.0050	.0191	.0800	.0064	.0092	.0810	.0780	.7061	.4117	.0800	.7871	.4507
	4	.0047	.0102	.0787	.0061	.0073	.0801	.0799	.6601	.3775	.0821	.7323	.4186
	5	.0047	.0070	.0773	.0061	.0066	.0783	.0790	.6015	.3479	.0800	.6728	.3888
	6	.0045	.0057	.0751	.0060	.0062	.0762	.0750	.6091	.3500	.0765	.6585	.3781
	7	.0045	.0052	.0737	.0060	.0059	.0748	.0756	.5987	.3447	.0767	.6421	.3723
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
asym	2	.0063	.0582	.0933	.0075	.0145	.0977	.0889	.9141	.5487	.0942	.9960	.5989
	3	.0056	.0250	.0869	.0067	.0100	.0904	.0864	.8282	.4791	.0889	.8350	.4842
	4	.0051	.0167	.0801	.0065	.0083	.0807	.0837	.7365	.4475	.0836	.7600	.4626
	5	.0049	.0109	.0791	.0062	.0073	.0793	.0816	.6637	.3932	.0797	.6881	.4043
	6	.0048	.0092	.0765	.0061	.0067	.0771	.0804	.6682	.3958	.0782	.6810	.3972
	7	.0047	.0082	.0769	.0060	.0063	.0762	.0816	.6545	.3906	.0800	.6566	.3862

Note. Dis. = distribution type, Cat. = number of categories, sym = symmetric distribution, and asym = slightly asymmetric distribution. MLR = robust maximum likelihood and DWLS = diagonally weighted least squares. λ_con = factor loadings of continuous observed variables, λ_cat = factor loadings of categorical observed variables, φ = inter-factor correlations, γ = structural paths between latent endogenous and latent exogenous variables, and β = structural paths among latent endogenous variables

Table 2 The average relative mean squared error ($RMSE_A$) for model parameters (n = 500)

		Measurement model parameters						Structural model parameters					
		MLR			DWLS			MLR			DWLS		
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
sym	2	.0023	.0389	.0358	.0027	.0049	.0370	.0355	.2924	.1695	.0367	.3136	.1876
	3	.0020	.0153	.0327	.0025	.0035	.0333	.0346	.2490	.1429	.0349	.2640	.1525
	4	.0019	.0070	.0319	.0024	.0029	.0323	.0333	.2264	.1296	.0340	.2419	.1421
	5	.0019	.0039	.0312	.0024	.0026	.0315	.0336	.2093	.1214	.0341	.2242	.1308
	6	.0018	.0029	.0305	.0023	.0024	.0308	.0322	.2072	.1205	.0327	.2194	.1291
	7	.0018	.0024	.0298	.0024	.0023	.0301	.0324	.2056	.1197	.0330	.2169	.1279
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
asym	2	.0027	.0525	.0366	.0029	.0056	.0379	.0374	.3170	.1833	.0395	.3285	.1934
	3	.0024	.0206	.0345	.0026	.0039	.0353	.0356	.2740	.1588	.0365	.2770	.1602
	4	.0021	.0128	.0329	.0025	.0032	.0332	.0356	.2551	.1516	.0361	.2609	.1575
	5	.0020	.0073	.0317	.0024	.0029	.0317	.0343	.2313	.1353	.0345	.2326	.1359
	6	.0019	.0059	.0313	.0024	.0026	.0311	.0335	.2295	.1342	.0330	.2277	.1345
	7	.0019	.0053	.0306	.0024	.0025	.0303	.0340	.2266	.1308	.0340	.2238	.1289

Note. Dis. = distribution type, Cat. = number of categories, sym = symmetric distribution, and asym = slightly asymmetric distribution. MLR = robust maximum likelihood and DWLS = diagonally weighted least squares. λ_con = factor loadings of continuous observed variables, λ_cat = factor loadings of categorical observed variables, φ = inter-factor correlations, γ = structural paths between latent endogenous and latent exogenous variables, and β = structural paths among latent endogenous variables

in the measurement model and 2.6%~3.32% in the SEM model). This bias dropped with increasing sample size. Specifically, the values of RB_A obtained from DWLS were between .1% and 1.25% in the sample size n = 500 conditions and between −.08% and .67% in the sample size n = 1000 conditions. Generally speaking, the RB_A for inter-factor

Table 3 The average relative mean squared error ($RMSE_A$) for model parameters (n = 1000)

		Measurement model parameters						Structural model parameters					
		MLR			DWLS			MLR			DWLS		
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
sym	2	.0012	.0374	.0172	.0014	.0024	.0178	.0178	.1417	.0802	.0184	.1551	.0886
	3	.0010	.0139	.0156	.0013	.0017	.0159	.0169	.1226	.0669	.0170	.1339	.0724
	4	.0009	.0058	.0154	.0012	.0015	.0155	.0163	.1115	.0619	.0165	.1218	.0683
	5	.0009	.0029	.0149	.0012	.0013	.0151	.0158	.1037	.0582	.0160	.1127	.0632
	6	.0009	.0019	.0147	.0012	.0012	.0148	.0155	.1040	.0568	.0157	.1123	.0615
	7	.0009	.0015	.0146	.0012	.0012	.0147	.0156	.1026	.0563	.0157	.1108	.0611
Dis. Cat.		λ_con	λ_cat	φ	λ_con	λ_cat	φ	φ	γ	β	φ	γ	β
asym	2	.0015	.0506	.0175	.0015	.0028	.0181	.0183	.1541	.0869	.0193	.1646	.0935
	3	.0013	.0189	.0167	.0013	.0019	.0170	.0176	.1356	.0769	.0179	.1378	.0783
	4	.0010	.0115	.0156	.0012	.0016	.0158	.0171	.1239	.0694	.0169	.1284	.0728
	5	.0010	.0061	.0154	.0012	.0014	.0154	.0167	.1159	.0651	.0165	.1191	.0663
	6	.0010	.0048	.0154	.0012	.0013	.0153	.0164	.1156	.0640	.0161	.1180	.0651
	7	.0010	.0041	.0152	.0012	.0013	.0149	.0163	.1120	.0623	.0160	.1114	.0617

Note. Dis. = distribution type, Cat. = number of categories, sym = symmetric distribution, and asym = slightly asymmetric distribution. MLR = robust maximum likelihood and DWLS = diagonally weighted least squares. λ_con = factor loadings of continuous observed variables, λ_cat = factor loadings of categorical observed variables, φ = inter-factor correlations, γ = structural paths between latent endogenous and latent exogenous variables, and β = structural paths among latent endogenous variables

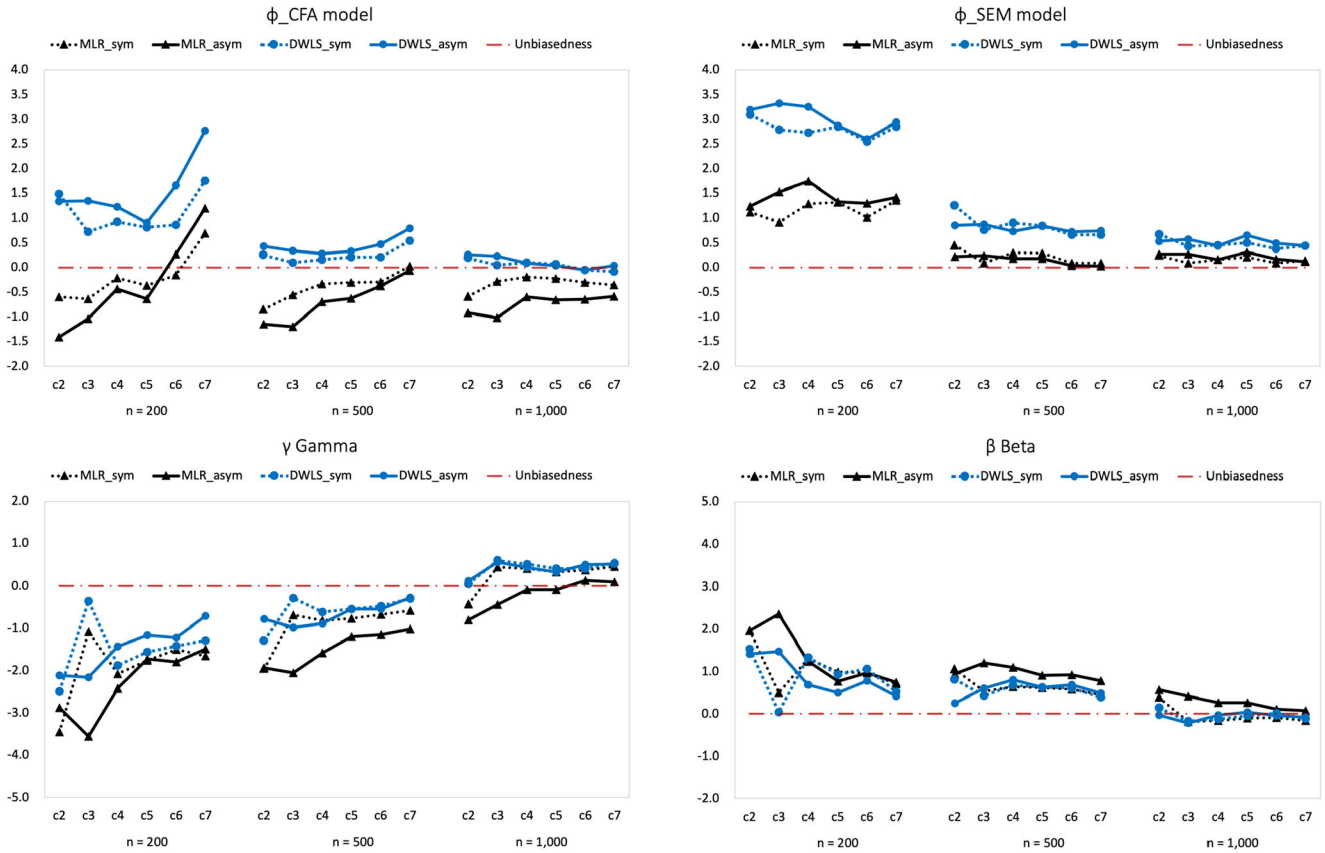

Fig. 5 The average relative bias (RB_A) of inter-factor correlations and structural paths

correlations in the measurement model was relatively smaller for DWLS than for MLR across nearly all conditions, except some conditions with the smallest sample size of 200. In the SEM model, the RB_A for inter-factor correlations was consistently smaller for MLR than for DWLS across all conditions, particularly for the conditions with the smallest sample size of 200. This finding implies that MLR generally yields more accurate inter-factor correlation estimates than DWLS when mixed scale observed variables are used in an SEM model or a measurement model with a small sample size of 200; however, DWLS can perform better than MLR at estimating inter-factor correlations in a measurement model with a sample size of 500 or more. It is noted that these differences between the two estimation methods, not more than 2%, may be regarded as negligible.

Regarding the overall quality of the estimated inter-factor correlations, the $RMSE_A$ decreased with increasing sample size but was not clearly a function of the number of observed variables' categories. No sizable difference was observed between the two estimation methods, in terms of $RMSE_A$. Although MLR, on average, produced more accurate inter-factor correlation estimates than DWLS, MLR estimates exhibited relatively larger variations (i.e., less efficient) than DWLS counterparts. For example, a relatively high positive bias in estimating inter-factor correlations was obtained using DWLS in the SEM model, but the overall quality of DWLS estimates was slightly better than that of MLR estimates in the conditions of asymmetric categorical data with more than five observed variables' categories. This finding suggests that DWLS and MLR can yield comparable inter-factor correlation estimates in a measurement model with mixed scale observed variables, in terms of the overall quality (accuracy and precision) of inter-factor correlation estimates.

Structural paths

The lower panel of Fig. 5 presents the average relative bias (RB_A) of structural paths for gamma and beta coefficients. Averaging over the structural path estimates, the bias in structural coefficients (including structural paths in both matrices **B** and **Γ**) obtained with the two estimators was, on average, trivially biased. The amount of bias in gamma and beta estimates was within 1% either in a positive or negative direction for both estimators when sample size *n* reached 500, except some conditions in MLR with asymmetric categorical observed variables or with symmetric categorical data having only two categories. In the conditions with a sample size of 200, negative bias in gamma estimates (−3.57%~−1.08% for MLR and −2.49%~−.37% for DWLS) and positive bias in beta estimates (.49%~2.35% for MLR and .04%~1.53% for

DWLS) generally decreased with increasing the number of observed variables' categories. Comparing the two estimators, there was no remarkable distinction in terms of the absolute value of RB_A. MLR, however, consistently introduced a relatively larger amount of bias into the estimates of both structural paths than did DWLS across nearly all conditions, except a relatively smaller amount of bias in MLR structural path estimates between endogenous and exogenous latent variables of the $\boldsymbol{\Gamma}$ matrix in the conditions with a sample size of 1000. However, the estimation differences between the two methods became negligible as sample size increased. To sum up, these findings support the aforementioned expectation that DWLS produces more accurate structural path estimates, both gamma and beta coefficients, than MLR in an SEM model with mixed scale observed variables.

The $RMSE_A$ for structural path estimates decreased with increasing sample size and the number of observed variables' categories. That is, structural coefficient estimation improved when sample size and the number of observed variables' categories increased. In general, there was no remarkable evidence suggesting that one estimator is inferior or superior to the other, in terms of $RMSE_A$. Although DWLS, on average, produced more accurate structural path estimates than did MLR across nearly all conditions, DWLS displayed greater variability in estimates than MLR. For example, in spite of a relatively high bias in estimating MLR structural paths in SEM models, the overall quality of MLR estimates was slightly better than that of DWLS estimates across nearly all conditions. This finding shows that MLR and DWLS can perform equivalently well at estimating structural paths in an SEM model with mixed scale observed variables, in terms of the overall quality (accuracy and precision) of structural path estimates.

Standard error bias

As shown in the upper panel of Fig. 6, standard errors of factor loadings exhibited, on average, slight downward bias (from −4.88% to −.72%) for DWLS with the smallest sample size n =200, reflecting that robust corrections to standard errors were not upward-adjusted enough to compensate for the loss of efficiency in the sample size of 200 conditions. Not surprisingly, this standard error underestimation improved (−2.61%~.40%) when sample size increased. Standard errors of factor loadings obtained from DWLS were constantly more biased for categorical observed variables than for continuous observed variables across all conditions. In contrast, a relatively small amount of bias was observed in MLR estimation across all simulation conditions (ranged from −2.26% to 1.79%). When the sample size reached $n = 500$, the differences in standard error bias between the two estimators greatly shrank. As soon as the sample size increased up to $n = 1000$, there was no notable discrepancy between MLR and DWLS in estimating standard errors of loading estimates for either continuous or categorical observed variables.

In the middle panel of Fig. 6, robust standard errors of inter-factor correlations exhibited, on average, moderate downward bias (from −9.15% to −2.71%) for DWLS with the smallest sample size of 200. This negative bias was only marginally lessened when sample size increased (from −8.35% to −2.71% in the sample size $n = 500$ conditions and from −4.11% to .18% in the sample size $n = 1000$ conditions). For MLR estimation, a small amount of bias was evident across all simulation conditions (from −3.80% to .91%), except for a relatively high bias (from −6.18% to −3.01%) in the SEM model with a sample size of 500. This exception is because nearly the same magnitude of negative bias was observed for the two sample size levels ($n = 200$ and 500), but relatively small values of empirical standard errors were obtained when sample size was 500, which produced a larger average relative bias in the sample size of 500 conditions. Therefore, the performance in estimating standard errors of inter-factor correlations was considered equivalent between DWLS and MLR when the sample size was equal to 1000. That is, standard errors of inter-factor correlations obtained from MLR performed slightly better than those from DWLS (by 1%~2%) in the conditions of sample size $n = 500$ and were considerably less biased than those from DWLS (by 4%~6%) in the sample size $n = 200$ conditions.

Similarly, robust standard errors of structural paths exhibited, on average, moderate downward bias (from −8.70% to −4.03%) for DWLS with the smallest sample size of 200, as depicted in the lower panel of Fig. 6. This negative bias was significantly reduced with increasing sample size (from −2.86% to .15% in the sample size $n = 500$ conditions and from −1.44% to .46% in the sample size $n = 1000$ conditions). A small amount of bias was, again, observed for MLR estimation across all simulation conditions (from −4.01% to .84%), with some relatively high bias typically occurring in the model with a small sample size of 200. The performance of DWLS in estimating standard errors of structural paths was apparently inferior to that of MLR in the conditions with a sample size of 200, trivially weaker than that of MLR in the sample size $n = 500$ conditions, but as good as that of MLR when the sample size reached 1000. On the whole, the performance of MLR surpassed that of DWLS in estimating standard errors across most conditions, as predicted. However, there was no remarkable distinction between MLR and DWLS in the conditions with a large sample size of 1000.

Chi-square statistics

Figure 7 presents findings for relative bias of chi-square goodness-of-fit statistics and rejection rates associated with the likelihood ratio (LR) test for the two model types, respectively. Generally speaking, the corrected chi-square test

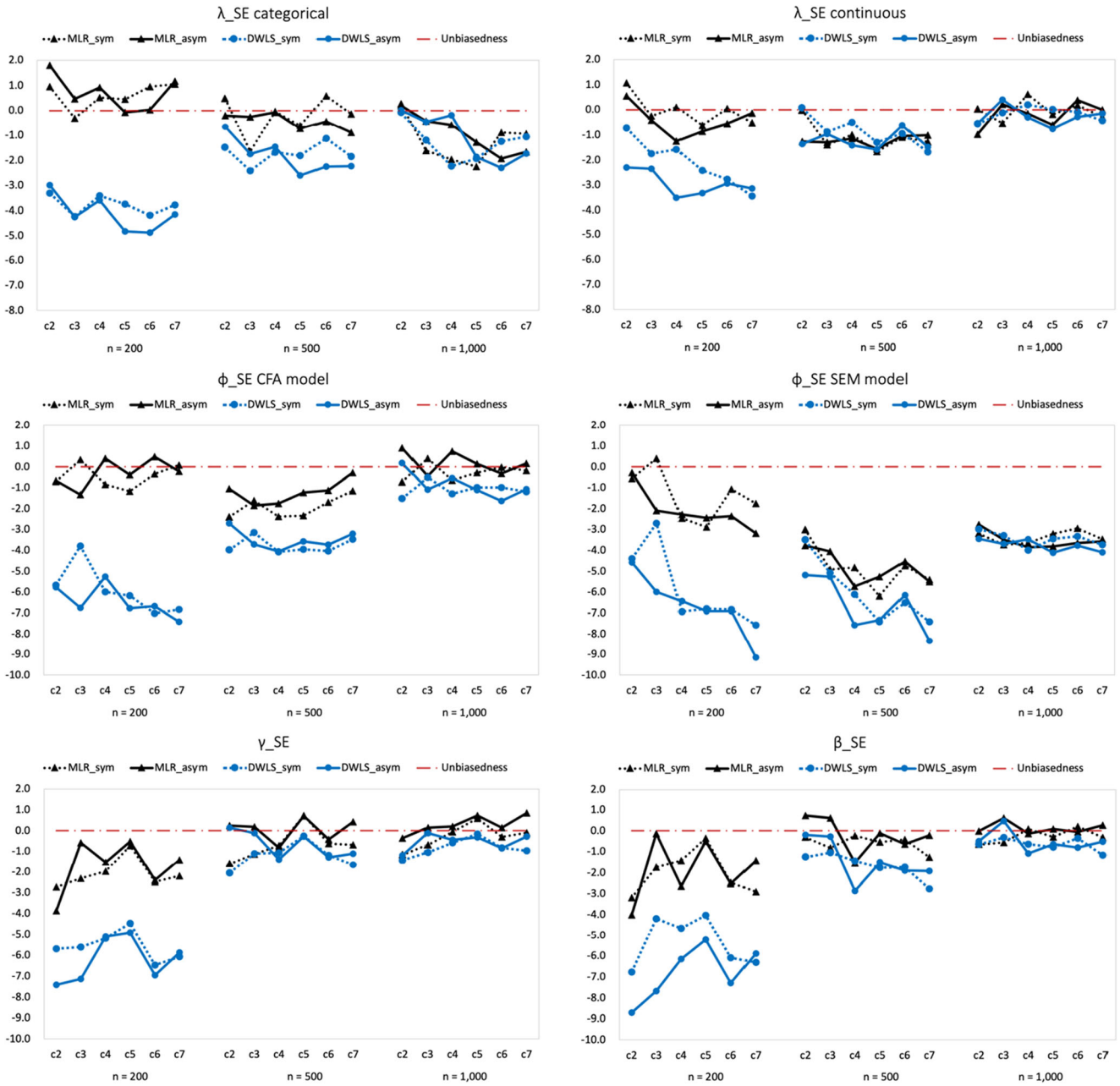

Fig. 6 The average relative bias (RB_A) for standard errors of model parameters (factor loadings, inter-factor correlations, and structural paths)

statistics tended to be positively biased across all experimental conditions in the upper panel of Fig. 7, with the MLR correction being particularly unstable. Chi-square statistics obtained from DWLS were minimally biased (−.66%~3.23%) across all conditions. In general, MLR was prone to yield moderately inflated chi-square statistics (4.26%~8.15%) in the conditions with a small sample size of $n = 200$ and suffered from moderate to substantial inflation (4.14%~9.37%) in the conditions with asymmetric categorical observed variables having fewer than four categories. It is not surprising that the former inflation was reduced with increasing sample size and the latter inflation dropped with increasing the number of observed variables' categories for both estimation methods. Compared to MLR, the bias in chi-square statistics was consistently smaller for DWLS, leading to a better performance of DWLS on the evaluation of overall model fit.

For DWLS, the empirical Type I error rates of testing overall model fit for measurement and SEM models were all within the range of .025 (lower bound) and .075 (upper bound), very close to the nominal Type I error (alpha =. 05), as shown in the lower panel of Fig. 7. On the other hand, MLR appeared to be systematically inferior in controlling the Type I error rate for testing overall model fit across nearly all measurement models, unless observed variables had at least four categories with the largest

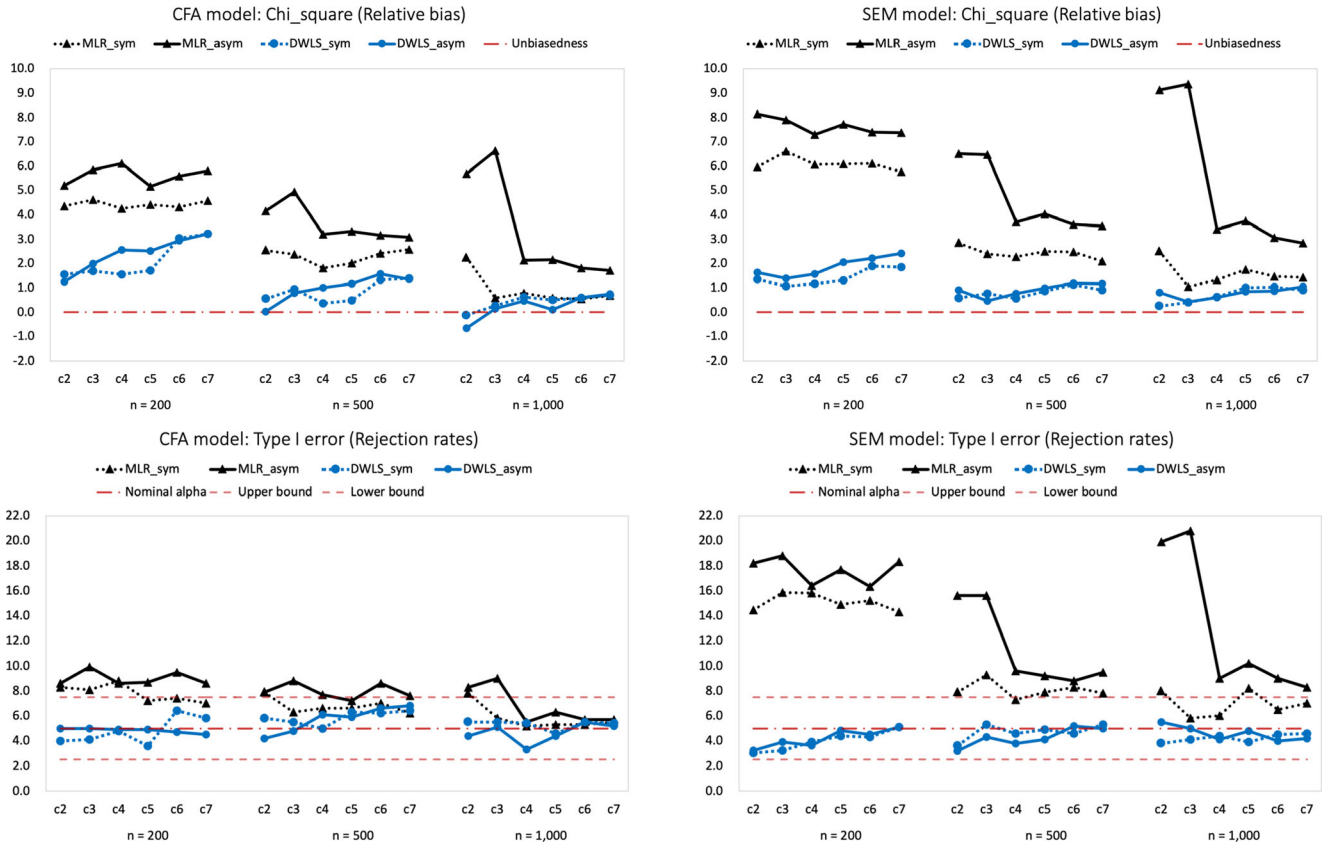

Fig. 7 The relative bias (RB) of chi-square statistics and rejection rates associated with the LR test

sample size of n =1000 and or symmetric observed variables had five or more categories. MLR even performed very poorly for the SEM models at the smallest sample size of $n = 200$ or in the conditions with asymmetric observed variables made up of two or three categories. More specifically, MLR seemed to over-reject the hypothesized SEM model much beyond expectation (approximately 3 to 5 times as often as the nominal Type I error rate would suggest), indicating that robust chi-square statistics obtained from MLR may have been substantially inflated in the case of a small sample size of $n = 200$ or in the case of asymmetric data with three or fewer categories. However, the degree of inflation diminished as sample size or the number of observed variables' categories increased. In general, in terms of controlling the Type I error rate in both measurement and SEM models, the performance of DWLS surpassed that of MLR on the evaluation of overall model fit across all conditions, as predicted.

Discussion

The main purpose of this study was to extend previous simulation studies on structural equation models with categorical observed variables by comparing the relative performance of the MLR and DWLS estimation methods in structural equation models with a mixture of continuous and categorical observed variables across different experimental conditions (i.e., numbers of observed variables' categories, categorical observed distributions, and sample sizes). Statistical estimation of SEM models with mixed scale observed variables has been examined repeatedly in the Bayesian framework, but never systematically for the two most popular frequentist methods. The primary contribution of this article is to compare the estimation performance of MLR and DWLS in a model with mixed continuous and categorical observed variables, which further informs research practice and our knowledge about the importance of selecting appropriate estimators in different research conditions. Findings from this study are of particular interest to applied researchers and methodologists in the social and behavioral sciences. Several general findings are summarized and discussed in this section.

First, results indicated neither estimation method was subject to convergence failures across any tested experimental condition, except two cases of nonconvergence in SEM models where the number of observed variables' categories (i.e., 2 or 3) and sample size (i.e., $n = 200$) were both small. However, in the SEM estimation, both MLR and DWLS produced a few inadmissible solutions in the smallest sample size $n = 200$ conditions, with a slightly higher likelihood of producing inadmissible solutions in the DWLS estimation. This finding is in line with results from previous simulation studies

in which inadmissible solutions more frequently occur with small sample sizes (Herzog et al., 2007; Li, 2016b; Rhemtulla et al., 2012; Forero et al., 2009). Therefore, it is generally recommended that a sample size of 500 observations is sufficient to produce converged and admissible solutions for the both estimation methods in an SEM model with a mixture of continuous and categorical variables, regardless of the number of observed variables' categories, categorical observed distribution shape, and sample size.

Second, this study replicated previous findings that DWLS consistently surpasses MLR in estimating factor loadings of categorical observed variables (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009; Li, 2016a; Rhemtulla et al., 2012), while MLR and DWLS factor loading estimates of continuous observed variables were comparable, in terms of biasness and precision. Interestingly, on the basis of this simulation study, a clear superiority of DWLS over MLR in estimating factor loadings for a measurement model with both continuous and categorical observed variables was evident, irrespective of number of observed variables' categories, categorical observed distribution shape, and sample size. Additionally, a distinct finding from this study adding value to the existing literature is that DWLS outperforms MLR at estimating factor loadings of continuous observed variables with a high proportion of categorical observed variables (i.e., 75%) within the same latent construct. Another important contribution in this paper is to clearly demonstrate the practical advantage of DWLS over MLR at estimating factor loadings as long as any categorical observed variable is present in a measurement model, which has never been validated in previous simulation studies. These findings call into question the appropriateness of using MLR estimation and directly translate into practical benefits of using DWLS estimation when there are mixed item response types within the same measurement model.

Third, when a mixture of continuous and categorical observed variables was employed in structural equation models, DWLS generally yielded slightly more biased inter-factor correlations but relatively less biased structural paths across nearly all conditions, compared to MLR. However, DWLS and MLR appeared to be comparable in estimating inter-factor correlations and structural paths, considering relative mean squared error. In order to gain a deeper understanding of this scenario, the value of $RMSE_A$ was partitioned into two components: squared bias and sampling variance. As shown in Table 4, DWLS generally displayed a higher bias in inter-factor correlation estimates than MLR (about 4~6 times more) in the combination of asymmetric categorical data and sample size $n = 200$, despite lower $RMSE_A$ values obtained from DWLS estimation. Specifically, DWLS produced more biased, but less variable, inter-factor correlation estimates, indicating that the estimates obtained in any given replications are likely to be close to each other, but relatively far from the population value. DWLS, on the other hand, exhibited higher $RMSE_A$ values than MLR due to a great deal of sampling variance in the condition with asymmetric data and sample size $n = 200$, despite a generally smaller amount of bias in structural path estimates produced by DWLS estimation (about 1.5~4 times less) in Table 4. Although DWLS produced less biased structural path estimates, it had relatively higher sampling variance than MLR, indicating that its estimates across replications tend to be not far from the population value, but are spread out to some degree. Such occurrences disappeared as sample size increased, reflecting that a large sample size can wash out the difference between two estimation methods. Overall, it is advisable to use either MLR or DWLS in an SEM model with a mixture of continuous and categorical variables if capturing the structural relationships (inter-factor correlations and structural paths) is the primary concern for applied researchers in their study.

Fourth, after applying robust corrections to standard errors, it was observed that MLR consistently gave less biased standard error estimates than DWLS across most conditions. This finding contradicts the existing literature, showing that DWLS performs better at estimating the standard errors of inter-factor correlations than MLR does (Rhemtulla et al., 2012; Yang-Wallentin et al., 2010), partly because SEM models with mixed continuous and categorical observed variables were analyzed in this study. The accuracy of standard error estimates improved with increasing sample size. That is, robust standard errors obtained from MLR and DWLS are comparable when sample size reaches 500 or more, which is similar to the conclusion of Li (2016b) that at least a medium sample size of $n = 500$ is needed to obtain stable standard error estimates from DWLS. However, robust standard error estimates were not so sensitive to the shape of categorical observed distributions and the number of observed variables' categories in this study. More importantly, applied researchers should be cautious about unstable standard error estimates and potentially unreliable model parameter inference when using DWLS in the sample size of 200 conditions, particularly for small structural coefficients. This study also contributes to the current literature by highlighting potential pitfalls of using DWLS in a small sample, such as biased standard errors of model parameters. An additional analysis of comparing DWLS and MLR standard error estimates of model parameters is warranted in research practice.

Fifth, in the evaluation of overall model fit using robust chi-square statistics, DWLS had empirical rejection rates within the acceptable range of 2.5% and 7.5%, close to the nominal Type I error $\alpha = .5$, whereas MLR tended to over-reject the hypothesized SEM models more often than expected. MLR produced unacceptable rejection rates due to moderate to substantial inflation of the chi-square statistics. Acceptable rejection rates associated with MLR chi-square statistics might not be observed until the sample size reaches

Table 4 Partition of the average relative mean squared error ($RMSE_A$) for structural model parameters (n = 200)

			Structural model parameters WLSMV $RMSE_A$								
			φ			γ			β		
Dis.	Cat.	Est.	Squared bias	Sampling variance	$RMSE_A$	Squared bias	Sampling variance	$RMSE_A$	Squared bias	Sampling variance	$RMSE_A$
asym	2	MLR	0.000154	0.088746	0.0889	0.000829	0.913271	0.9141	0.000380	0.548320	0.5487
		DWLS	0.001018	0.093182	0.0942	0.000380	0.995620	0.9960	0.000199	0.598701	0.5989
	3	MLR	0.000234	0.086166	0.0864	0.001274	0.826926	0.8282	0.000552	0.478548	0.4791
		DWLS	0.001102	0.087798	0.0889	0.000552	0.834448	0.8350	0.000216	0.483984	0.4842
	4	MLR	0.000303	0.083397	0.0837	0.000586	0.735914	0.7365	0.000151	0.447349	0.4475
		DWLS	0.001056	0.082544	0.0836	0.000151	0.759849	0.7600	0.000046	0.462554	0.4626
	5	MLR	0.000177	0.081423	0.0816	0.000299	0.663401	0.6637	0.000059	0.393141	0.3932
		DWLS	0.000824	0.078876	0.0797	0.000135	0.687965	0.6881	0.000024	0.404276	0.4043
	6	MLR	0.000169	0.080231	0.0804	0.000324	0.667876	0.6682	0.000094	0.395706	0.3958
		DWLS	0.000676	0.077524	0.0782	0.000151	0.680849	0.6810	0.000061	0.397139	0.3972
	7	MLR	0.000202	0.081398	0.0816	0.000225	0.654275	0.6545	0.000055	0.390545	0.3906
		DWLS	0.000864	0.079136	0.0800	0.000052	0.656548	0.6566	0.000017	0.386183	0.3862

Note. Dis. = distribution type, Cat. = number of categories, Est. = estimation methods, and asym = slightly asymmetric distribution. MLR = robust maximum likelihood and DWLS = diagonally weighted least squares. φ = inter-factor correlations, γ = structural paths between latent endogenous and latent exogenous variables, and β = structural paths among latent endogenous variables

1000. This finding seemingly raises another empirical question about the necessity of using some supplemental fit index (e.g., RMSEA, CFI, TLI) as an alternative to evaluate model plausibility when MLR is employed in applications. Future research assessing the performance of mean- and variance-adjusted maximum likelihood (MLMV) on overall model inference is still suggested, given that MLMV was considered the optimal choice for goodness-of-fit testing under non-normal data conditions (Maydeu-Olivares, 2017). Yet, the use of DWLS is practically effective when applied researchers use the likelihood ratio test to assess overall model fit across all conditions investigated in this study.

Finally, there are numerous combinations to manipulate in a single simulation study, but one can only focus on certain factors of particular interest to make the research design feasible and manageable. One drawback of carrying out a Monte Carlo study is that results are conditional on the simulation design. This study shares the same limitation as all Monte Carlo simulation studies, in that generalizations are constrained by the specification of the experimental conditions employed in this study. Although the trends found within the experimental conditions may behave in a predictable way outside the scope of the study's design, such predictions can become less reliable as conditions depart further from those investigated in the study. Generalizing any results beyond the scope of this study should be done with caution. For example, the present study does not pursue the potential effects of model specification errors; a worthy topic for future research is to evaluate the performance of different estimation methods on chi-square goodness of fit statistics and ad hoc fit indices under different levels of model misspecification (see, e.g., Bandalos, 2014; Yang-Wallentin et al., 2010). Moreover, as the Bayesian approach has recently been applied to many complex SEM models in small samples, a natural extension of this study might compare the relative performance of Bayesian and frequentist methods for statistical estimation of SEM models (see, e.g., Holtmann et al., 2016; Lee & Song, 2004; Liang & Yang, 2014) when mixed scale observed variables are specified. Lastly, this study did not empirically examine the effects of violation of normality assumption in the latent variables or continuous observed variables. Although previous simulation studies have suggested that MLR exhibits mild robustness to non-normal observed data (Asparouhov & Muthén, 2005; Maydeu-Olivares, 2017; Yuan et al., 2005) and DWLS is deemed robust against moderate violations of underlying normal distributions (Coenders et al., 1997; Flora & Curran, 2004; Liang & Yang, 2014), further investigation into various violations of normality assumption is suggested to deepen our understanding of statistical estimation performance under these conditions.

Conclusions

Many applications of measurement models in social and behavioral research use a mixture of continuous and categorical observed variables. DWLS has been intentionally developed to deal with categorical data, and existing scholarship on categorical CFA and SEM models suggesting DWLS with its robust corrections as a viable estimator has made important

advances, but its performance when continuous and categorical observed variables are mixed within a measurement model was previously not well understood. MLR, on the other hand, was expected to produce better statistical estimates when observed variables had mixed scale types than when they were exclusively categorical. By carrying out a Monte Carlo simulation study, this first attempt was made to address research gaps in the established literature about the impact of mixed item scale types (continuous and categorical observed variables) on parameter estimates (factor loadings, inter-factor correlations, and structural paths), standard errors, and chi-square statistics using the two popular estimation methods MLR and DWLS.

This study provides evidence that DWLS performs better than MLR in many conditions. DWLS merely requires a small sample size (e.g., $n = 200$) for the recovery of population factor loadings and structural coefficients, and to evaluate overall model fit using robust chi-square goodness of fit statistics in an SEM model with a mixture of continuous and categorical observed variables. Although this study suggests that DWLS can be generally recommended in research practice, it is worthwhile to point out that DWLS indeed has its own limitations. For example, this study has revealed that DWLS may not produce stable standard errors estimates unless at least a medium sample (e.g., $n = 500$) is used, which may potentially undermine the reliability of statistical inference for parameter estimates. In addition, some applied researchers may be limited in the choice of software programs or by estimation availability of certain software programs that they are familiar with. Compared to the popularity of maximum likelihood in applications, DWLS is currently implemented in M*plus*, LISREL, SAS, and R but unavailable in some statistical programs (e.g., EQS, Amos, and STATA).

Generally speaking, the moderate to substantial underestimation of factor loadings of categorical observed variables, and considerable inflation of chi-square goodness of fit statistics make MLR less attractive and favorable in an SEM model with mixed scale observed variables. However, the small amount of bias in structural coefficient and standard error estimates makes MLR practically recommendable when applied researchers are primarily concerned with structural relationships among latent variables. It is also worth noting that once applied researchers confront the problem of missing data or conduct the analysis of latent variable interaction in an SEM model, MLR with full information estimation is considered as a promising approach to handling missing data or moderating effects properly. Yet, the treatment of missing data and latent moderated structures in DWLS remains technically underdeveloped (Muthén & Muthén, 1998–2017).

In closing, this study has shifted the previous simulation focus on categorical data by exploring a mixture of continuous and categorical observed variables in the latent variable modeling framework. This distinction is theoretically and practically important because the promise of this study can extend this line of inquiry and vertically advance scholarly understanding of the performance of these two frequentist estimation methods in the context of structural equation modeling with a mixture of continuous and categorical variables.

It is expected to inform the work of applied researchers in suboptimal circumstances where observed variables of mixed scale types are needed, emphasizing the distinct advantage of latent variable modeling techniques.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01547-z.

References

Anderson, R. D. (1996). An evaluation of the Satorra-Bentler distribution misspecification correction applied to the McDonald fit index. *Structural Equation Modeling, 3,* 203–227.

Asparouhov, T., & Muthén, B. O. (2005). Multivariate statistical modeling with survey data. Retrieved from: http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf

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

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood. *Structural Equation Modeling, 21*, 102–116.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. Mueller (Eds.), Structural equation modeling: A second course (pp. 385–426). Information Age.

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

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

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling, 20,* 518–540.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 58,* 430–450.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29,* 468–508.

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.

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.

Diemer, M. A., & Li, C. (2012). Longitudinal roles of precollege contexts in low-income youths' postsecondary persistence, *Developmental Psychology, 48,* 1686–1693.

Ding, L., Velicer, W. F., Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2,* 119–144.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23,* 225–241.

Ethington, C. A., (1987). The robustness of LISREL estimates in structural equation models with categorical variables. *The Journal of Experimental Education, 55,* 80–88.

Fahrmeir, L., & Raach, A. (2007). A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika, 72,* 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(4)*, 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.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research, 20,* 225–271.

Gueorguieva, R. V., & Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine, 25,* 1307–1322.

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A Comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research, 51*, 661–680

Hom, P. W., Tsui, A. S., Wu, J. B., Lee, T. W., Zhang, A. Y., Fu, P. P., & Li, L. (2009). Explaining employment relationships with social exchange and job embeddedness. *Journal of Applied Psychology, 94,* 277–297.

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., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6–23.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Scientific Software International.

Lee, S. Y., & Zhu, H. T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 53,* 209–232.

Lee, S. Y., & Xu, L. (2003). Case-deletion diagnostics for factor analysis models with continuous and ordinal categorical data. *Sociological Methods & Research, 31,* 389–419.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavior Research, 39,* 653–686.

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

Li, C. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21,* 369–387.

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education, 2,* 17–38.

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

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.

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal, 44,* 1102–1121.

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. (1998–2017). *Mplus user's guide.* Muthén & Muthén.

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.

Morisky, D. E., Tiglao, T. V., Sneed, C. D., Tempongko, S. B., Baltazar, J. C., Detels, R., & Stein, J. A. (1998). The effects of establishment practices, knowledge and attitudes on condom use among Filipina sex workers. *AIDS Care, 10,* 213–220.

Olsson, U. H. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443–460.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7,* 557–595.

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.

Quinn, K. M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis, 12,* 338–353.

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 472–492). The Guildford Press.

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

Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods, 15*, 352–367.

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

Samani, E. B., & Ganjali, M. (2011). Bayesian latent variable model for mixed continuous and ordinal responses with possibility of missing responses. *Journal of Applied Statistics, 38,* 1103–1116.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). *Journal of the Royal Statistical Society. Series B (Methodological), 59,* 667–678.

Song, X. Y., & Lee, S. Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *British Journal of Mathematical and Statistical Psychology, 54,* 237–263.

Song, X. Y., Lee, S. Y., Ng, M. C. Y., So, W. Y., & Chan, J. C. N. (2007). Bayesian analysis of structural equation models with multinomial variable and an application to Type 2 diabetic nephropathy. *Statistics in Medicine, 26,* 2348–2369.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17,* 392–423.

Yuan, K. H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology, 59,* 397–417.

Yuan, K. H., & Schuster, C. (2013). Overview of statistical estimation methods. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods: Volume 1* (pp. 361–387). Oxford University Press.

Yuan, K. H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods and Research, 34,* 249–258.

Zhou, L., Lin, H. Z., Song, X. Y., & Li, Y. (2014). Selection of latent variables for multiple mixed-outcome models. *Scandinavian Journal of Statistics, 41,* 1064–1082.