Behavior Research Methods (2019) 51:409–428
https://doi.org/10.3758/s13428-018-1055-2

BRIEF REPORT

RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods

Yan Xia[1] · Yanyun Yang[2]

Published online: 4 June 2018
© Psychonomic Society, Inc. 2018

Abstract

In structural equation modeling, application of the root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker–Lewis index (TLI) highly relies on the conventional cutoff values developed under normal-theory maximum likelihood (ML) with continuous data. For ordered categorical data, unweighted least squares (ULS) and diagonally weighted least squares (DWLS) based on polychoric correlation matrices have been recommended in previous studies. Although no clear suggestions exist regarding the application of these fit indices when analyzing ordered categorical variables, practitioners are still tempted to adopt the conventional cutoff rules. The purpose of our research was to answer the question: Given a population polychoric correlation matrix and a hypothesized model, if ML results in a specific RMSEA value (e.g., .08), what is the RMSEA value when ULS or DWLS is applied? CFI and TLI were investigated in the same fashion. Both simulated and empirical polychoric correlation matrices with various degrees of model misspecification were employed to address the above question. The results showed that DWLS and ULS lead to smaller RMSEA and larger CFI and TLI values than does ML for all manipulated conditions, regardless of whether or not the indices are scaled. Applying the conventional cutoffs to DWLS and ULS, therefore, has a pronounced tendency not to discover model–data misfit. Discussions regarding the use of RMSEA, CFI, and TLI for ordered categorical data are given.

Keywords Structural equation modeling · Ordered categorical data · Diagonally weighted least squares · Unweighted least squares · Maximum likelihood · Fit index

"Does the hypothesized model fit the data well?" This is a critical question in almost every application of structural equation modeling (SEM). The model chi-square statistic and several fit indices are commonly reported to address this question. Three model fit indices that are widely applied are considered in this article, all of which are based on a fit function given a specific estimation method. They are the root mean square error of approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), comparative fit index (CFI; Bentler, 1990), and Tucker–Lewis index (TLI; Bentler & Bonett, 1980; Tucker & Lewis, 1973). RMSEA is an absolute fit index, in that it assesses how far a hypothesized model is from a perfect model. On the contrary, CFI and TLI are incremental fit indices that compare the fit of a hypothesized model with that of a baseline model (i.e., a model with the worst fit).

The application of RMSEA, CFI, and TLI is heavily contingent on a set of cutoff criteria. Earlier research (e.g., Browne & Cudeck, 1993; Jöreskog & Sörbom, 1993) suggested that an RMSEA value of < .05 indicates a "close fit," and that < .08 suggests a reasonable model–data fit. Bentler and Bonett (1980) recommended that TLI > .90 indicates an acceptable fit. However, these suggestions are largely based on intuition and experience rather than on any statistical justification (see Marsh, Hau, & Wen, 2004). To address the lack of statistical justification of these recommendations, Hu and Bentler (1999) conducted a simulation study to investigate the rejection rates under correct and misspecified models, by applying various cutoff values for many fit indices, including RMSEA, CFI, and TLI. Hu and Bentler suggested that an RMSEA smaller than .06 and a CFI and TLI larger than .95

✉ Yan Xia
yxia@asu.edu

Yanyun Yang
yyang3@fsu.edu

[1] T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ, USA

[2] College of Education, Florida State University, Tallahassee, FL, USA

indicate relatively good model–data fit in general. Hu and Bentler's study has become highly influential, and their recommended cutoffs have been adopted in many SEM practices. However, Hu and Bentler's study only concerns continuous data that are analyzed using the normal-theory maximum likelihood (ML). Hu and Bentler cautioned that the suggested cutoff values might not generalize to conditions that were not manipulated in their study, nor to estimation methods other than ML.

In psychological research the data are frequently ordered categorical (e.g., data collected using Likert scales). Applying the normal-theory ML to the covariance matrix of ordered categorical data can result in biased parameter estimates, inaccurate standard errors, and a misleading chi-square statistic, especially when the number of categories is below five and the categorical distribution is highly asymmetric (e.g., Beauducel & Herzberg, 2006; Johnson & Creech, 1983; Rhemtulla, Brosseau-Liard, & Savalei, 2012). To address the categorical nature of data, the diagonally weighted least squares (DWLS) estimator based on the polychoric correlation matrix has become the most popular method (Savalei & Rhemtulla, 2013). M*plus* (L. K. Muthén & Muthén, 2015) by default implements DWLS when variables are specified as being categorical, with mean- and variance-adjusted chi-square statistics and standard errors. (DWLS with such robust corrections is termed *mean- and variance-adjusted weighted least squares*—i.e., WLSMV; B. O. Muthén, du Toit, & Spisic, 1997.)

RMSEA, CFI, and TLI are based on a fit function that is specific to a chosen estimation method. Because the chi-square statistic is a function of the fit function, RMSEA, CFI, and TLI are also functions of the chi-square statistic. When the scaled chi-square statistic is used in calculating the DWLS fit indices (e.g., M*plus* and lavaan in R; see Eqs. 8–10), we denote the resulting fit indices as *scaled* fit indices—that is, $\mathrm{RMSEA}_\mathrm{S}$, CFI_S, and TLI_S. Although the scaled fit indices are widely applied, no theoretical justification exists for the use of robust chi-square in calculating the fit indices (Brosseau-Liard & Savalei, 2014; Brosseau-Liard, Savalei, & Li, 2012). When the fit indices are calculated as functions of the unscaled chi-square statistic, we denote the unscaled fit indices as $\mathrm{RMSEA}_\mathrm{U}$, CFI_U, and TLI_U. More details regarding the scaled and unscaled fit indices are described in the next section.

When DWLS is applied to ordered categorical data, many studies have questioned whether the conventional cutoff values for RMSEA, CFI, and TLI suggested by Hu and Bentler (1999) can be applied similarly. Yu (2002) conducted a simulation study in which the sample size, type of model misspecification, and type of variable (i.e., binary, normal, and nonnormal data) were varied, and she concluded that no universal cutoff values seem to work across all conditions for DWLS-scaled model fit indices. Beauducel and Herzberg (2006) also found that the performance of the DWLS-scaled model fit indices is not clear and that different cutoff criteria are needed. Garrido, Abad, and Ponsoda (2016) and Yang and Xia (2015) investigated the utility of $\mathrm{RMSEA}_\mathrm{S}$, CFI_S, and TLI_S in determining the number of common factors in exploratory factor analysis, both suggesting that it is difficult to develop a set of cutoff values across all simulation conditions for DWLS. Nye and Drasgow (2011) found that DWLS produced smaller $\mathrm{RMSEA}_\mathrm{S}$ and larger CFI_S and TLI_S values than did the ML-fit indices using an empirical data set with 9,292 observations. Nye and Drasgow's simulation study showed similar results, leading them to conclude that the conventional cutoff values do not work when DWLS is applied. DiStefano and Morgan (2014) found that $\mathrm{RMSEA}_\mathrm{S}$ and CFI_S using DWLS produce problematic results under highly asymmetric categorical distributions, small sample sizes, and dichotomous data. Koziol (2010) found that the differences in $\mathrm{RMSEA}_\mathrm{S}$ and CFI_S between nested models using DWLS do not have sufficient power to detect noninvariance in the measurement invariance testing context. In a similar context, Sass, Schmitt, and Marsh (2014) suggested that the differences in CFI_S and TLI_S between nested models should be applied with caution, particularly for misspecified models, because these variables' performance is impacted by both sample size and model complexity.

One alternative to DWLS is the *unweighted least squares* (ULS) estimator (termed *mean- and variance-adjusted unweighted least squares*—i.e., ULSMV—in M*plus* when robust corrections are implemented; B. O. Muthén et al., 1997). ULS and DWLS with robust corrections were both proposed by B. O. Muthén et al., but the former method has been underutilized, as compared with the latter. We did a search in Google Scholar and located 16 empirical studies before 2016 that applied ULS to Likert-scale data (e.g., Currier & Holland, 2014; De Beer, Pienaar, & Rothmann, 2014; Stander, Mostert, & de Beer, 2014). Simulation studies have found that ULS results in higher nonconvergence rates than does DWLS (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009), but that it provides slightly more accurate parameter estimates and standard errors than DWLS (Forero et al., 2009; C. H. Li, 2014; Yang-Wallentin, Jöreskog, & Luo, 2010). Savalei and Rhemtulla (2013) further investigated the robust chi-square statistics and concluded that ULS with the mean- and variance-adjusted chi-square statistic outperforms DWLS regarding Type I error rates and power. The performance of $\mathrm{RMSEA}_\mathrm{S}$, CFI_S, and TLI_S under ULS has not been extensively evaluated. We found only one such study, performed by Maydeu-Olivares, Fairchild, and Hall (2017), who investigated how the number of categories could impact the ULS-$\mathrm{RMSEA}_\mathrm{S}$. They found that ULS-$\mathrm{RMSEA}_\mathrm{S}$ decreases, and thus results in the loss of power, with more data categories.

On the basis of our literature review, we concluded that the performance of DWLS-$\mathrm{RMSEA}_\mathrm{S}$, CFI_S, and TLI_S is elusive, and that no clear guideline exists for goodness-of-fit evaluation for ordered categorical data in SEM. However, researchers in substantive areas are still tempted to apply the conventional cutoff values to DWLS. It is very common to see statements in published articles like "DWLS [or WLSMV] was applied to analyze ordered categorical data. A good model–data fit is indicated by RMSEA < .06, CFI > .95, and TLI > .95 (Hu & Bentler, 1999)." All 16 empirical studies that we found that employed ULS also applied the conventional cutoffs to evaluate the model–data fit. Hu and Bentler and other researchers (e.g., Marsh et al., 2004) have cautioned against the universal use of conventional cutoff criteria, which, unfortunately, has not seemed to stop the widespread use of these cutoffs with other estimation methods such as DWLS and ULS. In this article we take an alternative perspective, to examine the model–data fit evaluation under different estimation methods when ordered categorical data are analyzed. We aimed to show the consequences for model–data fit evaluation that result from applying the conventional cutoffs to DWLS and ULS. Specifically, this article answers the following research question: When a hypothesized model is fitted to a population-level polychoric correlation matrix using ML and results in a specific RMSEA value (e.g., .06), what is the RMSEA value if the same model is fitted to the same matrix using DWLS or ULS? CFI and TLI were investigated in the same manner.

We conducted a simulation study to answer the above research question. Although in SEM applications fit indices are commonly used in sample-level SEM analysis, we decided to conduct the present simulation study at the population level, so that no confounding due to sample fluctuation would be introduced. We made this decision for the following two reasons. First, if the same misspecified model were fit to the same population data but resulted in different fit indices using different estimation methods, then it would be clear that the same value of a fit index tells a different story regarding model–data misfit using different estimation methods. A clear understanding of these fit indices at the population level will be necessary before the sampling properties can be investigated. All the simulation studies we reviewed and cited above were conducted under finite samples; therefore, it is not clear whether the different performance of the fit indices under different estimation methods was due to their population properties, sampling error, or both. Second, although the scaled indices are commonly reported for sample-level analysis in SEM applications, they are calculated without theoretical justification, as has been evidenced by Brosseau-Liard and Savalei (2014) and Brosseau-Liard et al. (2012). As we will show in formulas (see Eqs. 8–13), the unscaled fit indices from ULS and DWLS converge to the theoretical definitions of RMSEA, CFI, and TLI as the sample size increases to infinity, but the scaled fit indices converge to distorted asymptotic values. Therefore, we question the appropriateness of using the scaled fit indices currently implemented in software programs.

We employed Cudeck and Browne's (1992) simulation technique in order to generate population polychoric correlation matrices with predefined ML-RMSEA values as measures of degrees of model misspecification. The generated matrices were then fitted using DWLS and ULS in order to calculate the corresponding scaled and unscaled DWLS and ULS fit indices. We chose the ML fit indices based on the polychoric correlation matrices as the benchmark for comparison. At the population level, the polychoric correlation matrix is essentially the Pearson correlation matrix for generating continuous data, given that the underlying continuous variables follow a standard multivariate normal distribution. Therefore, the population values of the ML fit indices based on polychoric correlation matrices are the same as those based on the Pearson correlation matrices of continuous data. Because Hu and Bentler's (1999) cutoff values were developed using the covariance matrices for continuous variables, the population values of ML fit indices using polychoric correlation matrices are consistent with those in Hu and Bentler's study. It is also important to note that robust corrections are recommended to adjust the chi-square statistic and standard error when ML is employed for polychoric correlation matrices (e.g., Yang-Wallentin et al., 2010).

The next section briefly summaries the fit functions of ML, ULS, and DWLS, as well as the population definitions of RMSEA, CFI, and TLI. Because current software programs (e.g., M*plus* and lavaan in R) scale the fit indices such that the indices are functions of the scaled chi-square statistics, we will also compare the scaled fit indices with the unscaled versions. Thereafter, a comparison of RMSEA, CFI, and TLI across the ML, ULS, and DWLS methods is presented using the simulation technique outlined in Cudeck and Browne (1992). Following the simulation study, we take six empirical polychoric correlation matrices reported in published studies, analyze the matrices using ML, ULS, and DWLS, and compare the resulting fit indices across the three estimators. We will show that the conclusions of a model–data fit evaluation can be dramatically different, depending on which estimation method is applied. Finally, suggestions and discussions regarding the application of RMSEA, CFI, and TLI in SEM with ordered categorical data are given.

Fit function, RMSEA, CFI, and TLI

In SEM analysis, it is frequently assumed that a continuous variable underlies an ordered categorical variable (e.g., Olsson, 1979; Pearson, 1904; Tallis, 1962). This underlying continuous variable is categorized into the ordered categorical variable on the basis of a set of threshold values. Under this assumption, the measure of association that is of interest in the modeling is the correlation between the underlying continuous variables, which is termed the *polychoric correlation* (*tetrachoric correlation* is a special case in which both ordinal variables have two categories).

To analyze ordered categorical variables in SEM, a three-stage procedure can be applied. The first two steps estimate the thresholds and polychoric correlation matrix, which can be achieved by Olsson's (1979) two-step ML method, implemented by default in software programs (e.g., M*plus*). In the third stage, the estimated threshold values and polychoric correlation matrix are fitted to a hypothesized model using an estimation method (e.g., ML, DWLS, or ULS). The model parameter estimates are then obtained by minimizing a certain fit function, as described below.

The general form of a fit function can be written as

$$F = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\omega}))'\mathbf{W}^{-1}(\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\omega})). \tag{1}$$

In Eq. 1, $\mathbf{s}' = \left(\mathbf{s}_1', \mathbf{s}_2'\right)$, where $\mathbf{s}_1$ is a column vector containing the thresholds and $\mathbf{s}_2$ contains the nonduplicate unstructured polychoric correlations; $\boldsymbol{\sigma}'(\boldsymbol{\omega}) = \left(\boldsymbol{\sigma}_1'(\boldsymbol{\omega}), \boldsymbol{\sigma}_2'(\boldsymbol{\omega})\right)$, where $\boldsymbol{\omega}$ is a vector containing the model parameters and $\boldsymbol{\sigma}_1(\boldsymbol{\omega})$ and $\boldsymbol{\sigma}_2(\boldsymbol{\omega})$ are the vectors containing the model-implied thresholds and polychoric correlations, respectively; and $\mathbf{W}$ is a weight matrix that is specific to an estimation method. The fit function characterizes the discrepancy between $\mathbf{s}$ and $\boldsymbol{\sigma}(\boldsymbol{\omega})$. The vector $\boldsymbol{\omega}$ is estimated by minimizing F. This study only considered the models with unstructured thresholds such that $\mathbf{s}_1' - \boldsymbol{\sigma}_1'(\boldsymbol{\omega}) = \mathbf{0}$.

In Eq. 1, different forms of $\mathbf{W}$ lead to different estimators. The weight matrices for ULS and DWLS are

$$\mathbf{W}_{\mathrm{ULS}} = \mathbf{I}, \tag{2}$$

and

$$\mathbf{W}_{\mathrm{DWLS}} = N \cdot \mathrm{diag}(\boldsymbol{\Gamma}), \tag{3}$$

respectively. In Eq. 2, $\mathbf{I}$ is an identity matrix. In Eq. 3, N is the sample size, $\boldsymbol{\Gamma}$ is the asymptotic variance and covariance matrix of thresholds and polychoric correlations, and diag($\boldsymbol{\Gamma}$) is a diagonal matrix with all the diagonal elements the same as the diagonal elements in $\boldsymbol{\Gamma}$. The detailed calculation of $N \cdot$ diag($\boldsymbol{\Gamma}$), which is a function of the thresholds and polychoric correlations, is described in Olsson (1979). Although $\mathbf{W}_{\mathrm{DWLS}}$ is expressed as $N \cdot$ diag($\boldsymbol{\Gamma}$), $\mathbf{W}_{\mathrm{DWLS}}$ is not a function of N, because the formula of diag($\boldsymbol{\Gamma}$) has N in the denominator, and thus N in Eq. 3 is canceled out.

It is also possible to fit the polychoric correlation matrix using the ML fit function, but there is no theoretical justification for the resulting robust standard error and chi-square statistic (Yang-Wallentin et al., 2010). The fit function of ML can also be expressed as

$$F_{\mathrm{ML}} = \log|\Sigma(\omega)| + \mathrm{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\omega)\right) - \log|\mathbf{S}| - p, \tag{4}$$

where $\mathbf{S}$ is the unstructured polychoric correlation matrix, $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ is the model-implied polychoric correlation matrix, and p is the number of observed variables.

RMSEA, CFI, and TLI are defined on the basis of the fit function (Bentler, 1990; Bentler & Bonett, 1980; Steiger, 1990; Steiger & Lind, 1980; Tucker & Lewis, 1973). Let H and B denote the hypothesized model and the baseline model (i.e., a model assuming zero correlation between every pair of underlying continuous variables), respectively. $\widetilde{F_H}$ and $\widetilde{F_{\mathrm{B}}}$ represent the minimized fit functions of H and B at the population level, respectively, and df_{H} and df_{B} are the corresponding model degrees of freedom. In the population,

$$\mathrm{RMSEA_U} = \sqrt{\frac{\tilde{F}_{\mathrm{H}}}{df_{\mathrm{H}}}}, \tag{5}$$

$$\mathrm{CFI_U} = 1 - \frac{\tilde{F}_{\mathrm{H}}}{\tilde{F}_{\mathrm{B}}}, \tag{6}$$

and

$$\mathrm{TLI_U} = 1 - \frac{\tilde{F}_{\mathrm{H}}/df_{\mathrm{H}}}{\tilde{F}_{\mathrm{B}}/df_{\mathrm{B}}}. \tag{7}$$

The subscript U in Eqs. 5–7 means that the indices are unscaled. Because the mean- and variance-adjusted chi-square is applied in both WLSMV and ULSMV, current software programs (e.g., M*plus* and lavaan in R) compute the scaled fit indices as functions of the adjusted chi-square statistic. At the sample level, the scaled indices are calculated as

$$\mathrm{RMSEA}_{\mathrm{S},N} = \sqrt{\frac{\hat{a}_{\mathrm{H}}(N-1)\hat{F}_{\mathrm{H}} + \hat{b}_{\mathrm{H}}}{(N-1)df_{\mathrm{H}}} - \frac{1}{N-1}}, \tag{8}$$

$$\mathrm{CFI}_{\mathrm{S},N} = 1 - \frac{\hat{a}_{\mathrm{H}}(N-1)\hat{F}_{\mathrm{H}} + \hat{b}_{\mathrm{H}} - df_{\mathrm{H}}}{\hat{a}_{\mathrm{B}}(N-1)\hat{F}_{\mathrm{B}} + \hat{b}_{\mathrm{B}} - df_{\mathrm{B}}}, \tag{9}$$

and

$$\mathrm{TLI}_{\mathrm{S},N} = 1 - \frac{\hat{a}_{\mathrm{H}}(N-1)\hat{F}_{\mathrm{H}} + \hat{b}_{\mathrm{H}} - df_{\mathrm{H}}}{\hat{a}_{\mathrm{B}}(N-1)\hat{F}_{\mathrm{B}} + \hat{b}_{\mathrm{B}} - df_{\mathrm{B}}} \cdot \frac{df_{\mathrm{B}}}{df_{\mathrm{H}}}. \tag{10}$$

$\hat{a}$ and $\hat{b}$ converge to α and b (i.e., the scaling parameter and shifting parameter, respectively; Asparouhov & Muthén, 2010), and $\hat{F}$ converges to $\tilde{F}$ as N increases to infinity. Therefore, Eqs. 8–10 converge to

$$\mathrm{RMSEA_S} = \sqrt{\frac{a_{\mathrm{H}}\tilde{F}_{\mathrm{H}}}{df_{\mathrm{H}}}}, \tag{11}$$

$$\mathrm{CFI_S} = 1 - \frac{a_{\mathrm{H}}\tilde{F}_{\mathrm{H}}}{a_{\mathrm{B}}\tilde{F}_{\mathrm{B}}}, \tag{12}$$

and

$$\mathrm{TLI_S} = 1 - \frac{a_{\mathrm{H}}\tilde{F}_{\mathrm{H}}}{a_{\mathrm{B}}\tilde{F}_{\mathrm{B}}} \cdot \frac{df_{\mathrm{B}}}{df_{\mathrm{H}}}. \tag{13}$$

Equations 5 to 7 and 11 to 13 show that both scaled and unscaled RMSEA, CFI, and TLI are functions of the fit function, whose value is dependent on the chosen estimation method. Therefore, given a population polychoric correlation matrix and a hypothesized misspecified model, different estimators can lead to different values of the fit indices. These equations also show that even for the same estimator, the scaled fit indices converge to different population values from unscaled fit indices. In Study 1, we compared each fit index across the three estimators using computer-generated polychoric correlation matrices. In Study 2, we fit six polychoric correlation matrices reported in empirical research using the three estimators. For both studies, we highlighted the potential consequence on the model–data fit evaluation if the conventional cutoffs developed under ML are applied to DWLS and ULS.

Study 1: Comparing fit indices across estimators using population matrices

We employed the simulation technique outlined in Cudeck and Browne (1992) to generate polychoric correlation matrices with target degrees of model misspecification (as measured by RMSEA_{U}) under ML. We then specified the thresholds for each variable. For each generated matrix and threshold condition, we analyzed the same model using DWLS and ULS to obtain the corresponding model fit indices.

Generation of polychoric correlation matrices

Cudeck and Browne's (1992) method was applied in order to generate the polychoric correlation matrices with prespecified ML-RMSEA_{U} values. The resulting matrices had two attributes: When these matrices were analyzed by ML using a prespecified target model, (1) the resulting RMSEA_{U} value was the same as the prespecified RMSEA_{U}, and (2) the parameter estimates were the same as those in the prespecified target model. Below we describe the use of this method for generating polychoric correlation matrices in three steps.

The first step was to specify a target model-implied polychoric correlation matrix $\boldsymbol{\Sigma}(\boldsymbol{\omega})$. Four target CFA models with different numbers of latent factors (one and three) and observed indicators (9 and 18) were used to specify $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ on the basis of

$$\Sigma(\omega) = \wedge\Phi\wedge' + \Psi, \tag{14}$$

where $\wedge$ was the loading matrix, with

$$\begin{aligned}
&\text{Model 1}: \wedge' = [.7\ .7\ .7\ .7\ .7\ .7\ .7\ .7\ .7]\\
&\text{Model 2}: \wedge' = [.7\ .7\ .7\ .7\ .7\ .7\ .7\ .7\ .7.7\ .7\ .7\ .7\ .7\ .7\ .7\ .7\ .7]\\
&\text{Model 3}: \wedge' = \begin{bmatrix} .7\ .7\ .7\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ .7.7.7\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ .7\ .7\ .7 \end{bmatrix}\\
&\text{Model 4}: \wedge' = \begin{bmatrix} .7\ .7\ .7\ .7\ .7\ .7\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 7\ .7\ .7\ .7\ .7\ .7\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 7\ .7\ .7\ .7\ .7\ .7 \end{bmatrix}
\end{aligned} \tag{15}$$

$\boldsymbol{\Phi}$ was the factor covariance matrix, with

$$\begin{aligned}
&\text{Model 1 and}\quad \text{Model 2}: \boldsymbol{\Phi} = [1]\\
&\text{Model 3 and}\quad \text{Model 4}: \boldsymbol{\Phi} = \begin{bmatrix} 1 & .3 & .3 \\ .3 & 1 & .3 \\ .3 & .3 & 1 \end{bmatrix},
\end{aligned} \tag{16}$$

and $\boldsymbol{\Psi}$ was the residual matrix, with no covariance between any pair of residuals:

$$\begin{aligned}
&\text{Model 1 and Model 3}: \text{diag}(\boldsymbol{\Psi}) = [.51, .51, .51, .51, .51, .51, .51, .51, .51]\\
&\text{Model 3 and Model 4}: \text{diag}(\boldsymbol{\Psi}) = \begin{bmatrix} .51, .51, .51, .51, .51, .51, .51, .51, .51, \\ .51, .51, .51, .51, .51, .51, .51, .51, .51 \end{bmatrix}.
\end{aligned} \tag{17}$$

The second step was to specify a target ML-RMSEA_{U} as the measure of the degree of misspecification. We varied the target RMSEA_{U} from .02 to .12 with an interval of .02, so that the degree of misspecification for the target CFA models ranged from small to large according to the conventional cutoff suggested by Hu and Bentler (1999). We did not include the condition with the target RMSEA_{U} of 0 because the corresponding analysis model was correctly specified and the fit indices showed perfect fit (i.e., scaled and unscaled RMSEA = 0, CFI = 1, and TLI = 1) at the population level, regardless of the estimator used.

The third step was to construct an error matrix **E** to obtain the population polychoric correlation matrix **S:**

$$\mathbf{S} = \boldsymbol{\Sigma}(\boldsymbol{\omega}) + \mathbf{E}.$$

The details of the matrix operations to generate **E** were described in Cudeck and Browne (1992). Briefly described, **E** was generated such that two constraints were met: When **S** was fitted to the target CFA model using ML, the estimated parameters yielded the same values as those in Eqs. 15–17, and the resulting ML-RMSEA_{U} was equal to the prespecified ML-RMSEA_{U} value. Because the computation of **E** involved random-number generation, it was possible to generate many **S** matrices, each of which resulted in the prespecified ML-RMSEA_{U} value and parameter estimates when it was fitted to the target CFA model using ML. We generated 200 **S** matrices for each target ML-RMSEA_{U} value under each target CFA model.

SAS (SAS Institute, 2015) was used to implement these three steps for generating the polychoric correlation matrices. In sum, 200 (matrices under each condition) × 6 (target ML-RMSEA_U values) × 4 (target CFA models) = 4,800 matrices were generated.

Analyses

Each matrix was fitted to its corresponding target CFA model using ULS and DWLS, with the CALIS procedure in SAS/STAT 14.1 (SAS Institute, 2015). For the ULS scaled indices and DWLS, threshold values need to be specified in order to calculate the weight matrix ($\mathbf{W}_{\text{DWLS}}$), shown in Eq. 3, as well as the *a* and *b* parameters in Eqs. 11–13, which were determined using the population thresholds and polychoric correlations (Asparouhov & Muthén, 2010; B. O. Muthén et al., 1997; Olsson, 1979).[1] We varied the number of categories to be either two or four. All ordered categorical variables under each condition had the same set of thresholds. With two categories, the thresholds were either [0] or [1.5] across the variables. With four categories, the thresholds were varied at [– 1, 0, 1] and [0, 1, 2]. The values of [0] and [– 1, 0, 1] resulted in symmetric categorical distributions, whereas [1.5] and [0, 1, 2] resulted in high levels of asymmetry.

The thresholds were not needed to compute the ULS-RMSEA_U, CFI_U, and TLI_U at the population level, for two reasons. First, the threshold structure was saturated—that is, $\mathbf{s}_1 - \boldsymbol{\sigma}_1(\boldsymbol{\omega}) = \mathbf{0}$ in Eq. 1. Second, the weight matrix shown in Eq. 2 is an identity matrix and is not impacted by the threshold values.

The resulting unscaled and scaled variants of RMSEA, CFI, and TLI based on DWLS and ULS were calculated using Eqs. 5–7 and 11–13. The matrices were not analyzed using ML, because Cudeck and Browne's (1992) method ensured that the ML-RMSEA_U was the same as the prespecified ML-RMSEA_U. In addition, the ML-CFI_U and TLI_U could be calculated using matrix operations. The corresponding ML-CFI_U and TLI_U for each target model are presented in Table 1. Note that, given an ML-RMSEA_U, the resulting ML-CFI_U and TLI_U were expected to differ across different population models (e.g., Lai & Green, 2016).

Results

Comparing ML with ULS Figure 1 plots the ULS-RMSEA_U for each generated polychoric correlation matrix against the prespecified ML-RMSEA_U. The horizontal and vertical dotted lines represent the conventional cutoff value of RMSEA (i.e., .06). For all the matrices, the ULS-RMSEA_U was lower than the ML-RMSEA_U, suggesting that the models had a greater tendency to be considered acceptable using ULS. Figure 2 compares ULS-RMSEA_S with ML-RMSEA_U. The left panel presents the results when each variable had the threshold of [0], and the right panel has the threshold equal to [1.5]. Similar to ULS-RMSEA_U, ULS-RMSEA_S was consistently smaller than ML-RMSEA_U. With more extreme thresholds, ULS-RMSEA_S became even lower, and thus suggesting a better fit.

Table 1 ML-RMSEA_U, -CFI_U, and -TLI_U under each simulation condition

	ML-RMSEA_U					
	.02	.04	.06	.08	.10	.12
Model 1: One Factor With 9 Items						
ML-CFI_U	.997	.989	.975	.956	.934	.907
ML-TLI_U	.996	.985	.967	.942	.911	.876
Model 2: One Factor With 18 Items						
ML-CFI_U	.994	.977	.950	.914	.872	.826
ML-TLI_U	.993	.974	.943	.903	.855	.803
Model 3: Three Factors With 9 Items						
ML-CFI_U	.996	.982	.961	.933	.890	.860
ML-TLI_U	.993	.973	.941	.900	.848	.790
Model 4: Three Factors With 18 Items						
ML-CFI_U	.992	.969	.932	.886	.833	.775
ML-TLI_U	.991	.964	.922	.868	.806	.740

Using RMSEA < .06 as the cutoff, many matrices (those marked as crosses in Figs. 1 and 2) resulted in poor fit when analyzed by ML but acceptable fit according to either ULS-RMSEA_U or ULS-RMSEA_S. For example, when the ML-RMSEA_U was .10, all the matrices in the one-factor models and in the three-factor models with 18 items showed ULS-RMSEA_U and RMSEA_S < .06. In addition, for a given ML-RMSEA_U, the one-factor models tended to result in a lower ULS-RMSEA_U and -RMSEA_S than did the three-factor models. This pattern suggests that ULS-RMSEA_U appeared to be less likely to detect model–data misfit in more parsimonious models than in more complex models if the same cutoff was applied.

Figures 3 and 4 plot the ULS-CFI_U and -CFI_S, respectively, against the ML-CFI_U. Both the ULS-CFI_U and -CFI_S were higher than the ML-CFI_U. The majority of the ULS-CFI_U and -CFI_S values were above .95 (except for several analyses with three factors and nine items), whereas the corresponding ML-CFI_U values fell below .95. For example, under the three-factor model with 18 items, when ML-CFI_U was approximately .77, the corresponding ULS-CFI_U and -CFI_S values were generally above .95. This pattern suggests that ULS-CFI_U and -CFI_S were much less sensitive to model misspecification than

[1] $\mathbf{W}_{\text{DWLS}}$ was calculated according to Olsson's (1979) Eq. 23; the *a* and *b* parameters for each simulation condition were approximated by generating a large dataset (1,000,000 observations) and then analyzing this dataset using the lavaan package in R.

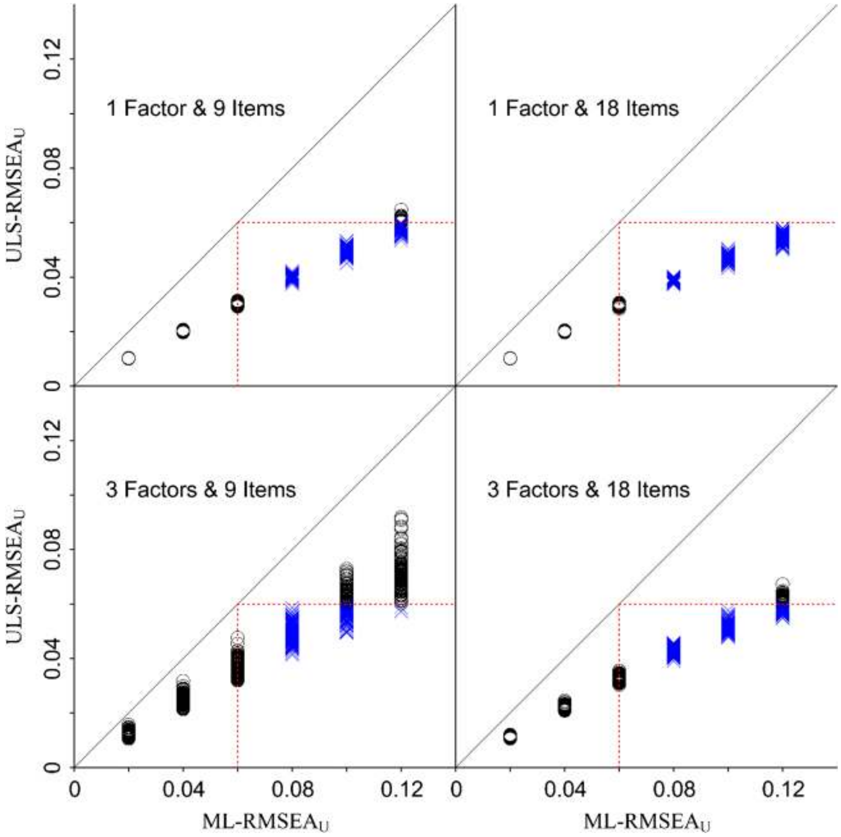

Fig. 1 Comparison between ML-RMSEA$_U$ and ULS-RMSEA$_U$. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if RMSEA < .06 is used as the criterion of acceptable fit.

was ML-CFI if the same cutoff was applied. The results for TLI had very similar patterns, and thus are not reported.

Comparing ML with DWLS Figures 5 and 6 show the values of DWLS-RMSEA$_U$ and -RMSEA$_S$, respectively, for data

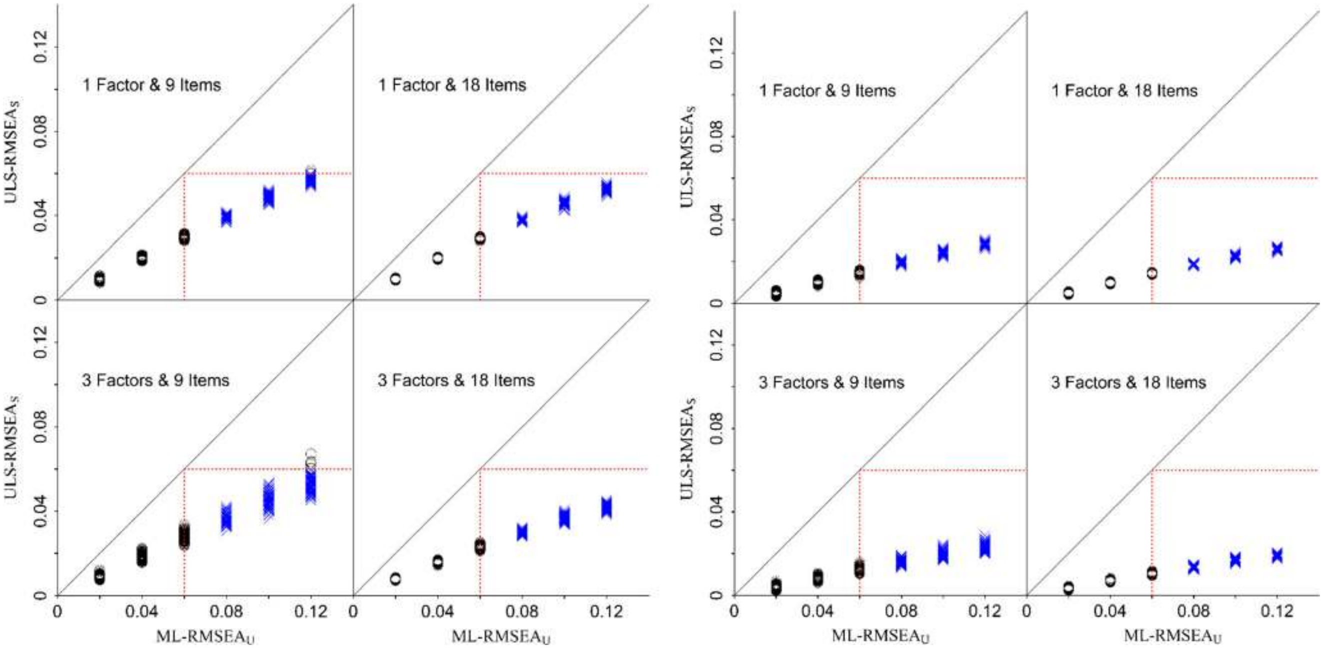

Fig. 2 Comparison between ML-RMSEA$_U$ and ULS-RMSEA$_S$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if RMSEA < .06 is used as the criterion of acceptable fit.

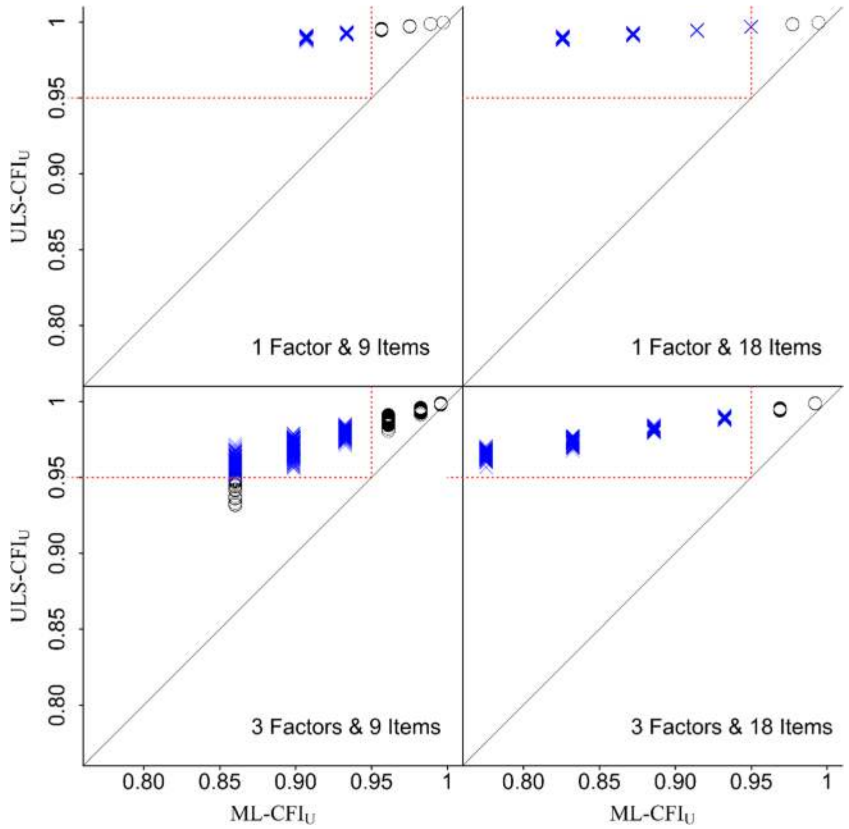

Fig. 3 Comparison between ML-CFI$_U$ and ULS-CFI$_U$. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if CFI > .95 is used as the criterion of acceptable fit.

with two categories. The patterns of DWLS-RMSEA$_U$ and -RMSEA$_S$ were similar to those from ULS, in that all the values were lower than those for ML-RMSEA$_U$. DWLS-RMSEA$_U$ and -RMSEA$_S$ were also less likely to discover

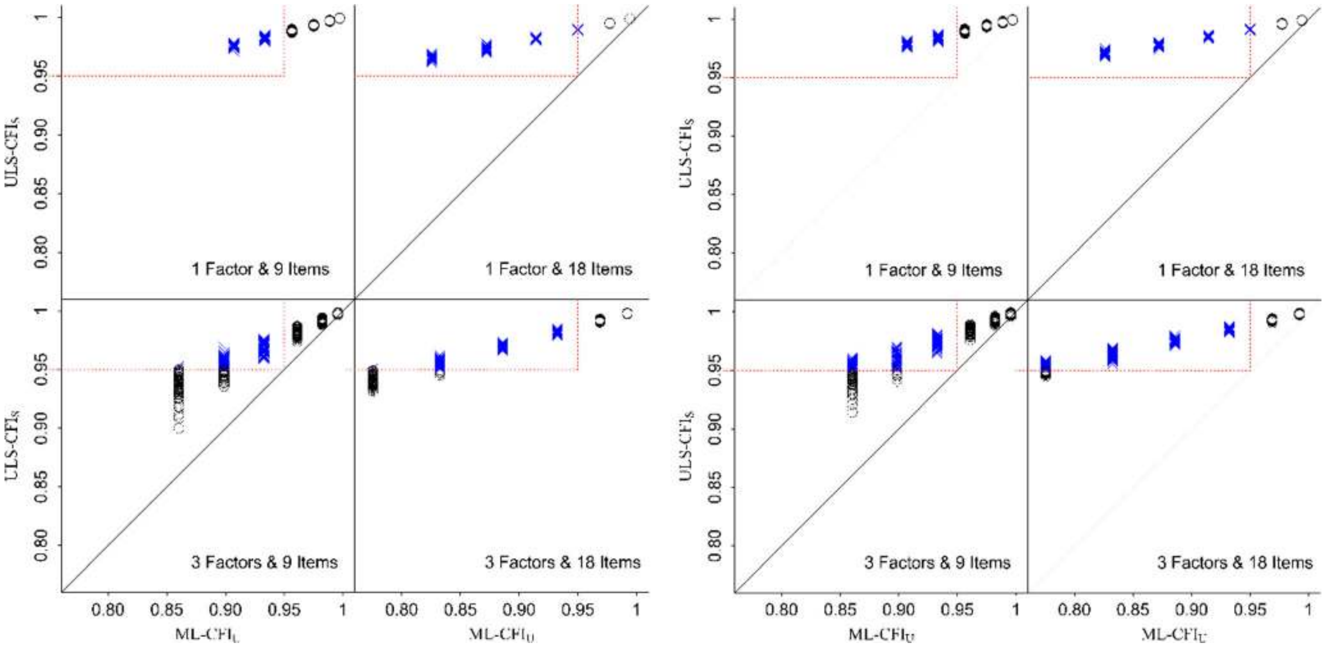

Fig. 4 Comparison between ML-CFI$_U$ and ULS-CFI$_S$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if CFI > .95 is used as the criterion of acceptable fit.

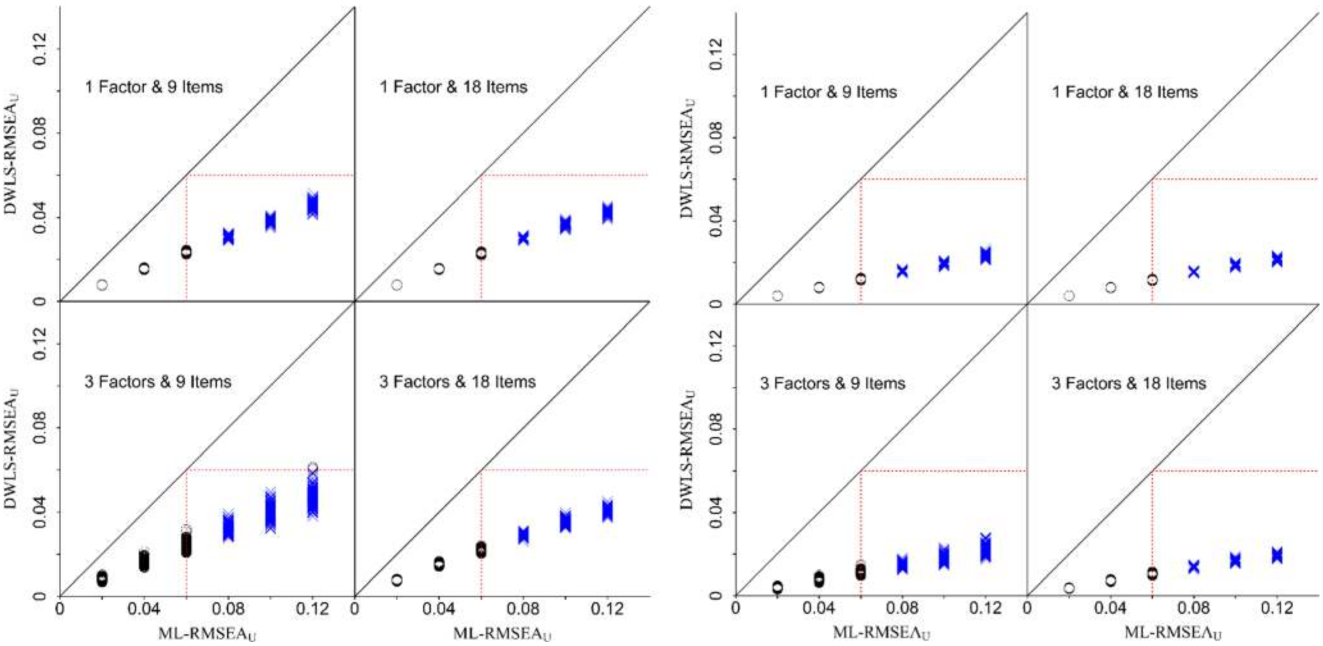

Fig. 5 Comparison between ML-RMSEA$_U$ and DWLS-RMSEA$_U$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if RMSEA < .06 is used as the criterion of acceptable fit.

model–data misfit with 18 than with nine items and with the one-factor than with the three-factor models. DWLS-RMSEA$_U$ and -RMSEA$_S$ were both dependent upon the categorical distribution, because its values decreased (all below .03) as the categorical distribution became asymmetric. Appendix 1 shows DWLS-RMSEA$_U$ and -RMSEA$_S$, respectively, for data with four categories. Similarly, the resulting values were smaller than those for ML-RMSEA$_U$, especially when the target models had 18 items and the categorical distribution was asymmetric. The impact of the number of latent factors on the DWLS indices was less clear. DWLS-RMSEA$_U$ was lower in the one-

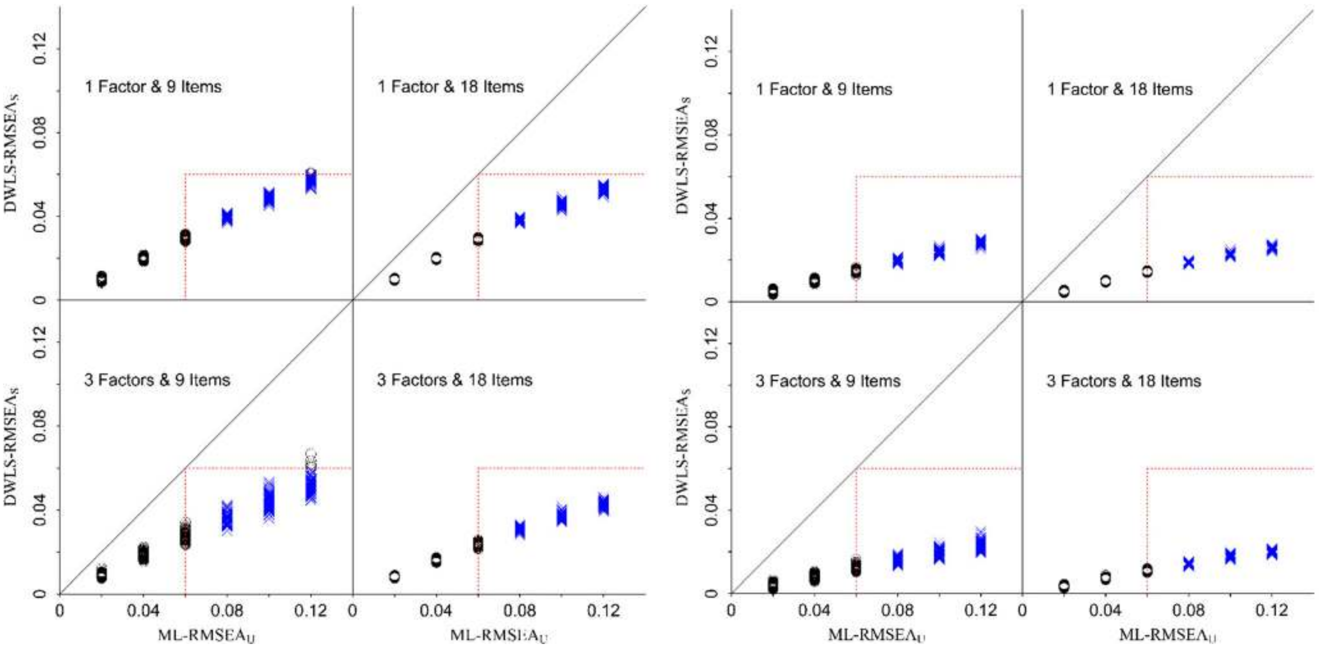

Fig. 6 Comparison between ML-RMSEA$_U$ and DWLS-RMSEA$_S$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if RMSEA < .06 is used as the criterion of acceptable fit.

factor than in the three-factor models, except for the analyses with 18 items and ML-RMSEA$_U$ = .12.

Figures 7 and 8 present the DWLS-CFI$_U$ and -CFI$_S$, respectively, when the data had two categories. ML-CFI$_U$ ranged from .77 to .99, but DWLS-CFI$_U$ and -CFI$_S$ were mostly greater than .95. DWLS-CFI$_U$ and -CFI$_S$ appeared to be relatively stable across conditions with different thresholds and numbers of categories. Again, if CFI > .95 was adopted as the indication of an acceptable model, the model was more likely to be judged a good model when DWLS rather than ML was used, especially for the target models with one factor or with nine items. The results for TLI were extremely similar to those for CFI, and thus are not reported.

Implications of Study 1

On the basis of the conditions we manipulated in Study 1, ULS- and DWLS-RMSEA$_U$ and -RMSEA$_S$ were smaller than those from ML-RMSEA$_U$, and ULS- and DWLS-CFI$_U$ values were larger than those from ML-CFI$_U$. Therefore, both scaled and unscaled RMSEA, CFI, and TLI from ULS and DWLS are more likely than ML to indicate better model–data fit when the same misspecified model is analyzed. If we continue applying Hu and Bentler's (1999) cutoffs developed under ML to ULS and DWLS in SEM analyses with ordered categorical variables, in the long run we would expect that more models that should have been considered poor fit would accumulate in the published literature and be considered acceptable fit. In addition, both unscaled and scaled ULS- and DWLS-CFI and -TLI were mostly clustered at large values (i.e., > .95), making them less useful in differentiating "unacceptable" from "acceptable" models. Because we investigated population-level fit indices, the conclusions above apply to data with large-enough sample sizes.

Study 2: Comparing fit indices across estimators using empirical polychoric correlation matrices

In Study 1 we employed polychoric correlation matrices using simulated matrices. To further demonstrate the consequences of applying the conventional cutoffs to analyses with ULS and DWLS, in Study 2 we applied six empirical polychoric correlation matrices (labeled as M1–M6) that have been reported in published research articles (Fernandez & Moldogaziev, 2013; Iglesias, Burnand, & Peytremann-Bridevaux, 2014; MacInnis, Lanting, Rupert, & Koehle, 2013; Martínez-Rodríguez et al., 2016; Nguyen et al., 2016; Pettersen, Nunes, & Cortoni, 2016) in the behavioral sciences (e.g., sexual behavior, employee performance, and aggression). These articles were located by a nonexhaustive search using Google Scholar from 2013 to 2016 and the search term "polychoric correlation matrix." All six of the articles reported polychoric correlation matrices and applied SEM with ordered categorical variables. We treated these matrices as the population matrices and fitted them to the models that were considered to be acceptable in their corresponding articles, using ML, ULS, and DWLS. The six polychoric correlation matrices are available in Appendix 2.

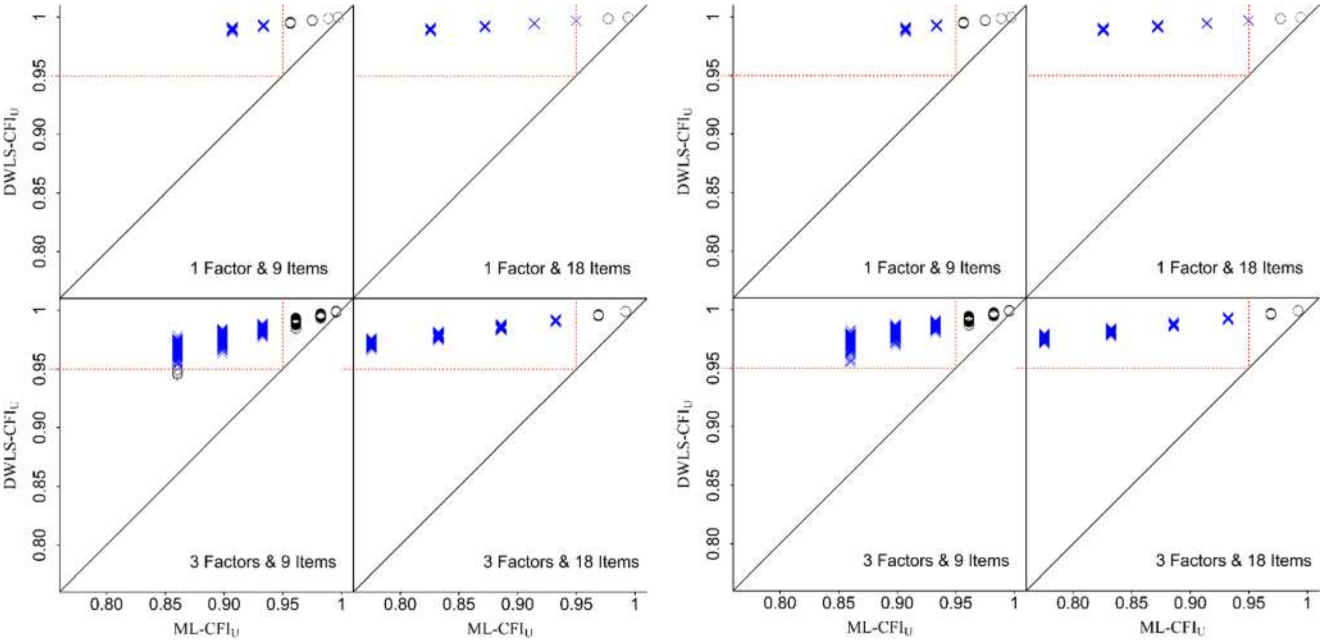

Fig. 7 Comparison between ML-CFI$_U$ and DWLS-CFI$_U$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if CFI > .95 is used as the criterion of acceptable fit.

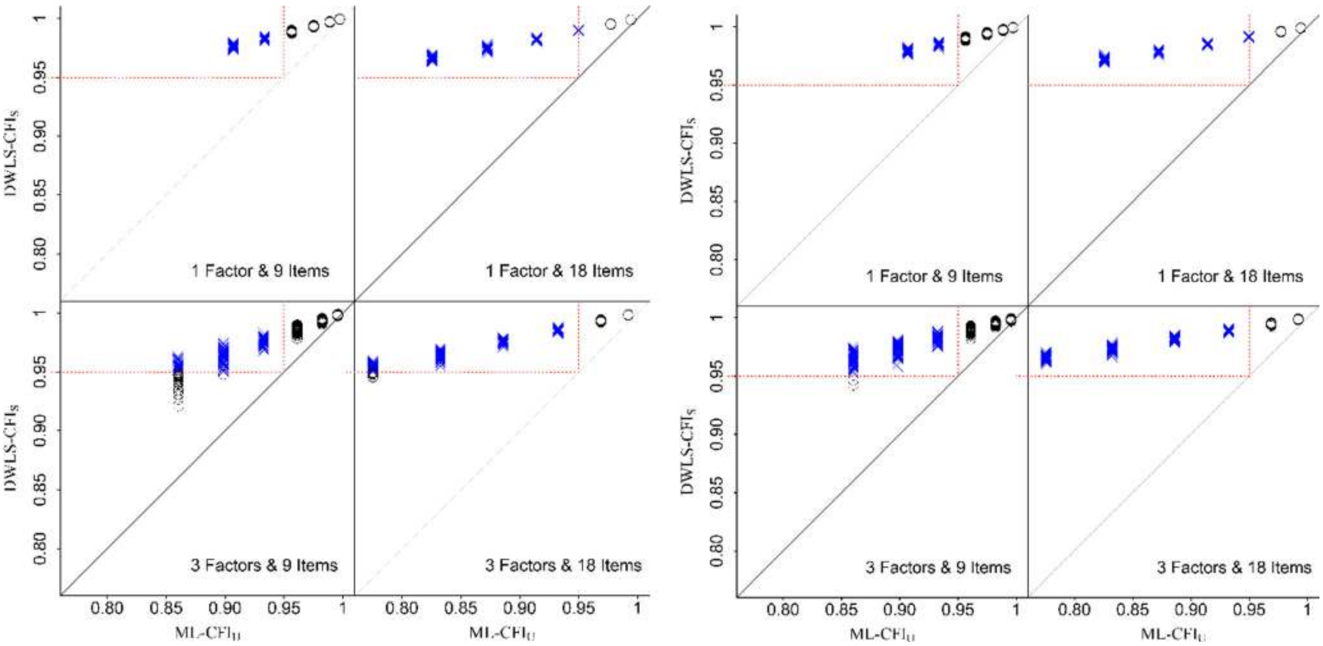

Fig. 8 Comparison between ML-CFI$_U$ and DWLS-CFI$_S$ when the data are binary. The left panel has threshold = [0], and the right panel has threshold = [1.5]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if CFI > .95 is used as the criterion of acceptable fit.

Our goal was to replicate the conclusion from Study 1 using real polychoric correlation matrices reported in empirical research. That is, we aimed to show that both unscaled and scaled ULS- and DWLS-RMSEA, -CFI, and -TLI tend to indicate a better model–data fit than does ML when analyzing ordered categorical variables.

Table 2 summarizes the target models that were used to analyze M1–M6, which include five CFA models and one full structural model, with different numbers of factors (one to four) and different test lengths (5–12 items). For the ML and ULS unscaled indices, the threshold values were not needed to calculate the fit indices at the population level. However, for the ULS scaled indices and for DWLS the threshold values were required, and we manipulated them in the same way as in Study 1.[2] The resulting RMSEA, CFI, and TLI values are presented in Table 2. The results were consistent with those from Study 1, in that the DWLS- and ULS-RMSEA$_U$ and -RMSEA$_S$ values were smaller than those from ML-RMSEA$_U$, and the DWLS- and ULS-CFI$_U$ and -CFI$_S$ values were larger than those from ML-CFI$_U$. The patterns for TLI were similar to those for CFI. For example, for M1, the ML-RMSEA$_U$, -CFI$_U$, and -TLI$_U$ values were .245, .919, and .879, all suggesting severe misfit. However, ULS and DWLS both produced indices that suggested much better fit, especially for CFI and TLI (> .998, reaching the ceiling and suggesting an excellent fit), regardless of whether or not they were scaled.

We also selected M1, M2, and M4 to investigate whether the conclusions above could be observed in the sample-level data. The sample size was fixed at 500, which is considered a moderate sample size in behavioral research. The results and more details of the simulation are presented in Appendix 3, consistently showing that ULS and DWLS resulted in overoptimistic unscaled and scaled fit indices as compared with ML.

Discussion

Given the lack of investigation of the ULS and DWLS fit indices, in this article we compared them with their ML counterparts at the population level. Study 1 used Cudeck and Browne's (1992) simulation technique and showed that both the unscaled and scaled RMSEA values from ULS and DWLS were smaller than the ML-RMSEA$_U$ values. CFI and TLI, in contrast, from ULS and DWLS had values larger than those from ML. In Study 2 we employed six polychoric correlation matrices reported in published research and found consistent results. In summary, the ULS- and DWLS-RMSEA, -CFI, and -TLI values, scaled or not, are more likely to indicate a better model–data fit than are the ML fit indices when the same misspecified model is analyzed and when the same sets of conventional cutoff values are adopted. Therefore, applying the conventional cutoffs to ULS and DWLS can lead in the

[2] None of the six articles reported the threshold values or the cell frequencies that were needed to calculate the threshold values. We manipulated the threshold values such that they yielded either symmetric or highly asymmetric categorical distributions. When the level of asymmetry was in-between the levels we manipulated, we found the corresponding RMSEA, CFI, and TLI values were also in-between the values under the manipulated conditions.

Table 2 Model fit indices calculated using the six polychoric correlation matrices in Study 2

Matrix	Target Model	Index	ML	ULS				DWLS			
				[0]	[1.5]	[– 1, 0, 1]	[0, 1, 2]	[0]	[1.5]	[– 1, 0, 1]	[0, 1, 2]
# 1	CFA one factor 7 items	RMSEA$_U$	**.245**	.038	.038	.038	.038	**.063**	.035	**.101**	**.088**
		RMSEA$_S$	–	**.081**	.044	**.132**	**.116**	**.087**	.047	**.140**	**.123**
		CFI$_U$	**.919**	.999	.999	.999	.999	.999	.999	.999	.999
		CFI$_S$	–	.995	.995	.994	.994	.996	.996	.995	.995
		TLI$_U$	**.879**	.998	.998	.998	.998	.998	.998	.998	.998
		TLI$_S$	–	.992	.993	.991	.991	.994	.994	.993	.993
# 2	CFA three factors 12 items	RMSEA$_U$	**.113**	.045	.045	.045	.045	.036	.019	.053	.047
		RMSEA$_S$	–	.045	.022	**.071**	**.062**	.046	.023	**.071**	**.062**
		CFI$_U$	**.934**	.995	.995	.995	.995	.996	.996	.996	.996
		CFI$_S$	–	.983	.986	.975	.976	.989	.991	.985	.986
		TLI$_U$	**.914**	.993	.993	.993	.993	.995	.995	.995	.995
		TLI$_S$	–	.978	.981	.967	.969	.986	.988	.981	.982
# 3	CFA four factors 12 items	RMSEA$_U$	**.096**	.049	.049	.049	.049	.037	.019	.053	.047
		RMSEA$_S$	–	.046	.049	**.072**	**.062**	.047	.022	**.073**	**.063**
		CFI$_U$	**.930**	.992	.992	.992	.992	.993	.993	.993	.993
		CFI$_S$	–	.980	.992	.970	.972	.982	.983	.974	.976
		TLI$_U$	**.903**	.989	.989	.989	.989	.990	.991	.991	.991
		TLI$_S$	–	.972	.989	.959	.962	.975	.977	.965	.967
# 4	CFA one factor 11 items	RMSEA$_U$	**.112**	.052	.052	.052	.052	.040	.021	.059	.052
		RMSEA$_S$	–	.051	.025	**.079**	**.070**	.052	.025	**.081**	**.070**
		CFI$_U$	**.907**	.991	.991	.991	.991	.992	.993	.993	.993
		CFI$_S$	–	.978	.981	.968	.970	.989	.983	.972	.973
		TLI$_U$	**.884**	.989	.989	.989	.989	.990	.991	.991	.991
		TLI$_S$	–	.972	.976	.960	.962	.975	.979	.965	.967
# 5	Structural model four factors 10 items	RMSEA$_U$	**.078**	.029	.029	.029	.029	.024	.012	.036	.031
		RMSEA$_S$	–	.034	.016	.056	.048	.034	.017	.056	.048
		CFI$_U$	.977	.999	.999	.999	.999	.999	.999	.999	.999
		CFI$_S$	–	.995	.996	.993	.993	.997	.997	.995	.996
		TLI$_U$	.965	.998	.998	.998	.998	.998	.999	.999	.999
		TLI$_S$	–	.993	.994	.989	.990	.995	.996	.993	.993
# 6	CFA one factor 5 items	RMSEA$_U$	**.078**	.039	.039	.039	.039	.030	.015	.044	.038
		RMSEA$_S$	–	.035	.017	.057	.049	.038	.019	**.063**	.054
		CFI$_U$	.982	.997	.997	.997	.997	.997	.997	.997	.997
		CFI$_S$	–	.994	.995	.991	.992	.994	.995	.992	.993
		TLI$_U$	.963	.993	.993	.993	.993	.994	.994	.994	.994
		TLI$_S$	–	.988	.989	.982	.984	.989	.990	.984	.985

RMSEA > .06, CFI < .95, and TLI < .95 are bolded.

long run to the accumulation of models with severe misfit that are nonetheless considered acceptable, even in substantive research with sufficient sample sizes.

Because in this study we have primarily compared the *asymptotic* values of the fit indices across estimation methods, the conclusions above only apply to studies with sufficient sample sizes. For example, with a sample size of 500, ULS and DWLS yield overoptimistic fit indices, as is shown in Appendix 3. The results for the asymptotic values of the DWLS fit indices are consistent with those based on finite samples in Nye and Drasgow (2011). Nye and Drasgow first employed a dataset consisting of 9,292 examinees and showed that DWLS produced smaller RMSEA$_S$ and larger CFI$_S$ and TLI$_S$ values than did ML on the basis of polychoric correlation matrices, especially when the data were dichotomized. Nye and Drasgow further implemented a simulation study to examine what cutoff values would be appropriate for DWLS fit indices under finite samples (i.e., 400, 800, and 1,600) with dichotomous data, and found that DWLS-CFI$_S$ and TLI$_S$ were mostly .99 for both moderately and severely misspecified models. However, we did not explore the results with smaller sample sizes. Savalei and Rhemtulla (2013) recommended that samples of at least 150 observations are needed for either binary or three-category data. With a smaller sample in conjunction with asymmetric categorical distributions, the expected values of fit indices across random samples might be very different from the population values of the indices.

Strong arguments against the application of RMSEA, CFI, and TLI and their conventional cutoff values have been raised in the SEM literature (e.g., Barrett, 2007; Marsh et al., 2004; McIntosh, 2007). However, before better alternatives are proposed and accepted by SEM practitioners, the application of these fit indices will continue in most SEM studies. We do not aim to directly address the question of what new cutoff values should be employed, partly because empirical Type I error rates and power have not been investigated for finite samples. However, this article delivers two messages that are related to the use of cutoff values in SEM with ordered categorical variables. First, for both ULS scaled indices and DWLS, no universal cutoff values are appropriate; the values of the indices are contingent upon the number of categories and the threshold values. The cutoff values thus also need to be specific to the thresholds, which is impractical. Second, both scaled and unscaled ULS- and DWLS-CFI and -TLI are very insensitive to model misspecification, because they mostly cluster above .95 under our manipulated conditions. In addition, in Study 2, even when the ML indices suggested severe misfit, both the unscaled and scaled ULS- and DWLS-CFI and -TLI values were still above .989.

The general consensus is that a larger RMSEA and smaller CFI and TLI values indicate worse fit, which prompts researchers to modify their models and search for a better explanation of the relationship among variables. However, the current practice has evolved into a stage at which the fit indices serve as the criteria (and the sole criteria, in many situations) for determining whether to accept or reject a hypothesized model. As long as the values of the fit indices reach a "publishable level" (e.g., RMSEA < .06), model respecification may be terminated. Given that the DWLS and ULS fit indices tend to show a better model–data fit evaluation than do ML fit indices when the same misspecified model is analyzed, we argue that surpassing a set of cutoff values should not serve as the only justification for the acceptance of a model. It would be more appropriate to consider RMSEA, CFI, and TLI as diagnostic tools for model improvement. A statement such as "because the RMSEA, CFI, and TLI values suggest good fit, this model was chosen as the final model" is not sufficient. Achieving a set of desired values of RMSEA, CFI, and TLI (e.g., according to the conventional cutoffs) is only one marker showing that the model improvement may be successful. Thereafter, one still needs to explain whether other options exist to improve the model, why the options are or are not adopted, and, as was suggested by Barrett (2007), what are the substantive scientific consequences of considering this model to be the final one.

When implementing Study 2, we initially attempted to replicate the results on the basis of the six articles we found. However, we could not achieve this purpose, because neither the threshold values nor the weight matrix used to compute the fit function was reported. We have also found that many studies have fit their models to polychoric correlation matrices but reported Pearson correlation matrices instead by treating the ordered categorical variables as continuous. Misunderstanding exists regarding the application of polychoric correlation. To improve the transparency and reproducibility of published results, we recommend that researchers report the polychoric correlation matrices associated with the threshold values (or the proportions of observed responses in each category).

When a mean- and variance-adjusted chi-square statistic is employed in WLSMV and ULSMV, software programs by default compute scaled fit indices that do not converge to the definitions of RMSEA, CFI, and TLI. Brosseau-Liard, Savalei, and Li (2012) first raised this concern using continuous data analyzed by ML with Satorra and Bentler's (1994) robust correction. Brosseau-Liard and Savalei (2014), Brosseau-Liard et al. (2012), and Xia, Yung, and Zhang (2016) showed that the scaled RMSEA, CFI, and TLI values under continuous data implemented in SEM software programs (e.g., EQS, M*plus*, and the CALIS procedure in SAS/STAT 14.1; Bentler, 2008; L. K. Muthén & Muthén, 2015; SAS Institute, 2015) converge to values that are different from the population definitions of RMSEA, CFI, and TLI. Because a similar logic lies behind the robust corrections to ULS and DWLS (e.g., Savalei, 2014) when analyzing ordered categorical variables, the scaled fit indices also converge to population values that deviate from their definitions, as we have demonstrated in Eqs. 8–13. Our study evidences that both the unscaled and scaled fit indices for ULS and DWLS can be problematic, in that they all appear to be insensitive to model misspecification if Hu and Bentler's cutoff values are applied. Future studies will need to seek alternative methods (e.g., Yuan & Marshall, 2004; Zhang, 2008) for goodness-of-fit evaluation with ordered categorical data.

Appendix 1: Additional results for Study 1

The results for the fit indices for conditions with four categories were reported in Figs. 9, 10, 11, 12, 13 and 14. The patterns were similar to the conditions with two categories. That is, ULS and DWLS produced smaller unscaled and scaled RMSEA values and larger unscaled and scaled CFI and TLI values than did ML-RMSEA$_U$.

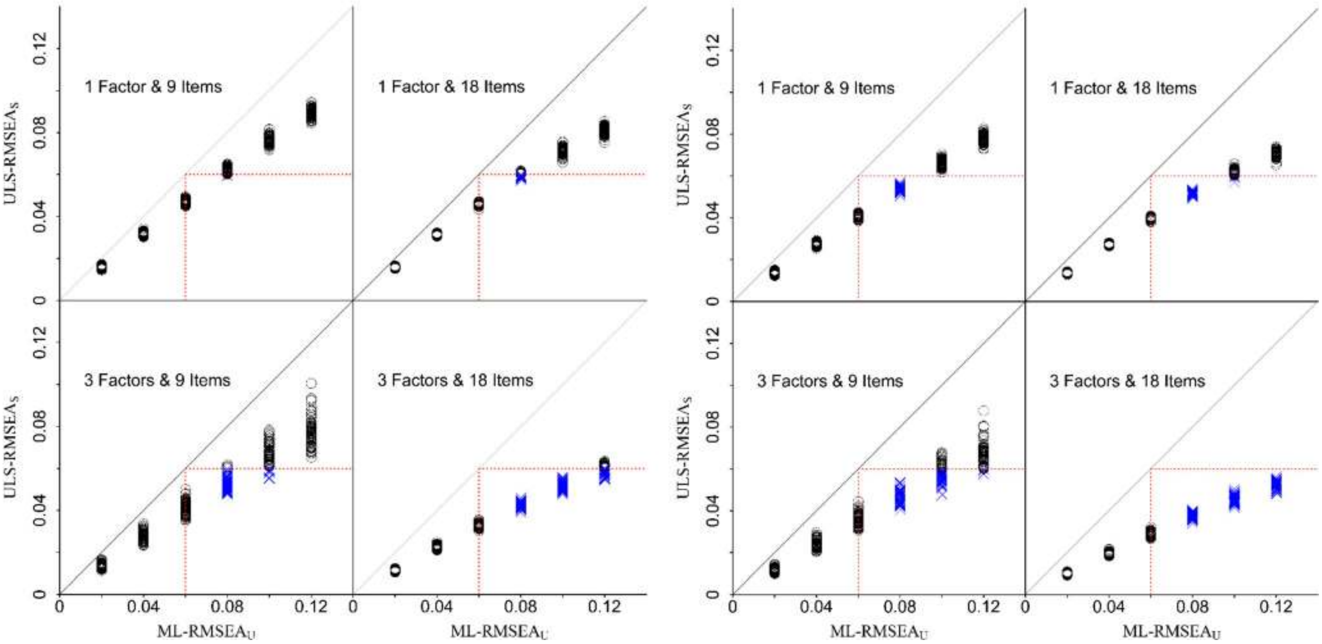

Fig. 9 Comparison between ML-RMSEA$_U$ and ULS-RMSEA$_S$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if RMSEA < .06 is used as the criterion of acceptable fit.

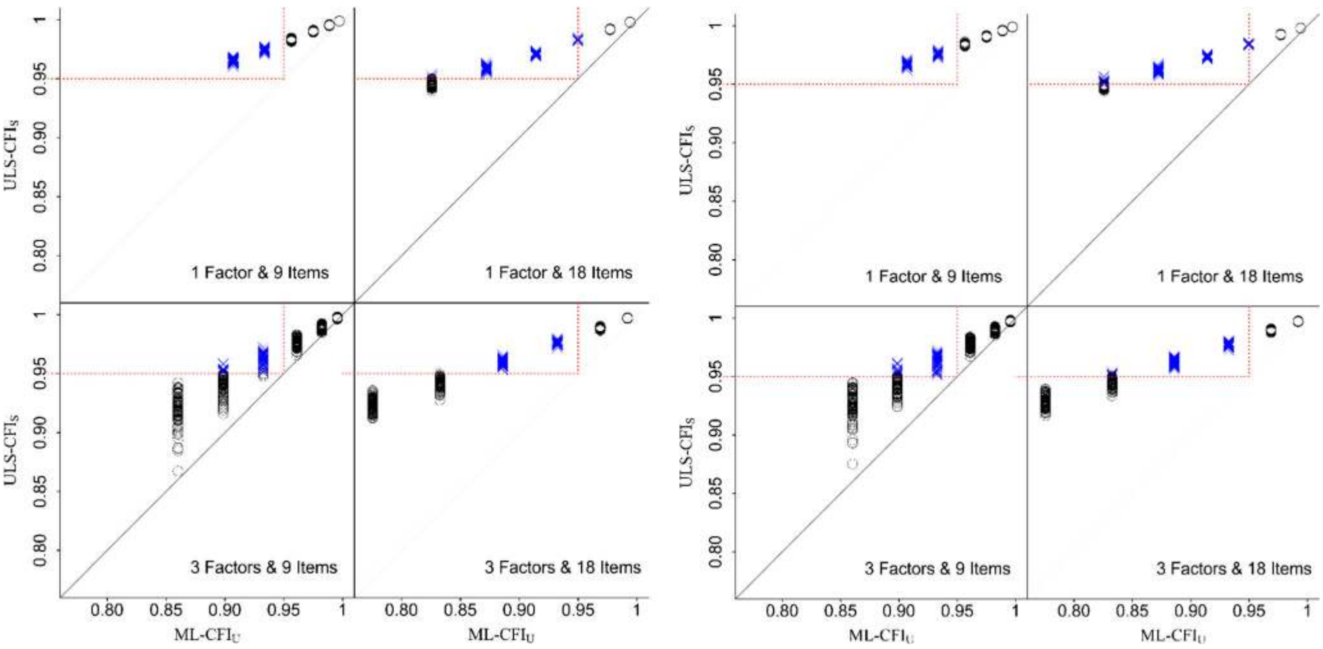

Fig. 10 Comparison between ML-CFI$_U$ and ULS-CFI$_S$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using ULS, if CFI > .95 is used as the criterion of acceptable fit.

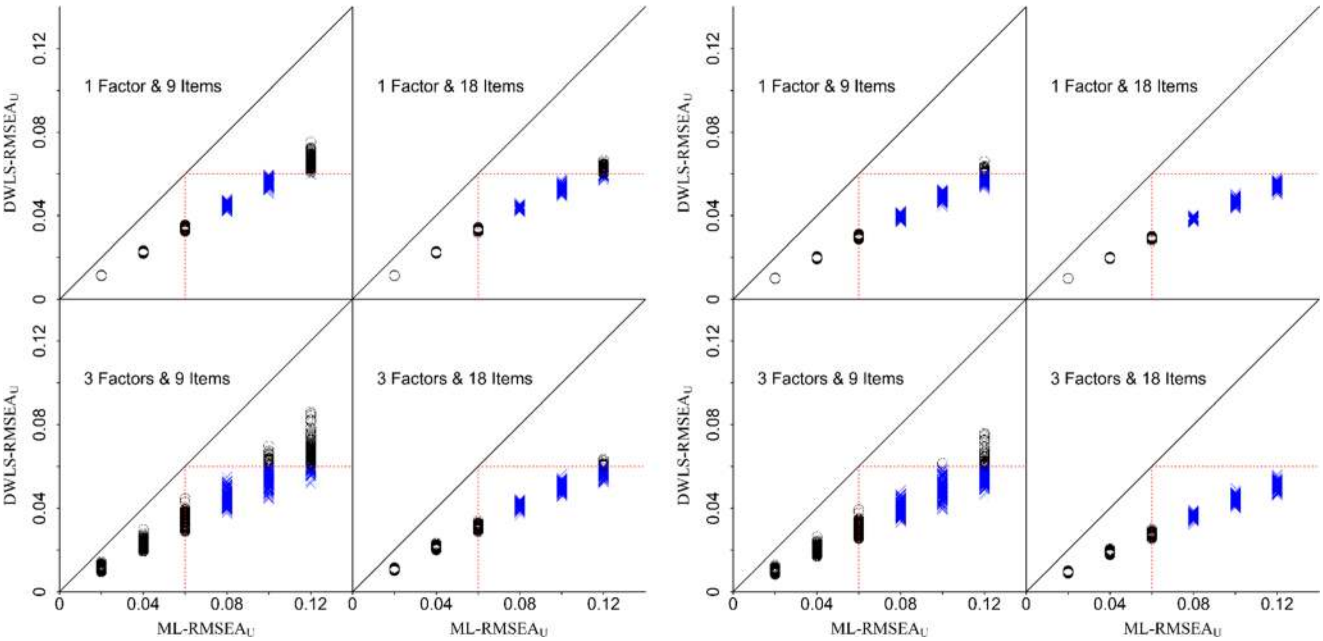

Fig. 11 Comparison between ML-RMSEA$_U$ and DWLS-RMSEA$_U$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if RMSEA < .06 is used as the criterion of acceptable fit.

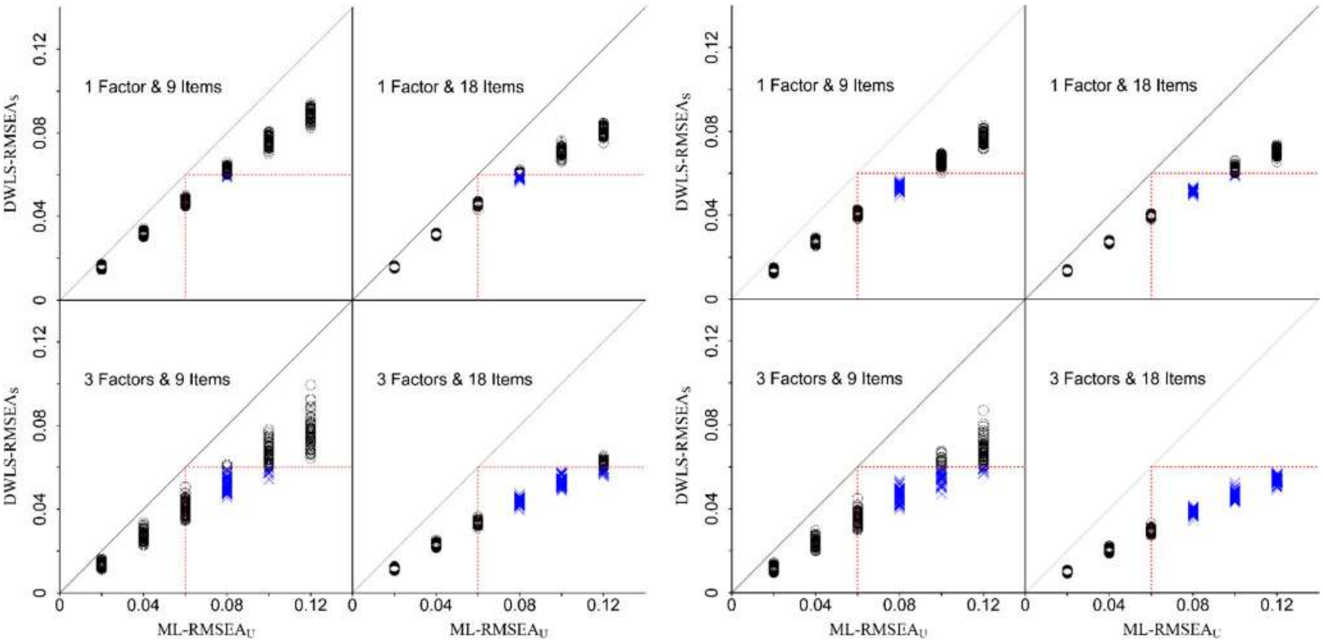

Fig. 12 Comparison between ML-RMSEA$_U$ and DWLS-RMSEA$_S$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if RMSEA < .06 is used as the criterion of acceptable fit.

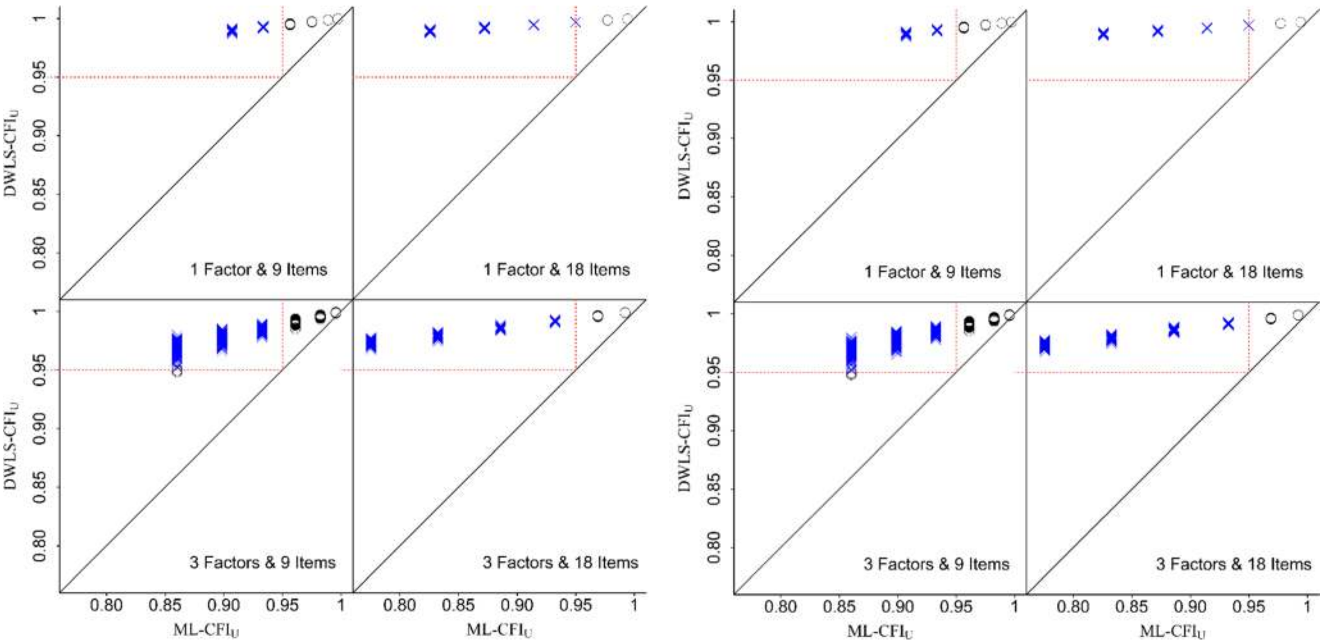

Fig. 13 Comparison between ML-CFI$_U$ and DWLS-CFI$_U$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if CFI > .95 is used as the criterion of acceptable fit.

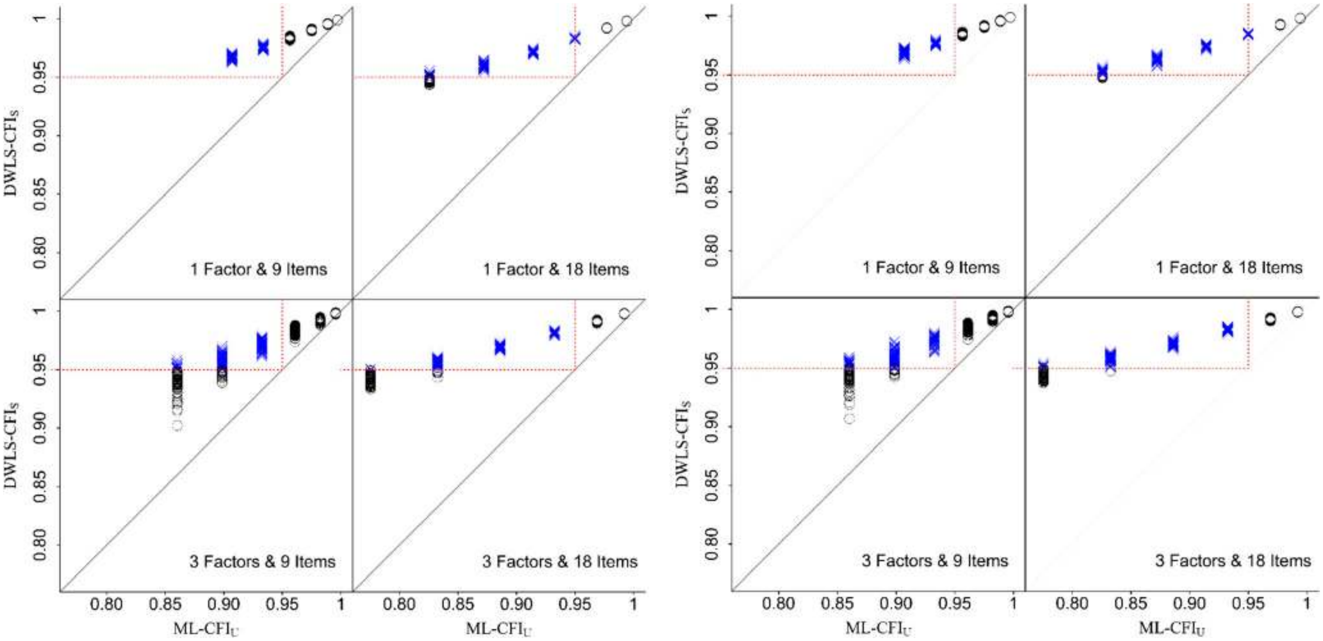

Fig. 14 Comparison between ML-CFI$_U$ and DWLS-CFI$_S$ when data are four-category. The left panel has threshold = [– 1, 0, 1], and the right panel has threshold = [0, 1, 2]. The crosses represent the matrices on the basis of which the hypothesized models are considered unacceptable using ML but acceptable using DWLS, if CFI > .95 is used as the criterion of acceptable fit.

Appendix 2: The six polychoric correlation matrices in Study 2

The six polychoric correlation matrices in Study 2 are presented in Tables 3, 4, 5, 6, 7 and 8. The analysis model for each matrix is described below. We use "F" to indicate latent factors, "V" to indicate items. "by" means "measured by" and "on" means "regressed on."

CFA model for M1:

F by V1–V7.

CFA model for M2:

F1 by V1–V5; F2 by V6–V9; F3 by V10–V12.

CFA model for M3:

F1 by V1–V3; F2 by V4–V6; F3 by V7–V9; F4 by V10–V12.

CFA model for M4:

F by V1–V11;

Nonrecursive structural model for M5:

F1 by V1–V4; F2 by V5–V6; F3 by V7–V8; F4 by V9–V10.
F2 on F1; F3 on F1; F4 on F1; F3 on F2; F2 on F3; F3 on F4.

CFA model for M6:

F by V1–V5.

Table 3 Polychoric Correlation Matrix 1 (Martínez-Rodríguez et al., 2016)

Items	1.	2.	3.	4.	5.	6.	7.
1.	1						
2.	.86	1					
3.	.80	.83	1				
4.	.84	.84	.89	1			
5.	.80	.79	.83	.94	1		
6.	.79	.79	.81	.87	.90	1	.91
7.	.79	.79	.83	.90	.86	.91	1

Table 4 Polychoric Correlation Matrix 2 (Nguyen et al., 2016)

Items	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1.	1											
2.	.59	1										
3.	.68	.59	1									
4.	.58	.58	.77	1								
5.	.65	.58	.74	.70	1							
6.	.41	.33	.43	.47	.40	1						
7.	.57	.49	.60	.57	.62	.70	1					
8.	.54	.42	.60	.54	.63	.69	.80	1				
9.	.49	.34	.58	.53	.59	.58	.69	.80	1			
10.	.37	.47	.47	.45	.36	.30	.41	.42	.33	1		
11.	.38	.46	.51	.42	.40	.34	.39	.41	.40	.83	1	.80
12.	.39	.46	.55	.44	.45	.36	.42	.40	.34	.79	.80	1

Table 5 Polychoric Correlation Matrix 3 (Pettersen et al., 2016)

Items	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1.	1											
2.	.66	1										
3.	.57	.68	1									
4.	.31	.24	.36	1								
5.	.49	.51	.55	.50	1							
6.	.34	.39	.46	.55	.56	1						
7.	.42	.43	.54	.53	.49	.56	1					
8.	.49	.55	.57	.46	.55	.60	.64	1				
9.	.47	.55	.56	.53	.52	.59	.66	.71	1			
10.	.24	.24	.28	.34	.39	.32	.36	.46	.35	1		
11.	.26	.34	.27	.38	.46	.35	.38	.47	.36	.27	1	.45
12.	.29	.31	.39	.41	.45	.41	.49	.54	.53	.38	.45	1

Table 6 Polychoric Correlation Matrix 4 (Iglesias et al., 2014)

Items	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1.	1										
2.	.36	1									
3.	.53	.50	1								
4.	.48	.28	.62	1							
5.	.42	.22	.50	.51	1						
6.	.55	.49	.60	.37	.45	1					
7.	.57	.46	.62	.64	.50	.60	1				
8.	.57	.42	.59	.50	.44	.49	.67	1			
9.	.60	.41	.63	.54	.44	.62	.65	.70	1		
10.	.40	.32	.43	.53	.33	.40	.44	.48	.49	1	.38
11.	.46	.34	.51	.50	.43	.43	.54	.51	.56	.38	1

Table 7 Polychoric Correlation Matrix 5 (Fernandez & Moldogaziev, 2013)

Items	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1.	1									
2.	.72	1								
3.	.76	.72	1							
4.	.77	.72	.74	1						
5.	.67	.64	.72	.71	1					
6.	.30	.26	.32	.31	.41	1				
7.	.55	.57	.57	.56	.51	.29	1			
8.	.64	.59	.62	.62	.56	.35	.60	1		
9.	.69	.63	.70	.73	.66	.38	.58	.62	1	.85
10.	.75	.68	.70	.75	.66	.33	.58	.70	.85	1

Table 8 Polychoric Correlation Matrix 6 (MacInnis et al., 2013)

Items	1.	2.	3.	4.	5.
1.	1				
2.	.52	1			
3.	.55	.65	1		
4.	.55	.52	.60	1	.28
5.	.24	.20	.26	.28	1

Appendix 3: Results for the sample-level simulation

Sample-level simulation was conducted in order to further explicate the differences between ML, ULS, and DWLS fit indices. We selected three of the matrices (i.e., Matrices 1, 2, and 4) in Study 2 to generate sample data, because they yielded ML-RMSEA$_U$ > .10, ML-CFI$_U$ < .95, and ML-TLI$_U$ < .95. The procedure of this simulation is described below:

- We first generated continuous data following a standard multivariate normal distribution. The continuous data were analyzed by the normal-theory ML estimator. Hu and Bentler's (1999) cutoff values based on continuous data were for an ML analysis. Therefore, we used the resulting fit indices as the benchmarks for comparison.
- We categorized the continuous data using threshold values. For two-category data, we employed [0] or [1] as the threshold value for every item. For four-category data, the thresholds were either [– 1, 0, 1] or [0, 1, 2]. The threshold manipulation was consistent with that at the population level in Study 2.
- The ordered categorical data were analyzed using the lavaan package in R, using ULSMV and WLSMV. The unscaled and scaled indices were recorded, and their means are reported in Table 9.

The number of replication was 1,000, and sample size was fixed at 500.

The results for the sample-level simulation were consistent with those of Study 2. That is, ULS and DWLS produced overoptimistic unscaled and scaled fit indices as compared with ML, especially for CFI and TLI.

Table 9 Means of model fit indices when sample size is 500

Matrix	Target Model	Index	ML	ULS				DWLS			
				[0]	[1.5]	[– 1, 0, 1]	[0, 1, 2]	[0]	[1.5]	[– 1, 0, 1]	[0, 1, 2]
# 1	CFA	RMSEA$_U$	**.245**	.007	.026	.002	.003	.052	.013	**.094**	**.080**
	one factor	RMSEA$_S$	–	**.077**	.035	**.127**	**.110**	**.082**	.038	**.135**	**.117**
	7 items	CFI$_U$	**.919**	.999	.999	.999	.999	.999	.999	.999	.999
		CFI$_S$	–	.995	.995	.994	.995	.996	.997	.996	.996
		TLI$_U$	**.879**	.999	.998	.999	.999	.999	.999	.998	.998
		TLI$_S$	–	.993	.993	.992	.992	.994	.996	.993	.994
# 2	CFA	RMSEA$_U$	**.113**	.039	**.088**	.023	.027	.018	.002	.041	.034
	three factors	RMSEA$_S$	–	.043	.018	**.068**	.059	.044	.019	**.069**	**.060**
	12 items	CFI$_U$	**.934**	.996	.978	.998	.998	.999	.999	.998	.998
		CFI$_S$	–	.984	.984	.976	.978	.990	.992	.986	.987
		TLI$_U$	**.914**	.994	.971	.998	.997	.998	.999	.997	.998
		TLI$_S$	–	.979	.980	.969	.972	.987	.990	.982	.983
# 4	CFA	RMSEA$_U$	**.112**	.051	**.098**	.038	.041	.027	.004	.049	.042
	one factor	RMSEA$_S$	–	.049	.020	**.077**	**.067**	.050	.021	**.078**	**.068**
	11 items	CFI$_U$	**.907**	.991	.966	.995	.994	.996	.999	.995	.995
		CFI$_S$	–	.978	.978	.969	.972	.981	.985	.974	.976
		TLI$_U$	**.884**	.989	.958	.994	.993	.995	.999	.993	.994
		TLI$_S$	–	.973	.974	.962	.965	.976	.981	.967	.970

RMSEA > .06, CFI < .95, and TLI < .95 are bolded.

References

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

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815–824.

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

De Beer, L. T., Pienaar, J., & Rothmann, S. (2014). Job burnout's relationship with sleep difficulties in the presence of control variables: A self-report study. *South African Journal of Psychology*, *44*, 454–466.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (2008). *EQS structural equation modeling software*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting incremental fit indexes for nonnormality. *Multivariate Behavioral Research*, *49*, 460–470.

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, *47*, 904–930.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*, 357–369.

Currier, J. M., & Holland, J. M. (2014). Involvement in abusive violence among Vietnam veterans: Direct and indirect associations with substance use problems and suicidality. *Psychological Trauma: Theory, Research, Practice, and Policy*, *6*, 73–82.

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

Fernandez, S., & Moldogaziev, T. (2013). Employee empowerment, employee attitudes, and performance: Testing a causal model. *Public Administration Review*, *73*, 490–506.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indexes really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93–111.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Iglesias, K., Burnand, B., & Peytremann-Bridevaux, I. (2014). PACIC Instrument: Disentangling dimensions using published validation models. *International Journal for Quality in Health Care*, *26*, 250–260.

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48*, 398–407.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Koziol, N. A. (2010). *Evaluating measurement invariance with censored ordinal data: A Monte Carlo comparison of alternative model estimators and scales of measurement*. Unpublished master's thesis, University of Nebraska, Lincoln, NE.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51* 220–239.

Li, C. H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* (Doctoral dissertation). Michigan State University, East Lansing, MI.

MacInnis, M. J., Lanting, S. C., Rupert, J. L., & Koehle, M. S. (2013). Is poor sleep quality at high altitude separate from acute mountain sickness? Factor structure and internal consistency of the Lake Louise Score Questionnaire. *High Altitude Medicine and Biology*, *14*, 334–337.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Martínez-Rodríguez, S., Iraurgi, I., Gómez-Marroquin, I., Carrasco, M., Ortiz-Marqués, N., & Stevens, A. B. (2016). Psychometric properties of the Leisure Time Satisfaction Scale in family caregivers. *Psicothema*, *28*, 207–213.

Maydeu-Olivares, A., Fairchild, A. J., & Hall, A. G. (2017). Goodness of fit in item factor analysis: Effect of the number of response alternatives. *Structural Equation Modeling*, *24*, 495–505.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, *42*, 859–867.

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

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nguyen, T. Q., Poteat, T., Bandeen-Roche, K., German, D., Nguyen, Y. H., Vu, L. K., … Knowlton, A. R. (2016). The internalized homophobia scale for Vietnamese sexual minority women: Conceptualization, factor structure, reliability, and associations with hypothesized correlates. *Archives of Sexual Behavior*, *45*, 1329–1346.

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*, 548–570.

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

Pearson, K. (1904). *Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation* (Drapers' Co. Research Memoirs, Biometric Series, no. 1). Cambridge, UK: Cambridge University Press.

Pettersen, C., Nunes, K. L., & Cortoni, F. (2016). Does the factor structure of the aggression questionnaire hold for sexual offenders? *Criminal Justice and Behavior*, *43*, 811–829.

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

SAS Institute. (2015). *SAS/STAT 14.1 user's guide*. Cary, NC: Author.

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

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

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

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*, 201–223.

Stander, F. W., Mostert, K., & de Beer, L. T. (2014). Organisational and individual strengths use as predictors of engagement and productivity. *Journal of Psychology in Africa*, *24*, 403–409.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors.* Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, *18*, 342–353.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Xia, Y., Yung, Y. F., & Zhang, W. (2016). Evaluating the selection of normal-theory weight matrices in the Satorra–Bentler correction of chi-square and standard errors. *Structural Equation Modeling*, *23*, 585–594.

Yang, Y., & Xia, Y. (2015). On the number of factors in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*, 756–772. https://doi.org/10.3758/s13428-014-0499-2

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

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model-fit indexes for latent variable models with binary and continuous outcomes* (Doctoral dissertation). University of California, Los Angeles, CA.

Yuan, K. H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika*, *31*, 67–90.

Zhang, W. (2008). A comparison of four estimators of a population measure of model fit in covariance structure analysis. *Structural Equation Modeling*, *15*, 301–326.