Fitting Large Factor Analysis Models With Ordinal Data

Educational and Psychological
Measurement

2019, Vol. 79(3) 417–436

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164418818242

journals.sagepub.com/home/epm

Christine DiStefano¹, Heather L. McDaniel¹,
Liyun Zhang¹, Dexin Shi¹ and Zhehan Jiang²

Abstract

A simulation study was conducted to investigate the model size effect when confirmatory factor analysis (CFA) models include many ordinal items. CFA models including between 15 and 120 ordinal items were analyzed with mean- and variance-adjusted weighted least squares to determine how varying sample size, number of ordered categories, and misspecification affect parameter estimates, standard errors of parameter estimates, and selected fit indices. As the number of items increased, the number of admissible solutions and accuracy of parameter estimates improved, even when models were misspecified. Also, standard errors of parameter estimates were closer to empirical standard deviation values as the number of items increased. When evaluating goodness-of-fit for ordinal CFA with many observed indicators, researchers should be cautious in interpreting the root mean square error of approximation, as this value appeared overly optimistic under misspecified conditions.

Keywords

ordinal data, WLSMV, large models

It is well known that estimation is an important topic in structural equation modeling (SEM), as the choice of estimator is responsible for producing parameter estimates, standard errors of parameter estimates, and estimates of fit—all of which are used for model evaluation. However, under many commonly encountered empirical research situations (e.g., use of ordinal data, nonnormally distributed data), some estimation techniques may encounter shortcomings which limit the ability to produce accurate

¹University of South Carolina, Columbia SC, USA

²University of Alabama, Tuscaloosa, AL, USA

Corresponding Author:

Christine DiStefano, University of South Carolina, 138 Wardlaw Hall, Columbia, SC 29208, USA.

Email: distefan@mailbox.sc.edu

estimates. To deal with such situations, robust corrections have become increasingly popular (Lei & Wu, 2012; Wirth & Edwards, 2007). Robust corrections are applied to an estimator to account for asymptotic inefficiency that arises when estimation is conducted under suboptimal conditions (e.g., nonnormality in data used with a normal theory estimator, use of an estimator that employs a diagonal weight matrix; Savalei, 2014).

The term *diagonal weighted least squares* (DWLS) has been used to broadly define a collection of estimation techniques for ordinal data that use the diagonal of the weight matrix for inversion and the full weight matrix to estimate parameter estimates and standard errors (Bandalos, 2008; Savalei, 2014). We use DWLS to broadly refer to the weighted least squares (WLS) estimator plus a correction for the mean (i.e., WLSM), or mean and variance (WLSMV), as the practice of utilizing only the diagonal of the WLS matrix with zeros in the off-diagonal elements is not recommended (Savalei, 2012). DWLS estimation with robust correction is recommended with ordinal data, especially if fewer than five ordered categories are present (DiStefano & Morgan, 2014; Muthén & Kaplan, 1992; Rhemtulla, Brosseau-Liard, & Savalei, 2012). When analyzing ordinal data, DWLS methods have been found to have additional advantages besides providing more accurate χ^2 values, including lower bias observed in standard errors of parameter estimates and improved performance of fit indices (DiStefano & Morgan, 2014). Additionally, DWLS estimation can be used to accommodate larger models with smaller sample sizes than were possible under full WLS (e.g., Bandalos, 2008; DiStefano & Morgan, 2014; Flora & Curran, 2004). Far fewer studies, however, have investigated effects of DWLS corrections with larger models and ordinal data. While model size may be interpreted as the number of items or the number of latent variables, we define a large model size by the number of items included in a study (Kenny & McCoach, 2003; Shi, Lee, & Terry, 2015).

Study of many ordinal variables is of interest as procedures used in the past to accommodate large numbers of ordinal variables (e.g., parceling), may no longer be necessary given ease of access to DWLS methods (Bandalos, 2008). In addition, the ability of estimation techniques to accommodate ordinal data allows a direct parallel to many common standardized testing situations where a large number of items are included on an assessment (e.g., licensure tests, state testing programs). Such situations are typically analyzed with item response theory (Forero & Maydeu-Olivares, 2009), but could alternatively be analyzed with SEM.

In the empirical literature, analysis of large models with SEM techniques is not uncommon (e.g., Kranzler, Fehling, Anestis, & Selby, 2016; Levant, Hall, Weigold, & McCurdy, 2016; McWhirter, Hackett, & Bandalos, 1998). However, many researchers have used normal theory estimators (e.g., maximum likelihood [ML]), even when ordinal data were present (e.g., Cinamon, 2016; Lau et al., 2016). When reviewing characteristics of empirical studies analyzing models with many variables, some commonalities emerged. For example, the selection of estimator was mainly based on considerations related to missing data treatments (e.g., full information ML

estimation), and limited focus was placed on the metric level of the analyzed data. In addition, distributional normality of data were not discussed (e.g., Cinamon, 2016; Kranzler et al., 2016). Given that the need to test models with many items is not uncommon and that most studies in social sciences use ordinal data, information regarding the performance of WLSMV with large models is of importance.

Structure of DWLS Estimation

In its general form, the WLS fit function may be written as $F = (\mathbf{s} - \hat{\boldsymbol{\sigma}})' \mathbf{W}^{-1} (\mathbf{s} - \hat{\boldsymbol{\sigma}})$, where $\mathbf{s}$ represents a vector of the nonduplicated elements in the sample covariance matrix ($\mathbf{S}$), $\hat{\boldsymbol{\sigma}}$ is a vector of the nonduplicated elements in the model implied covariance matrix, $[\sum(\hat{\boldsymbol{\theta}})]$, and $\mathbf{s} - \hat{\boldsymbol{\sigma}}$ represents a residual vector of the discrepancies between sample and model implied values (Finney & DiStefano, 2013). When considering ordinal data, the WLS formula is often written as: $F = (\mathbf{r} - \hat{\boldsymbol{\rho}})' \mathbf{W}^{-1} (\mathbf{r} - \hat{\boldsymbol{\rho}})$, where input consists of polychoric correlations and the objective is to replicate the correlation input through the relations identified with the estimated model (i.e., model-implied values). Residual vectors are weighted by an asymptotic covariance matrix, which contains the variance-covariances of the estimated variances and covariances while the weight matrix ($\mathbf{W}$) provides information about the sampling variability associated with the sample estimates (Wirth & Edwards, 2007). Under DWLS, the weight matrix contains only the diagonal elements of the asymptotic covariance matrix. Thus, the weight matrix used in the fit function is a simplified version of the full asymptotic covariance matrix. Although only the diagonal is inverted with the weight matrix ($\mathbf{W}^{-1}$), information from the full asymptotic covariance matrix is used to more accurately estimate the standard errors of parameter estimates and χ^2 (Wirth & Edwards, 2007). Corrections applied with DWLS estimators typically include a mean shift adjustment for nonnormality to compute a value which reflects the mean of the central χ^2 distribution (WLSM) or a correction to adjust the distribution of the χ^2 to reflect both the mean and the variance of the central χ^2 distribution (WLSMV). It is noted that parameter estimates under the different DWLS techniques are the same, but standard errors of parameter estimates, χ^2 values, and fit indices that use χ^2 in their calculations will likely differ among methods. Our focus is on the WLSMV estimator, as this is typically the default estimation technique when ordinal data are analyzed (e.g., lavaan package, R software, Rosseel, 2012; Mplus v. 8.1, Muthén & Muthén, 1998-2018).

Effects of Model Size on Model Results

Previous studies that have included many items in the tested model (e.g., Kenny & McCoach, 2003) have largely used ML-based estimators and focused investigations on the χ^2 test of global fit. Simulation studies have found an inflated χ^2 statistic, with the amount of Type I error increasing as the number of variables included in the model increased (Anderson & Gerbing, 1984; Beauducel & Herzberg, 2006; Herzog,

Boomsma, & Reinecke, 2007; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998; Moshagen, 2012; Shi, Lee, & Terry, 2018). A similar effect was observed with increasing model size when data were categorical and the “full” WLS technique was utilized as the estimator (Flora & Curran, 2004; Potthast, 1993).

Under WLSMV estimation, research investigating regarding the impact of large models has been mixed. In line with studies using the ML estimator, some studies found the WLSMV-based χ^2 global fit statistic to increase (i.e., smaller p value) as the size of the model increased (Bandalos, 2014; Beauducel & Herzberg, 2006; Flora & Curran, 2004). Conversely, others have found χ^2 rejection rates to drop below the nominal level as model size increased (Shi, DiStefano, McDaniel, & Jiang, 2018). Increasing the number of observed variables per factor has been associated with more accurate parameter estimation under ML, suggesting that there may be a trade-off with fit as model size increases (Marsh et al., 1998). In other words, increasing the number of parameters in the model has not only been associated with greater accuracy in parameter estimates but has also produced worse fit. It is unclear if this trade-off exists with large models analyzed via WLSMV.

In addition, when categorical data were analyzed, few studies have examined the impact of model size on parameter estimates, standard errors of parameter estimates, or fit indices other than χ^2 . In terms of parameter values, Bandalos (2014) did not notice an effect of model size with WLSMV producing loading estimates with little bias; however, factor correlation estimates reported overestimation. Beauducel and Herzberg (2006) found similar results and found the amount of overestimation was reduced when there were more parameters in the model. Considering standard error of parameter estimates, WLSMV-based estimates did not differ greatly from ML-based estimates; instead, asymmetric data increased the amount of negative bias more than model size (Bandalos, 2014). Beauducel and Herzberg (2006) showed root mean square error of approximation (RMSEA) to increase with an increasing number of items, suggesting that increased model size was associated with an indication of poorer model fit.

In sum, while there are advantages of using WLSMV with ordinal data, far fewer studies have investigated the impact of the estimation technique on results when model size (based on the number of items) is large. Given the ease of access to the WLSMV technique, there is a gap in the literature in this area. The purpose of the study was to investigate large models using many indicators to investigate the impact of model size on parameter estimates, standard error of parameter estimates and selected fit indices.

Method

A simulation study was conducted using a confirmatory factor analysis (CFA) model; CFA was selected, as it is the most widely used SEM model in empirical studies (Maydeu-Olivares, 2017). The WLSMV estimator and the Mplus software package (v. 7.4; L. K. Muthén & Muthén, 1998-2015) were used for all analyses.

The population model consisted of a congeneric CFA with three factors. Loading values were held constant at a standardized value of 0.70 and correlation between factors was set at 0.40; factor variances were set to 1.0 for identification. Model size, nonnormality of item-level distributions, sample size ratio, and model misspecification were varied to examine the impact on recovery of parameter estimates, standard errors, power, and tests of model fit. Details of the simulation study are provided.

Number of Observed Items

The total number of items included five conditions: 15, 36, 60, 90, and 120 ordinal items, with an equal number of items loading on each of the three factors. Specifically, five, 12, 20, 30, and 60 items per factor were modeled.

Tested Models

The three-factor model was analyzed under two conditions: a correctly specified model, where all paths in the estimated model were specified to match the population model, and a misspecified model, where the correlation between two factors was collapsed, resulting in a two-factor model. The correctly specified model may be considered as a baseline for comparisons. The misspecified model condition was selected to match a situation where a larger “test” is analyzed. We recognize that this is a relatively severe level of misspecification, as the factors have little overlap ($r = 0.40$); however, this condition can inform researchers of the effects of a severe misspecification on the selected outcomes.

Number of Categories and Item Distributions

Using the population characteristics, multivariate normal data were generated and categorized by applying threshold values. Two conditions were created: two ordered categories (i.e., dichotomous data) or five ordered categories. The dichotomous level condition was selected given that many standardized measures (e.g., achievement tests) produce dichotomous data. Five categories were selected as this is a popular choice for many survey-type questionnaires (e.g., Likert-type scales).

Item threshold values were selected to achieve target levels of asymmetry, as noted by the percentage of cases responding per category. Item distributions were considered either (approximately) symmetric or asymmetric. Symmetric conditions considered roughly 50% of cases in each category for dichotomous data and for five-category data, 6%, 20%, 48%, 20%, and 6% per category, respectively. The asymmetric condition consisted of approximately 10% and 90% for the two-category situation and 2%, 4%, 8%, 16%, and 70% per category under five-category data. Three distributional conditions were tested: (1) all items symmetric, (2) half symmetric/half asymmetric, and (3) all items asymmetrically distributed. In the half symmetric/half asymmetric situation, items with different distributions were included on the

same factor. When there were five items per factor, three items were considered asymmetric, and two symmetric; equal number of symmetric/asymmetric items were present for remaining conditions.

Sample Size

Sample size is important to ensure stable results and power, given factors such as model size and loading strength (Wolf, Harrington, Clark, & Miller, 2013). With a large number of items to be analyzed, we felt a constant sample size may be unreasonable. Therefore, $N:p$ ratios of 10:1 and 20:1 were selected, where the number of cases varied depending on the number of items included in the model, yet the ratio remained constant. The 10:1 sample size condition yielded samples of size 150, 360, 600, 900, and 1200 and the 20:1 condition resulted in sample sizes of 300, 720, 1200, 1800, and 2400. We recognize that the large model size in concert with the other design factors may yield unstable results, especially at smaller sample sizes; however, lower $N:p$ ratios may be encountered in applied settings.

In summary, the simulation study consisted of a fully crossed design with 120 cells: 5 levels of model size (i.e., 15, 36, 60, 90, 120 items) $\times$ 2 tested models (correctly specified, misspecified) $\times$ 2 category levels (2-category, 5-category) $\times$ 3 distributional conditions (all symmetric, half symmetric/half asymmetric, all asymmetric) $\times$ 2 sample sizes (10:1, 20:1). One thousand replications were run for each design cell; replications that showed nonconvergence or improper solutions were removed from further analysis.

Data Analysis

To determine the impact of the simulation conditions, we examined recovery of parameter estimates (i.e., factor loadings and factor correlations), standard errors of parameter estimates, power of the χ^2 to reject a misspecified model, and select model fit indices (RMSEA and weighted root mean square residual [WRMR]). To examine the recovery of parameter estimates, first, the average factor loading and factor correlation estimates per cell were computed and compared with the population values. The difference between computed average values and population values were multiplied by 100 to obtain the percentage of relative bias (RB; Bandalos & Leite, 2013). Positive bias values indicated overestimation of the parameter estimate; negative values indicated underestimation of the population value.

In addition, the pooled standard deviation of factor loading estimates and correlation estimates were computed for each cell in the design. This value provides an estimate of the variability across the parameter estimates (Bandalos & Leite, 2013). For standard errors, the mean relative bias was calculated by comparing the mean estimated standard error ($\overline{SE}_\theta$) to the standard deviations of the estimates (sd_θ) over the replications: $\frac{\overline{SE}_\theta - sd_\theta}{sd_\theta}$ (Bandalos & Leite, 2013). Relative bias percentage levels less than 10% were judged to be at an acceptable level (e.g., Hoogland & Boomsma,

1998; Shi, Song, & Lewis, 2017) and values less than 5% were considered trivial bias.

We also examined the p values associated with the χ^2 test. For correctly specified models, these values provide empirical Type I errors; for the misspecified model, p values provide empirical power rates. We evaluated the Type I error rates against a nominal alpha of .05 and power against an ideal cutoff of .80. Average estimates of the RMSEA were analyzed as this index is often recommended for use with ordinal data (e.g., Bandalos, 2014). Values of .08 or lower for RMSEA average demonstrated “adequate” fit and .06 or lower as “good” fit (Hu & Bentler, 1999). In addition, the WRMR was investigated, as this index was developed for use with ordinal data. Values of WRMR equal to or less than 1 indicated good fit (DiStefano, Liu, Jiang, & Shi, 2018).

Results

Prior to summarizing results, we examined the number of admissible solutions by cell. Overall, admissible solution rates were generally high—except when five-category, asymmetric data were present with fewer variables (i.e., 15) and a small sample size (small $N:p$ ratio). While the number of inadmissible solutions was higher under asymmetric data conditions (49% of replications converged), convergence problems were also observed when half of the data were asymmetric (65% convergence rate). No convergence problems were observed when asymmetric data were analyzed with many variables ($p > 36$) or when sample size ($N:p$ ratio) increased. In such situations, convergence rates were equal to or greater than 97%. Replications with inadmissible solutions were removed and only cells exhibiting proper solutions were included in the summarization. Admissible solution rates are provided in Table 1.

Relative Bias in Parameter Estimates

Table 1 provides the percentage of relative bias in factor loading estimates and factor correlation(s) for correctly specified and misspecified models. Considering correctly specified models, the relative bias observed with factor loading estimates was trivial—less than 1% for all situations. Factor correlations were also acceptable, illustrating small amounts of relative bias across all design cells. Larger levels of positive bias were observed when at least half of the variables were asymmetric and dichotomous data were present; however, even in such situations, the amount of relative bias was still negligible. These results mirror findings of other researchers, who have noted WLSMV parameter estimates to be accurate when few categories and/or non-normal data are analyzed, and models are correctly specified (e.g., Bandalos, 2014; DiStefano & Morgan, 2014; Flora & Curran, 2004). Here, findings are extended to situations where symmetric and asymmetric data are analyzed on the same factor. For both factor loading and factor correlation parameters, the amount of relative bias

Table 1. Admissible Solution Percentages, Parameter Estimates.

Level	Ratio	p	Correctly Specified						Misspecified					
			C = 2			C = 5			C = 2			C = 5		
			ADM	LD	COR	ADM	LD	COR	ADM	LD	COR	ADM	LD	COR
Normal	10	15	100	0.45	3.62	100	0.30	1.59	100	-9.67	22.07	100	-9.63	19.52
		36	100	0.31	1.07	100	0.20	0.52	100	-9.11	17.65	100	-8.84	16.18
		60	100	0.16	0.83	100	0.09	0.33	100	-9.05	16.91	100	-8.73	15.55
	20	90	100	0.10	0.30	100	0.10	0.10	100	-9.10	16.30	100	-8.70	15.10
		120	100	0.10	0.40	100	0.10	0.20	100	-9.00	16.20	100	-8.60	15.20
		15	100	0.26	1.76	100	0.14	0.86	100	-9.91	19.96	100	-9.70	18.34
Half normal	10	36	100	0.13	0.78	100	0.09	0.40	100	-9.28	17.26	100	-8.92	16.27
		60	100	0.07	0.39	100	0.04	0.09	100	-9.15	16.33	100	-8.75	15.14
		90	100	0.00	0.00	100	0.00	-0.10	100	-9.10	16.00	100	-8.70	15.00
		120	100	0.10	0.30	95	0.00	0.10	100	-9.00	16.00	91	-8.60	15.10
		15	100	0.39	8.41	65	0.75	4.12	100	-10.00	29.38	65	-9.19	22.42
		36	100	0.89	2.25	99	0.24	0.65	100	-8.70	19.54	99	-8.93	16.66
	20	60	100	0.54	1.14	100	0.09	0.36	100	-8.80	17.48	100	-8.84	15.75
		90	100	0.40	0.50	100	0.10	0.20	100	-8.90	16.80	100	-8.80	15.50
		120	100	0.40	0.60	100	0.10	0.30	100	-8.90	16.80	100	-8.70	15.50
		15	100	0.79	3.47	97	0.11	1.31	100	-9.62	23.30	97	-9.88	19.38
		36	100	0.42	1.23	100	0.08	0.47	100	-9.14	18.22	100	-9.03	16.20
		60	100	0.26	0.60	100	0.05	0.13	100	-9.09	16.93	100	-8.86	15.41
		90	100	0.20	0.10	100	0.00	0.00	100	-9.10	16.50	100	-8.80	15.30
		120	100	0.10	0.30	100	0.10	0.20	100	-9.10	16.60	100	-8.70	15.40

(continued)

Table 1. (continued)

Level	Ratio	<i>p</i>	Correctly Specified						Misspecified					
			C = 2			C = 5			C = 2			C = 5		
			ADM	LD	COR	ADM	LD	COR	ADM	LD	COR	ADM	LD	COR
Nonnormal	10	15	100	-0.63	7.03	49	0.76	5.95	100	-10.45	25.18	49	-8.82	23.92
		36	100	0.63	4.46	98	0.19	0.65	100	-7.67	19.15	98	-8.79	16.35
		60	100	0.31	2.04	100	0.03	0.33	100	-7.93	16.28	100	-8.73	15.41
	20	90	100	0.30	1.30	100	0.10	0.20	100	-7.90	15.30	100	-8.60	15.10
		120	100	0.20	1.00	100	0.10	0.40	100	-8.00	15.30	100	-8.60	15.00
		15	100	-0.04	4.55	97	0.01	1.34	100	-9.17	21.50	97	-9.75	18.96
		36	100	0.27	1.96	100	0.06	0.33	100	-8.15	16.35	100	-8.88	15.63
		60	100	0.15	0.97	100	0.02	0.16	100	-8.16	15.22	100	-8.72	15.12
		90	100	0.10	0.60	100	0.00	0.00	100	-8.10	14.80	100	-8.60	15.00
		120	100	0.10	0.50	100	0.10	0.30	100	-8.10	14.90	100	-8.60	15.20

Note. Bias levels (greater than/equal sign) 10.00 are shown in boldface. Level = distributional characteristics of item level data; Ratio = *N/p* ratio; *p* = number of indicators; ADM = percentage of admissible solutions; LD = percentage of relative bias in factor loading estimates; COR = percentage of relative bias in factor correlation estimates.

decreased as model size increased; however, the decrease was more pronounced for factor correlations.

When examining relative bias in parameter estimates for misspecified models, patterns were similar to those of correctly specified models. Considering bias in factor loadings, values were underestimated and, as found in prior studies, the amount of RB observed was still negligible (i.e., less than 10%; Bandalos, 2008; Flora & Curran, 2004). Relative bias approached the 10% cutoff in many conditions, but only exceeded the cutoff when analyzing dichotomous data, with small $N:p$ ratio, few variables, and asymmetric data. Factor correlations were consistently overestimated for all situations, with the highest levels of bias approximating 20% overestimation when fewer variables ($p = 15$) were analyzed. Also, the amount of relative bias in the factor correlations was slightly more pronounced when two-category data were analyzed. As with correctly specified models, relative bias in both factor loading and correlation estimates decreased as model size increased; however, bias levels in the factor correlations remained sizable for all conditions.

Standard Error of Parameter Estimates

Table 2 provides the average of the relative bias of the standard errors of parameter estimates across all conditions. As with other studies involving DWLS estimators, standard errors of the parameter estimates reported negative bias (DiStefano, 2002; DiStefano & Morgan, 2014; Flora & Curran, 2004; Lei, 2009; Potthast, 1993; Yu & Muthén, 2002). Table 2 shows that all relative biases are negative, meaning that the estimated standard errors were smaller than the associated empirical standard deviations. Under correctly specified models, the relative bias of the standard error of factor loadings was generally observed to be low ($RB \leq |.10|$); however, standard errors were noticeably underestimated with conditions when the $N:p$ ratio was small, there were few observed indicators, and at least half of items followed asymmetric distributions. In addition, under misspecified conditions, the relative bias of standard error for factor loadings was consistently underestimated when five-category data were analyzed.

Regarding the average relative bias of standard error for the factor correlations, values were substantially underestimated in conditions in which there were few observed variables. In fact, across correctly specified conditions in which there were 15 observed indicators, the average relative bias observed with the factor correlation standard error term was problematic for the majority of conditions. As the number of observed items increased, relative bias of standard error for factor correlations decreased greatly in magnitude and was generally acceptable (except in the case of five-category data, small $N:p$, and all data following nonnormal distributions). Thus, for factor correlations, increasing the number of items past 15 reduced bias to a mild level of underestimation in the standard errors. Findings for factor correlations under misspecified conditions followed a similar pattern.

Table 2. Percentage of Bias, Standard Errors of Parameter Estimates.

Level	Ratio	<i>p</i>	Correctly Specified				Misspecified			
			C = 2		C = 5		C = 2		C = 5	
			LD	COR	LD	COR	LD	COR	LD	COR
Normal	10	15	-5.39	-6.48	-7.10	-50.84	-14.12	-8.71	-17.95	-42.16
		36	-2.38	-4.22	-4.00	-6.83	-9.86	-4.40	-13.97	-6.53
		60	-1.30	-2.06	-1.79	-4.12	-8.78	-1.19	-11.51	-4.81
		90	-0.70	-2.70	-1.00	-2.10	-7.70	-4.50	-10.50	-4.30
		120	-0.60	0.90	-0.80	-0.40	-7.70	1.70	-10.70	-0.60
	20	15	-2.54	-3.71	-3.31	-70.42	-11.10	-5.87	-14.19	-57.34
		36	-0.93	-2.85	-1.58	-4.04	-8.47	-3.28	-11.15	-5.01
		60	-0.59	-2.65	-0.95	-3.85	-8.08	-3.36	-10.66	-4.59
		90	-0.30	-3.20	-0.20	-1.40	-7.90	-4.10	-10.00	-1.70
		120	-0.20	1.70	-0.40	0.30	-7.20	1.20	-10.10	-0.60
Half normal	10	15	-9.47	-12.83	-10.61	-67.78	-17.12	-56.91	-20.94	-79.96
		36	-4.53	-5.67	-5.61	-8.60	-10.51	-5.84	-14.73	-8.47
		60	-2.90	-6.20	-2.46	-6.33	-8.85	-7.22	-11.58	-7.26
		90	-7.00	-2.00	-3.40	-2.40	-11.90	-4.30	-12.10	-4.80
		120	-6.00	8.40	-2.90	-1.70	-11.50	-0.10	-12.10	-2.10
	20	15	-4.77	-9.09	-5.65	-75.12	-11.52	-78.23	-16.07	-87.85
		36	-2.56	-3.29	-2.67	-4.52	-8.66	-2.93	-12.10	-3.30
		60	-1.62	-4.77	-1.25	-4.81	-7.55	-4.64	-10.54	-5.05
		90	-5.90	-2.00	-2.20	-2.10	-11.10	-1.70	-11.00	-2.60
		120	-6.80	3.30	-2.40	-0.60	-10.80	-1.10	-11.60	-2.40
Nonnormal	10	15	-19.38	-16.82	-12.81	-67.16	-25.79	-30.08	-21.74	-83.60
		36	-7.74	-7.75	-7.32	-9.92	-13.34	-8.10	-15.61	-10.42
		60	-4.10	-5.16	-3.80	-8.17	-9.92	-8.13	-12.12	-10.29
		90	-2.90	-3.50	-2.50	-4.30	-9.10	-4.70	-11.10	-6.30
		120	-2.40	-1.70	-1.90	-3.50	-8.80	4.10	-10.20	-3.80
	20	15	-8.30	-18.16	-7.54	-68.68	-15.13	-42.70	-17.03	-83.98
		36	-3.90	-6.99	-3.47	-5.96	-10.26	-5.07	-12.22	-5.79
		60	-2.38	-4.26	-1.97	-5.55	-8.62	-5.85	-10.64	-6.25
		90	-1.30	-2.30	-0.80	-2.70	-7.60	-1.50	-9.50	-2.20
		120	-1.30	-0.90	-0.80	-1.40	-8.20	-3.70	-10.00	-4.70

Note. Bias levels (greater than/equal sign) 10.00 are shown in boldface. Level = distributional characteristics of item level data; Ratio = *N:p* ratio; *p* = number of indicators; LD = percentage of relative bias in standard errors of factor loading estimates; COR = relative bias in standard errors of factor correlation estimates.

Type I Errors and Normed χ^2

Table 3 provides empirical Type I errors for the χ^2 goodness-of-fit statistic. Because models included differing numbers of items and thus, different expected values, the normed χ^2 value (i.e., $\chi^2/\text{degrees of freedom}$) was computed for each condition, where values lower than 2 were representative of acceptable fit (Ullman, 2001). Again, trends in the results are observed. For example, all correctly specified models

Table 3. Average χ^2 Values, Normed χ^2 Values, and Average Numbers of Rejections per Cell.

Level	Ratio	<i>p</i>	Correctly Specified						Misspecified					
			C = 2			C = 5			C = 2			C = 5		
			χ^2	χ^2/df	rej	χ^2	χ^2/df	rej	χ^2	χ^2/df	rej	χ^2	χ^2/df	rej
Normal	10	15	89	1.03	0.04	91	1.05	0.08	161	1.81	0.98	244	2.75	1.00
		36	597	1.01	0.02	601	1.02	0.05	1328	2.24	1.00	2044	3.45	1.00
		60	1718	1.01	0.02	1723	1.01	0.04	4398	2.57	1.00	6762	3.96	1.00
		90	3921	1.00	0.00	3932	1.01	0.00	11328	2.89	1.00	17227	4.40	1.00
		120	7028	1.00	0.00	7038	1.00	0.00	21986	3.13	1.00	32933	4.69	1.00
	20	15	88	1.01	0.05	89	1.02	0.06	240	2.70	1.00	404	4.53	1.00
		36	596	1.01	0.04	595	1.01	0.04	2236	3.77	1.00	3717	6.27	1.00
		60	1714	1.00	0.03	1719	1.01	0.04	7848	4.59	1.00	12619	7.38	1.00
		90	3917	1.00	0.00	3921	1.00	0.00	20921	5.35	1.00	32512	8.31	1.00
		120	7022	1.00	0.00	7029	1.00	0.00	41326	5.89	1.00	62423	8.89	1.00
Half normal	10	15	89	1.03	0.02	97	1.12	0.16	116	1.30	0.55	206	2.31	1.00
		36	605	1.02	0.02	609	1.03	0.07	943	1.59	1.00	1690	2.85	1.00
		60	1732	1.01	0.02	1730	1.01	0.04	3016	1.76	1.00	5553	3.25	1.00
		90	3943	1.01	0.00	3942	1.01	0.00	7608	1.94	1.00	14210	3.63	1.00
		120	7042	1.00	0.00	7050	1.00	0.00	14632	2.08	1.00	27377	3.90	1.00
	20	15	91	1.04	0.05	91	1.05	0.08	152	1.71	0.98	310	3.48	1.00
		36	602	1.02	0.04	602	1.02	0.07	1414	2.38	1.00	2957	4.99	1.00
		60	1722	1.01	0.02	1723	1.01	0.05	4878	2.85	1.00	10206	5.97	1.00
		90	3930	1.00	0.00	3928	1.00	0.00	13032	3.33	1.00	26651	6.81	1.00
		120	7037	1.00	0.00	7038	1.00	0.00	25950	3.70	1.00	51807	7.38	1.00
Nonnormal	10	15	89	1.02	0.01	99	1.14	0.19	105	1.18	0.25	186	2.09	1.00
		36	602	1.02	0.00	614	1.04	0.09	731	1.23	0.99	1371	2.31	1.00
		60	1726	1.01	0.00	1737	1.02	0.05	2198	1.29	1.00	4364	2.55	1.00
		90	3938	1.01	0.00	3948	1.01	0.00	5256	1.34	1.00	10972	2.80	1.00
		120	7048	1.00	0.00	7060	1.01	0.00	9770	1.39	1.00	20893	2.98	1.00
	20	15	90	1.04	0.01	93	1.07	0.09	123	1.38	0.73	265	2.97	1.00
		36	601	1.02	0.01	604	1.02	0.05	928	1.56	1.00	2260	3.81	1.00
		60	1720	1.01	0.01	1726	1.01	0.05	2948	1.73	1.00	7649	4.48	1.00
		90	3928	1.00	0.00	3933	1.01	0.00	7386	1.89	1.00	19792	5.06	1.00
		120	7038	1.00	0.00	7044	1.00	0.00	14187	2.02	1.00	38334	5.46	1.00

Note: Level = distributional characteristics of item level data; Ratio = $N:p$ ratio; p = number of indicators; C = number of categories; rej = average number of rejections per cell.

yielded normed χ^2 values, which approximated 1 across all conditions, illustrating that the size of the model did not affect the global χ^2 test of goodness of fit. Empirical rejection rates were at or lower than 0.05 when two-category data were analyzed under all conditions. When five-category data were analyzed, increasing the sample size helped to reduce rejection rates to the nominal level. For all conditions, as the number of items increased, the χ^2 rejection rate decreased. In addition,

Type I error rates fell below the nominal level (0.05) when model size increased and dichotomous data were analyzed.

When misspecified models were analyzed, results yielded the opposite conclusion. normed χ^2 indices increased as the number of items increased. In addition, χ^2 rejection rates were high for all conditions, and at 100% for the majority of situations. When dichotomous data were analyzed, with fewer items and low $N:p$ ratios, the normed χ^2 values were lower than 2, indicating acceptable fit. Interestingly, when all variables followed asymmetric distributions, normed χ^2 values were within acceptable regions with dichotomous data; none of the average normed χ^2 values were at an acceptable level when five-category data were analyzed.

Average Fit Indices

Table 4 provides average fit indices of RMSEA and WRMR across conditions. Considering correctly specified conditions, all values indicated an adequately fitting model. As expected, RMSEA decreased with increasing model size (Bandalos, 2014) and WRMR increased with increasing model size (DiStefano et al., 2018).

Under misspecification, RMSEA values did not always indicate a poorly fitting model. In fact, RMSEA indicated adequate ($\overline{\text{RMSEA}} \leq .08$) fit in 28 of the 36 misspecified conditions and good fit ($\overline{\text{RMSEA}} \leq .06$) in 17 of the 36 misspecified conditions, with better fit estimates observed when the size of the model was very large ($p = 60$). According to its definition, the RMSEA penalizes model complexity by incorporating the degree of freedom in the formulation, and measures the discrepancy due to approximation per degree of freedom. Therefore, for the fixed misspecified parameter(s), the population RMSEA is expected to decrease as p increases, because including more variables is typically associated with larger degrees of freedom.

The WRMR appeared to be a better indicator of a misspecified model than the RMSEA. Using a cutoff at or less than 1 as indicative of an adequately fitting model (DiStefano et al., 2018), condition averages here indicated poor fit. WRMR increased with increasing numbers of observed indicators in the model, suggesting greater misspecification with increasing observed items.

Discussion

As few studies in the literature have examined large models with ordinal data, this study investigated the effect that increasing model size, as defined by many ordinal items, had on selected outcomes: convergence rates, relative bias of parameter estimates, standard errors of parameter estimates, performance of the model χ^2 and selected ad hoc model fit indices. The unique feature of the study was to investigate performance of the under WLSMV estimator under larger models than previously examined in the literature. Three-factor CFA population models, with increasing numbers of items ($p = 15, 36, 60, 120$) were examined and characteristics of sample size, number of ordered categories, and item-level distribution were manipulated.

Table 4. Average Values, Selected Ad Hoc Fit Indices.

Level	Ratio	<i>p</i>	Correctly Specified				Misspecified			
			C = 2		C = 5		C = 2		C = 5	
			RMSEA	WRMR	RMSEA	WRMR	RMSEA	WRMR	RMSEA	WRMR
Normal	10	15	0.014	0.733	0.017	0.606	0.072	1.092	0.107	1.183
		36	0.006	0.816	0.006	0.719	0.058	1.523	0.082	1.877
		60	0.003	0.837	0.004	0.755	0.051	1.869	0.070	2.428
		90	0.002	0.845	0.002	0.773	0.046	2.228	0.061	2.985
		120	0.001	0.85	0.002	0.782	0.042	2.528	0.055	3.447
	20	15	0.009	0.726	0.010	0.587	0.074	1.336	0.108	1.512
		36	0.004	0.813	0.004	0.708	0.062	1.978	0.085	2.528
		60	0.002	0.834	0.003	0.750	0.055	2.496	0.073	3.327
		90	0.001	0.844	0.001	0.768	0.049	3.026	0.064	4.13
		120	0.001	0.848	0.001	0.779	0.045	3.469	0.057	4.797
Half normal	10	15	0.013	0.785	0.024	0.655	0.042	0.965	0.092	1.089
		36	0.007	0.868	0.008	0.750	0.040	1.287	0.071	1.679
		60	0.004	0.887	0.004	0.781	0.036	1.519	0.061	2.137
		90	0.003	0.892	0.003	0.797	0.032	1.764	0.054	2.607
		120	0.002	0.891	0.002	0.805	0.03	1.973	0.049	2.999
	20	15	0.011	0.788	0.012	0.620	0.048	1.104	0.090	1.323
		36	0.005	0.861	0.005	0.736	0.044	1.582	0.074	2.207
		60	0.003	0.878	0.003	0.773	0.039	1.943	0.064	2.892
		90	0.002	0.886	0.002	0.791	0.036	2.321	0.057	3.577
		120	0.001	0.889	0.001	0.801	0.034	2.637	0.052	4.147
Nonnormal	10	15	0.012	0.759	0.026	0.668	0.032	0.903	0.084	1.038
		36	0.006	0.861	0.009	0.762	0.025	1.107	0.060	1.512
		60	0.004	0.879	0.005	0.789	0.022	1.257	0.051	1.888
		90	0.002	0.885	0.003	0.802	0.019	1.414	0.045	2.278
		120	0.002	0.887	0.002	0.81	0.018	1.555	0.041	2.605
	20	15	0.010	0.773	0.013	0.629	0.034	0.992	0.080	1.228
		36	0.004	0.854	0.005	0.742	0.028	1.290	0.062	1.939
		60	0.002	0.869	0.003	0.777	0.025	1.531	0.054	2.513
		90	0.001	0.877	0.002	0.794	0.022	1.784	0.047	3.091
		120	0.001	0.881	0.001	0.804	0.021	2.004	0.043	3.573

Note. Level = distributional characteristics of item level data; Ratio = *N:p* ratio; *p* = number of indicators; C = number of categories; RMSEA = root mean square error of approximation; WRMR, weighted root mean square residual.

As the number of ordinal variables increased, convergence rates and accuracy of parameter estimates increased for both correctly specified and misspecified models. As with previous studies, increasing the number of items in a tested model was associated with a greater number of admissible solutions and also more accurate parameter estimates, (e.g., Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Marsh et al., 1998). Parameter estimates for correctly specified models had been found to be accurate from previous studies when using WLSMV (e.g., Flora & Curran, 2004).

Here, we noted that parameter estimates for loading values reported minimal levels of bias, even under severe misspecification. However, interfactor correlation estimates reported large levels of negative bias in parameter estimates and interfactor correlation estimates were closer to the population value as p increased. While prior research has identified underestimation in factor correlations (e.g., Bandalos, 2014; DiStefano, 2002; DiStefano & Morgan, 2014; Potthast, 1993), previous studies have not included as many items as the current study. Thus, including more items tends to improve the accuracy of factor correlation estimates.

Accuracy in parameter estimates was not impacted by the number of categories or sample size used; however, convergence rates were impacted, with fewer admissible solutions observed when ordinal data included more categories and smaller sample sizes were used. Such results have been observed previously, as WLSMV has encountered estimation problems when empty cells are present (DiStefano & Morgan, 2014; Savalei, 2012).

Previous studies including WLS estimators have noted extreme amounts of negative bias in parameter estimates—especially with factor intercorrelations, asymmetric distributions, and when fewer categories were analyzed (DiStefano, 2002; DiStefano & Morgan, 2014; Potthast, 1993). With the exception of dichotomous, symmetric data, standard errors were extremely underestimated when the number of observed items were small ($p = 15$) matching findings from previous research (e.g., DiStefano & Morgan, 2014). However, a new finding was that standard errors of parameter estimates became closer to the empirical standard deviation values (i.e., smaller relative bias in standard errors) as the number of items increased. This finding was pronounced with the relative bias associated with factor correlations, as bias decreased substantially with more items. Also, the reduction in the relative bias was observed for both correctly specified and misspecified models. As studies typically concentrate neither on standard errors of parameter estimates nor on increasing model size, the findings are noteworthy.

In terms of fit indices, the 5% empirical rejection rate for the overall model χ^2 decreased and approached zero as model size increased. This means that even with a large number of ordinal variables, researchers would tend to never reject a correctly specified model. Such findings are consistent with previous literature (e.g., Rhemtulla et al., 2012; Savalei & Rhemtulla, 2013; Shi, DiStefano, et al., 2018). Shi, DiStefano, et al. (2018) noted similar results with other estimators when robust corrections for both mean and variance were present (e.g., ULSMV, WLSMV). In other words, the mean- and variance-corrections may “overcorrect” the estimated χ^2 value to produce values that are excessively positive of the model–data fit. WLSMV-based χ^2 values were able to indicate poor fit (Bandalos, 2014) under misspecification, where values generally indicated poor model–data fit, regardless of the number of variables, categories, or sample size. It was only in the least ideal combination of these conditions that power fell conspicuously below a desired cutoff of 0.80. That is when two-category, asymmetric data was analyzed with few variables.

An additional strength of the current study is the investigation of ad hoc fit indices beyond the χ^2 value. As the number of variables increased, sample RMSEA values

decreased, indicating better fit as p increased. This pattern was in opposition to findings from CFA with continuous outcomes under ML estimation (Anderson & Gerbing, 1984; Ding, Velicer, & Harlow, 1995; Kenny & McCoach, 2003; Shi, Lee, & Maydeu-Olivares, 2018). However, we expected WRMR to increase, even under correctly specified models, as this index is sensitive to sample size (DiStefano et al., 2018). Few studies have investigated WRMR at all, thus, these findings contribute to further understanding of this index.

Under misspecified models, most additional fit indices examined in the current study (e.g., normed χ^2 , RMSEA, and WRMR) indicated worse model–data fit as the number of items increased, the normed χ^2 increased, indicating worse fit. Unlike other indices, the sample RMSEA decreased, indicating better fit under misspecification when more ordinal items were included. The findings are partially consistent with results from CFA with continuous outcomes (Maydeu-Olivares, Shi, & Rosseel, 2018; Shi, Lee, & Maydeu-Olivares, 2018). Comparing the behaviors of the fit indices, as p increased, WRMR noted poorer fit whereas RMSEA illustrated better fit as the number of items increased.

Based on the study, we are able to make recommendations to applied researchers interested in fitting an ordinal CFA model with many variables and using WLSMV to accommodate for the ordinal nature of the data. First, including more observed variables in the measurement models could yield better convergence rates, especially when ordinal data are asymmetrically distributed. In addition, using more items in the measurement model is associated with more accurate parameter estimates and smaller relative bias in the standard error of the parameter estimates. Even when the models were misspecified, including more observed variables in the measurement models yielded parameter estimates closer to the true value, and standard errors closer to the empirical standard deviations. We understand that these findings are limited to the misspecified models studied here; however, the trend showed a strong decrease in the standard errors as the number of variables increased. This finding relates to increased accuracy associated with parameter estimates and greater trustworthiness in significance tests of parameters.

Finally, when evaluating goodness of fit for ordinal CFA, researchers should be cautious in interpreting RMSEA, especially when p is large, the data are asymmetric, and few categories are used. Under such situations, severely misspecified models can yield an average RMSEA of less than 0.05. For conditions considered in the current study, WRMR seemed to perform acceptably with the commonly used cutoff of 1. However, as the performance of fit indices has been shown to differ when ordinal data are used and, thus, not follow the recommendations commonly used with continuous data (e.g., Nye & Drasgow, 2011), future studies are needed to gain a greater understanding of fit when ordinal data are present.

Limitations and Future Directions

We recognize limitations with the current study. First, we note that the findings are limited to the situations manipulated here. We recognize that an infinite number of

conditions could be manipulated within a simulation study and other situations not covered in this study may be of interest to applied researchers. The conditions here represent a sampling of select situations and different combinations of conditions may be extended in future studies.

Second, this study considered only the WLSMV estimator and the Mplus (v. 7.4) software package. While these conditions were selected to analyze data as recommended using default procedures for ordinal data, future studies should investigate the effect of model size under different estimation methods (e.g., ULSMV with dichotomous data; MLMV with 5-category data). Also, we recognize that the model misspecification used here may be considered extreme in many situations. Thus, different types and levels of model misspecification (e.g., omitting cross-loadings) may be of interest to examine. Finally, we note that the model size could affect the estimation of commonly used robust ad hoc fit indices in terms of their population values, or/and their sample estimates. Future studies should explore the effects of estimates obtained from a tested model to their population counterparts.

Given the strengths of robust estimation, the ability of SEM packages to accommodate large numbers of variables efficiently, and the availability of large amounts of test data, it is likely that researchers may be faced with the need to estimate large models. The WLSMV technique is largely able to accommodate such situations. We look forward to future research in this line of research to best assist empirical research involving many ordinal items.

Authors' Note

Heather L. McDaniel is now affiliated with University of Virginia, Charlottesville, VA, USA.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Zhehan Jiang <https://orcid.org/0000-0002-1376-9439>

References

- Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49, 155-173. doi:10.1007/BF02294170

- Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling, 15*, 211-240. doi:10.1080/10705510801922340
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling, 21*, 102-116. doi:10.1080/10705511.2014.859510
- Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625-666). Greenwich, CT: 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*, 186-203. doi:10.1207/s15328007sem1302_2
- Cinamon, R. G. (2016). Integrating work and study among young adults: Testing an empirical model. *Journal of Career Assessment, 24*, 527-542. doi:10.1177/1069072715599404
- 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*, 119-143. doi:10.1080/10705519509540000
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327-346. doi:10.1207/S15328007SEM0903_2
- DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling, 25*, 453-466. doi:10.1080/10705511.2017.1390394
- 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. doi:10.1080/10705511.2014.915373
- Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439-492). Charlotte, NC: Information Age.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491. doi:10.1037/1082-989X.9.4.466
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*, 275-299. doi:10.1037/a0015825
- 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. doi:10.1080/10705510903203573
- Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling, 14*, 361-390. doi:10.1080/10705510701301602
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*, 329-367. doi:10.1177/0049124198026003003
- 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. doi:10.1080/10705519909540118

- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10, 333-351. doi:10.1207/S15328007SEM1003_1
- Kranzler, A., Fehling, K. B., Anestis, M. D., & Selby, E. A. (2016). Emotional dysregulation, internalizing symptoms, and self-injurious and suicidal behavior: structural equation modeling analysis. *Death Studies*, 40, 358-366. doi:10.1080/07481187.2016.1145156
- Lau, Y., Tha, P. H., Wong, D. F. K., Wang, Y., Wang, Y., & Yobas, P. K. (2016). Different perceptions of stress, coping styles, and general well-being among pregnant Chinese women: A structural equation modeling approach. *Archives of Women's Mental Health*, 19, 71-78. doi:10.1007/s00737-015-0523-2
- Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity*, 43, 495-507. doi:10.1007/s11135-007-9133-z
- Lei, P. W., & Wu, X. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164-180). New York, NY: Guilford Press.
- Levant, R. F., Hall, R. J., Weigold, I. K., & McCurdy, E. R. (2016). Construct validity evidence for the Male Role Norms Inventory–Short Form: A structural equation modeling approach using the bifactor model. *Journal of Counseling Psychology*, 63, 534-542. doi:10.1037/cou0000171
- Marsh, H. W., Hau, K. T., 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. doi:10.1207/s15327906mbr3302_1
- 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.
- Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling*, 25, 389-402.
- McWhirter, E. H., Hackett, G., & Bandalos, D. L. (1998). A causal model of the educational plans and career expectations of Mexican American high school girls. *Journal of Counseling Psychology*, 45, 166-181. doi:10.1037/0022-0167.45.2.166
- Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, 19, 86-98. doi:10.1080/10705511.2012.634724
- 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. doi:10.1111/j.2044-8317.1992.tb00975.x
- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, 14, 548-570. doi:10.1177/1094428110368562
- Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, 46, 273-286. doi:10.1111/j.2044-8317.1993.tb01016.x

- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354-373. doi: 10.1037/a0029315
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). 10.18637/jss.v048.i02
- Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, 72, 910-932. doi:10.1177/0013164412452564
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21, 149-160. doi:10.1080/10705511.2013.824793
- Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, 66, 201-223. doi: 10.1111/j.2044-8317.2012.02049.x
- Shi, D., DiStefano, C., McDaniel, H., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling*, 25, 924-945. doi:10.1080/10705511.2018.1449653
- Shi, D., Lee, T., & Maydeu-Olivares, A. (2018). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, 79, 310-334.
- Shi, D., Lee, T., & Terry, R. A. (2015). Revisiting the model size effect in structural equation modeling (SEM). *Multivariate Behavioral Research*, 50, 142-142.
- Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling*, 25, 21-40.
- Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*. Advance online publication. doi:10.1177/1073191117711020.
- Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed., pp. 653-771). Needham Heights, MA: Allyn & Bacon.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79. doi:10.1037/1082-989X.12.1.58
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73, 913-934. doi: 10.1177/0013164413495237
- Yu, C. Y., & Muthén, B. (2002, April). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.