British Journal of Mathematical and Statistical Psychology (2012)

The British Psychological Society

www.wileyonlinelibrary.com

The performance of robust test statistics with categorical data

Victoria Savalei[1]* and Mijke Rhemtulla[2]

[1]University of British Columbia, Vancouver, Canada
[2]University of Kansas, USA

This paper reports on a simulation study that evaluated the performance of five structural equation model test statistics appropriate for categorical data. Both Type I error rate and power were investigated. Different model sizes, sample sizes, numbers of categories, and threshold distributions were considered. Statistics associated with both the diagonally weighted least squares (cat-DWLS) estimator and with the unweighted least squares (cat-ULS) estimator were studied. Recent research suggests that cat-ULS parameter estimates and robust standard errors slightly outperform cat-DWLS estimates and robust standard errors (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). The findings of the present research suggest that the mean- and variance-adjusted test statistic associated with the cat-ULS estimator performs best overall. A new version of this statistic now exists that does not require a degrees-of-freedom adjustment (Asparouhov & Muthén, 2010), and this statistic is recommended. Overall, the cat-ULS estimator is recommended over cat-DWLS, particularly in small to medium sample sizes.

1. Introduction

Structural equation modelling is a popular data modelling tool in many areas of the social and behavioural sciences. Among the most popular types of structural equation model are confirmatory factor analysis (CFA) models, which traditionally hypothesize a set of linear relationships between the observed indicator variables and the latent factors. However, when data are categorical, linear relationships between the observed categorical indicators and continuous latent factors are no longer possible. Instead, categorical CFA analysis assumes that there is a continuous latent variable underlying each observed categorical variable. The linear CFA model is then assumed to connect these underlying continuous indicators and the latent factors.

A popular class of approaches for fitting categorical CFA models are the so-called limited information methods (e.g., Maydeu-Olivares & Joe, 2005), which fit the model

*Correspondence should be addressed to Victoria Savalei, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T1Z4, Canada (e-mail: v.savalei@ubc.ca).

DOI:10.1111/j.2044-8317.2012.02049.x

only to the univariate and bivariate frequencies of the observed categorical data. Several such approaches exist. One method is first to estimate variables' thresholds and the matrix of polychoric correlations, and then to fit the CFA model to this matrix (Christoffersson, 1975; Jöreskog, 1994; Olsson, 1979; Muthén, 1978, 1984, 1993; Lee, Poon, & Bentler, 1990, 1995). This method is implemented, for example, in M*plus* 6.11 (Muthén & Muthén, 2010). The polychoric correlation matrix is computed under the assumption of multivariate normality of the underlying continuous indicators.

The three best-known limited information methods for categorical data are weighted least squares (cat-WLS), unweighted least squares (cat-ULS), and diagonally weighted least squares (cat-DWLS), which use different fit functions to fit the CFA model to the polychoric correlation matrix. All three of these approaches minimize a fit function that is a weighted sum of model residuals, that is, differences between polychoric correlations and model-estimated correlations. They differ in the weight matrix used. The oldest approach, cat-WLS, uses the inverse of the estimated covariance matrix of polychoric correlations as the weight matrix (e.g., Muthén, 1978, 1984). This method produces correct standard error estimates without any special corrections and an asymptotically chi-square distributed model test statistic (when the model is true). The method is not often used because it tends to be unstable and to produce biased results unless the sample size is very large (DiStefano, 2002; Dolan, 1994; Flora & Curran, 2004; Hoogland & Boomsma, 1998; Lei, 2009; Maydeu-Olivares, 2001; Potthast, 1993; Yang-Wallentin, Jöreskog, & Luo, 2010).

The two methods that perform best in small and medium samples are cat-ULS and cat-DWLS. Cat-ULS simply minimizes the sum of squared model residuals; that is, it uses the identity matrix as the weight matrix. Cat-DWLS uses a diagonal weight matrix, where the diagonal elements prior to inverting are obtained from the estimated covariance matrix of polychoric correlations. Recent evidence suggests that cat-ULS and cat-DWLS parameter estimates perform very similarly (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Yang-Wallentin *et al*., 2010), with cat-ULS performing slightly better. The default standard errors associated with cat-ULS and cat-DWLS are not correct and require corrections. So-called robust or sandwich standard errors can be computed for each method. The relative performance of these robust standard errors in terms of coverage is also very similar, with cat-ULS robust standard errors outperforming slightly (Forero *et al*., 2009). Because the finding that cat-ULS may be preferred over cat-DLWS is relatively new, cat-DWLS remains the most common method of analysis among practitioners.

This paper is concerned with model test statistics for categorical data. The default model test statistics associated with cat-ULS and cat-DWLS are also incorrect and require adjustments. Several robust test statistics can in principle be computed for each method; in practice, researchers' choices are limited by the options available in the popular software. In this paper, we used M*plus* 6.11, which offers the following options. A mean- and variance-adjusted chi-square is available for both cat-ULS and cat-DWLS estimators (activated, respectively, by ESTIMATOR: ULSMV and WLSMV), and a mean-corrected chi-square is available for cat-DWLS (activated by ESTIMATOR: WLSM), but not for cat-ULS. In addition, two slightly different computations of the mean- and variance-adjusted chi-square are available. Technical details for all these statistics are provided in Section 2.

While a few studies exist that compare the cat-ULS and cat-DWLS estimators and their associated robust standard errors, no study, to our knowledge, has comprehensively compared both mean- and mean- and variance-adjusted robust test statistics across these two categorical estimators. The goal of the present study is to compare all categorical

data test statistics available in M*plus* for cat-ULS and cat-DWLS estimators, in terms of both Type I error and power. Of interest are both the comparison of different test statistics within an estimator, and the comparison of the same type of statistic across estimators. The latter comparison may present a reason to prefer one estimation method over the other.

2. Robust test statistics for cat-ULS and cat-DWLS

Let y be a $p \times 1$ vector of categorical variables with k categories, and let y^* be the $p \times 1$ vector of the underlying continuous normally distributed variables with mean 0 and variance 1. Let $\tau_1, \ldots, \tau_{k-1}$ be the thresholds used to categorize y^* into y. Let ρ be the $\frac{1}{2}p(p-1) \times 1$ vector of population correlations among the variables y^*. Categorical CFA models assume that this vector is structured according to the model $\rho = \rho(\theta)$, where θ is the vector of q parameters that includes loadings and factor correlations.

Let r be the $\frac{1}{2}p(p-1) \times 1$ vector of polychoric correlations estimated from the observed categorical data. Assuming a saturated threshold structure, the cat-ULS parameter estimates $\hat{\theta}_{\text{LS}}$ are obtained by minimizing the fit function $F_{\text{ULS}} = (r - \rho(\theta))'(r - \rho(\theta))$. Cat-DWLS parameter estimates $\hat{\theta}_{\text{DWLS}}$ are obtained by minimizing the fit function $F_{\text{DWLS}} = (r - \rho(\theta))'\hat{D}^{-1}(r - \rho(\theta))$, where $\hat{D} = \text{diag}(\hat{V})$ is a diagonal matrix, and $\hat{V}$ is an estimate of the asymptotic covariance matrix of r, the vector of polychoric correlations. The default or 'naïve' test statistics are given by $T_{\text{ULS}} = (N-1)F_{\text{ULS}}(\hat{\theta}_{\text{ULS}})$ and $T_{\text{DWLS}} = (N-1)F_{\text{DWLS}}(\hat{\theta}_{\text{DWLS}})$ for cat-ULS and cat-DWLS, respectively. These statistics are not valid for inference, as neither is asymptotically chi-square distributed when the model is true. Some programs, such as M*plus*, no longer even print their values. Robust corrections to these statistics have been developed that adjust the test statistics to approximately follow a chi-square distribution.

The following five robust statistics are studied in this paper: $T_{\text{DWLS-M}}$ (the mean-adjusted statistic based on the cat-DWLS estimator). $T_{\text{DWLS-MV1}}$ and $T_{\text{DWLS-MV2}}$ (the original and new versions of the mean- and variance-adjusted statistics based on the cat-DWLS estimator), and $T_{\text{ULS-MV1}}$ and $T_{\text{ULS-MV2}}$ (the original and new versions of the mean- and variance-adjusted statistics based on the cat-ULS estimator). These are now defined.

The mean-adjusted statistic based on the cat-DWLS estimator is given by:

$$T_{\text{DWLS-M}} = \frac{df}{\text{tr}(\hat{U}_{\text{DWLS}}\hat{V})} T_{\text{DWLS}}, \tag{1}$$

where $df = \frac{1}{2}p(p-1) - q$, $\hat{U}_{\text{DWLS}} = \hat{D}^{-1} - \hat{D}^{-1}\hat{\Delta}_{\text{DWLS}}(\hat{\Delta}'_{\text{DWLS}}\hat{D}^{-1}\hat{\Delta}_{\text{DWLS}})^{-1}\hat{\Delta}'_{\text{DWLS}}\hat{D}^{-1}$, and

$$\hat{\Delta}_{\text{DWLS}} = \left.\frac{\partial \rho(\theta)}{\partial \theta'}\right|_{\hat{\theta}_{\text{DWLS}}}$$

is the $\frac{1}{2}(p-1)p \times q$ matrix of model derivatives (Satorra & Bentler, 1994; Muthén, 1993). This statistic is analogous to the so-called Satorra–Bentler scaled chi-square that is popular for continuous data. It is referred to a chi-square distribution with df degrees of freedom, χ^2_{df}, although this is only its approximate asymptotic distribution. The distribution of $T_{\text{DWLS-M}}$ matches χ^2_{df} in the mean; for this reason equation (1) is known as a first-order

adjustment. In principle the corresponding statistic for the ULS estimator, $T_{\mathrm{ULS-M}}$, could also be defined, but this statistic is not printed by M*plus*, thus precluding its study. Yang-Wallentin *et al*. (2010) compared the LISREL implementations of $T_{\mathrm{DWLS-M}}$ and $T_{\mathrm{ULS-M}}$ in samples of size 400 and greater, and found their rejection rates to be nearly identical.

With categorical data, the mean- and variance-adjusted statistics appear to perform better than mean-adjusted statistics in small samples (Maydeu-Olivares, 2001; Muthén, du Toit, & Spisic, 1997). The original mean- and variance-adjusted statistic based on the categorical DWLS estimator is defined as follows:

$$T_{\mathrm{DWLS-MV1}} = \frac{k_{\mathrm{DWLS}}}{tr(\hat{U}_{\mathrm{DWLS}}\hat{V})} T_{\mathrm{DWLS}}, \tag{2}$$

which is referred to a chi-square distribution with the new adjusted degrees of freedom k_{DWLS}, where

$$k_{\mathrm{DWLS}} \approx \frac{[\mathrm{tr}(\hat{U}_{\mathrm{DWLS}}\hat{V})]^2}{\mathrm{tr}(\hat{U}_{\mathrm{DWLS}}\hat{V}\hat{U}_{\mathrm{DWLS}}\hat{V})},$$

rounded to the nearest integer. The distribution of $T_{\mathrm{DWLS-MV1}}$ matches $\chi^2_{k_{\mathrm{DWLS}}}$ in the mean and the variance, and equation (2) provides a second-order adjustment. Equations (1) and (2) differ only in that the degrees of freedom in the numerator of (2) are redefined. The original mean- and variance-adjusted statistic based on the categorical ULS estimator is similar and is defined as follows:

$$T_{\mathrm{ULS-MV1}} = \frac{k_{\mathrm{ULS}}}{\mathrm{tr}(\hat{U}_{\mathrm{ULS}}\hat{V})} T_{\mathrm{ULS}}, \tag{3}$$

where $\hat{U}_{\mathrm{ULS}} = I - \hat{\Delta}_{\mathrm{ULS}}(\hat{\Delta}'_{\mathrm{ULS}}\hat{\Delta}_{\mathrm{ULS}})^{-1}\hat{\Delta}'_{\mathrm{ULS}}$,

$$\hat{\Delta}_{\mathrm{ULS}} = \left.\frac{\partial \rho(\theta)}{\partial \theta'}\right|_{\hat{\theta}_{\mathrm{ULS}}}, \quad \text{and} \quad k_{\mathrm{ULS}} \approx \frac{[\mathrm{tr}(\hat{U}_{\mathrm{ULS}}\hat{V})]^2}{\mathrm{tr}(\hat{U}_{\mathrm{ULS}}\hat{V}\hat{U}_{\mathrm{ULS}}\hat{V})},$$

rounded to the nearest integer. This statistic is referred to a chi-square distribution with degrees of freedom k_{ULS}.

The adjustment of the degrees of freedom in the statistics $T_{\mathrm{DWLS-MV1}}$ and $T_{\mathrm{ULS-MV1}}$ may be viewed as problematic. Researchers are used to thinking of degrees of freedom as the difference between the number of data points in the covariance or correlation matrix and the number of model parameters. Using these statistics may mean that the same model is referred to different degrees of freedom when estimated on different data sets. It may also mean that the test statistic has different degrees of freedom depending on the estimation method used – that is, k_{ULS} and k_{DWLS} may be different when computed on the same data set. Recently, Asparouhov and Muthén (2010) proposed a different way to implement a second-order adjustment, one that does not change the model's degrees of freedom. Under this approach, the new mean- and variance-adjusted statistic based on the cat-DWLS estimator is computed as follows:

$$T_{\mathrm{DWLS-MV2}} = a_{\mathrm{DWLS}} T_{\mathrm{DWLS}} - b_{\mathrm{DWLS}}, \tag{4}$$

where

$$a_{\text{DWLS}} = \sqrt{\frac{df}{\text{tr}(\hat{U}_{\text{DWLS}}\hat{V}\hat{U}_{\text{DWLS}}\hat{V})}}$$

and $b_{\text{DWLS}} = df - a_{\text{DWLS}}\text{tr}(\hat{U}_{\text{DWLS}}\hat{V})$. Similarly, the new mean- and variance-adjusted statistic based on the cat-ULS estimator is computed as follows:

$$T_{\text{ULS-MV2}} = a_{\text{ULS}}T_{\text{ULS}} - b_{\text{ULS}}, \tag{5}$$

where

$$a_{\text{LS}} = \sqrt{\frac{df}{\text{tr}(\hat{U}_{\text{ULS}}\hat{V}\hat{U}_{\text{ULS}}\hat{V})}}$$

and $b_{\text{ULS}} = df - a_{\text{ULS}}tr(\hat{U}_{\text{ULS}}\hat{V})$. The distribution of both statistics can be approximated by a χ^2_{df} distribution in both the mean and the variance. In a small simulation study, Asparouhov and Muthén (2010) found that Type I error rates for the cat-DWLS statistics (2) and (4) were extremely similar, with the new statistic $T_{\text{DWLS-MV2}}$ having slightly higher (typically less than 1%) rejection rates than the old statistic $T_{\text{DWLS-MV1}}$. The relative performance of the cat-ULS statistics (3) and (5) has not, to our knowledge, ever been evaluated.

3. Literature review

Several studies have evaluated the performance of cat-DWLS and/or cat-ULS with ordinal data, typically with either two or five categories. Both methods typically produce unbiased parameter estimates (Beauducel & Herzberg, 2006; Dolan, 1994; Flora & Curran, 2004; Forero *et al*., 2009; Lei, 2009; Muthén *et al*., 1997; Nussbeck, Eid, & Lischetzke, 2006; Rigdon & Ferguson, 1991; Yang-Wallentin *et al*., 2010). Very little bias has also been found in robust standard errors associated with either cat-DWLS or cat-ULS (Flora & Curran, 2004; Forero *et al*., 2009; Lei, 2009; Maydeu-Olivares, 2001; Nussbeck *et al*., 2006; Yang-Wallentin *et al*., 2010). Studies that have compared the two methods to each other have either reported no difference (Yang-Wallentin *et al*., 2010) or a slight advantage of cat-ULS over cat-DWLS (Forero *et al*., 2009; Maydeu-Olivares, 2001), in terms of both parameter estimates and robust standard errors.

When it comes to robust test statistics, which are the focus of the present paper, the literature is sparse. Yang-Wallentin *et al*. (2010) compared the performance of mean-adjusted cat-ULS and cat-DWLS statistics and found their Type I error rates to be both acceptable (near 5%) and similar to each other. However, only data for sample sizes greater than 400 were reported. Maydeu-Olivares (2001) compared the performance of the mean-adjusted and the mean- and variance-adjusted statistics associated with both cat-ULS and cat-DWLS methods in a small simulation study using very small models (either four or seven observed variables), data that had either 2 or 5 categories, and sample sizes of $N = 100$ or $N = 300$. He found that the mean- and variance-adjusted statistic outperformed the mean-adjusted statistic at $N = 100$ for both methods, and the performance of the two types of statistics was similar at $N = 300$. Cat-ULS and cat-DWLS

versions of the statistics performed very similarly. Several studies have found that the mean- and variance-adjusted statistic based on the cat-DWLS estimator performs well with 2- and 5-category data in samples of $N = 200$ or greater (Flora & Curran, 2004; Lei, 2009; Nussbeck *et al*., 2006; Muthén *et al*., 1997).

In summary, cat-ULS and cat-DWLS parameter estimates and standard errors have been found to perform similarly, with cat-ULS performing slightly better. Small differences make it difficult to recommend one method over the other. Cat-DWLS is the most popular choice among applied researchers. However, because cat-ULS does appear to have a slight advantage, some authors have advocated its use (Forero *et al*., 2009). This recommendation is incomplete without a thorough investigation of the relative performance of the corresponding robust test statistics, which has not been conducted. The current study aims to fill this gap in the literature and to provide such a comparison.

4. Method

A Monte Carlo simulation study was conducted to compare the performance of the five cat-ULS and cat-DWLS test statistics with categorical data. Normally distributed data were generated from a two-factor CFA model with either 5 or 10 indicators per factor. Factor loadings for each factor were .3, .4, .5, .6, and .7; when the factor had 10 indicators, these loadings repeated. These values have been used in previous simulation studies (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Flora & Curran, 2004). The factor correlation was set to .3. The variances of all observed and latent variables were set to 1. The data were then categorized to create ordinal variables. The following four variables were varied: model size ($p = 10$ or $p = 20$); number of categories (2–7); threshold type (symmetry; moderate asymmetry I, moderate asymmetry II, extreme asymmetry I, extreme asymmetry II, defined in Section 4.3); and sample size ($N = 100, 150, 350, 600$). The study had a total of 240 conditions, with 1,000 data sets generated per condition.[1] The four manipulated variables are now discussed in more detail.

4.1. Model size

Model 1 was a two-factor CFA model with 5 indicators per factor, for a total of 10 indicators. Model 2 was identical to model 1, but with 10 indicators per factor, for a total of 20 indicators. Model 1 had 34 degrees of freedom, while model 2 had 169 degrees of freedom. Note that for model 2, the degrees of freedom are greater than the two smallest studied sample sizes, and the behaviour of the test statistics may be particularly interesting in these conditions (e.g., Yuan & Bentler, 1998; Savalei, 2010).

4.2. Number of categories

Previous research that has compared cat-ULS and cat-DWLS statistics studied data with 2 and 5 categories (Maydeu-Olivares, 2001), or with 2, 5, and 7 categories (Yang-Wallentin *et al*., 2010). To better understand the effect of the number of categories on rejection

[1]The simulated data used in this study were a subset of the data generated by Rhemtulla, Brosseau-Liard, and Savalei (2012), who studied the relative performance of continuous and categorical data methods, but only examined one categorical estimator (cat-ULS) and one test statistic ($T_{\mathrm{ULS-MV1}}$).

rates of the test statistics, continuous latent response distributions were categorized into 2, 3, 4, 5, 6, or 7 categories.

4.3. Threshold type

Previous research has found that thresholds that were distributed asymmetrically around 0 led to less accurate cat-DWLS parameter estimates (Babakus, Ferguson, & Jöreskog, 1987; DiStefano, 2002; Dolan, 1994; Lei, 2009; Rigdon & Ferguson, 1991), and that highly asymmetric thresholds (e.g., 2-category data where more than 90% of the distribution fell into one category) resulted in biased robust standard errors for cat-DWLS, and to a lesser extent cat-ULS (Forero *et al*., 2009). When it comes to the effect of threshold asymmetry on test statistics, Lei (2009) found that threshold asymmetry led to higher Type I error rates for cat-DWLS mean-adjusted and mean- and variance-adjusted statistics. However, Yang-Wallentin *et al*. (2010), who only created mild threshold asymmetry, found that it made no difference for the rejection rates of mean- and mean- and variance-adjusted cat-ULS and cat-DWLS statistics. Thus, it may be that test statistics are robust to mildly asymmetric thresholds but not to extremely asymmetric ones. To investigate this, we created five threshold type conditions.

Table 1 summarizes the threshold values used. In the symmetry (S) condition, category thresholds were distributed symmetrically around 0. In the moderate asymmetry I (MA-I) condition, category thresholds were chosen such that the peak of the distribution fell to the left of centre. In the extreme asymmetry I (EA-I) condition, category thresholds were typically more skewed than in the MA-I condition and were also such that the lowest category would always contain the largest number of cases. As Table 1 illustrates, with 3 or more categories this means that the smallest category in the MA-I condition is smaller than the smallest category in the EA-I condition, and thus it is not as clear which threshold condition is more 'difficult'. In the S, MA-I, and EA-I conditions, all variables had the same threshold values. The remaining two conditions, moderate asymmetry II (MA-II) and extreme asymmetry II (EA-II), had identical threshold values to MA-I and EA-I, except that the direction of the asymmetry was reversed for half the variables. This situation is expected to make estimation of positive correlations particularly difficult.

4.4. Sample size

Four sample sizes were studied: $N = 100, 150, 350$, and 600. In structural equation modelling applications, sample sizes less than 200 are typically considered small. Thus, two small and two medium sample sizes are studied.

4.5. Data generation and analysis

Continuous normally distributed data were generated and automatically categorized using the simulation feature of EQS 6.1 (Bentler, 2008). Note that new data were generated for each of the 240 conditions – that is, the same continuous data were not categorized in more than one way.

Data in all 240 cells of the design were analysed ten times using M*plus* 6.11. The ten analyses differed in the following ways: the type of test statistic requested (five statistics, given by equations (1)–(5)); and whether the correct or an incorrect model was fitted to data. These are now discussed in more detail.

Table 1. Thresholds imposed on continuous data; proportion of the data falling into each category. In the MA-II and EA-II conditions, thresholds had opposite values for half the variables (these are not presented)

Threshold condition	Number of categories	Category thresholds as Z-scores						Proportion of values falling in each category						
S	2	0.00						50	50					
	3	−0.83	0.83					20	59	20				
	4	−1.25	0.00	1.25				11	39	39	11			
	5	−1.50	−0.50	0.50	1.50			7	24	38	24	7		
	6	−1.60	−0.83	0.00	0.83	1.60		5	15	30	30	15	6	
	7	−1.79	−1.07	−0.36	0.36	1.07	1.79	4	11	22	28	22	11	4
MA-I	2	0.36						64	36					
	3	−0.50	0.76					31	47	22				
	4	−0.31	0.79	1.66				38	41	17	5			
	5	−0.70	0.39	1.16	2.05			24	41	22	10	2		
	6	−1.05	0.08	0.81	1.44	2.33		15	38	26	14	7	1	
	7	−1.43	−0.43	0.38	0.94	1.44	2.54	8	26	31	18	10	7	1
MA-II	2	1.04						85	15					
	3	0.58	1.13					72	15	13				
	4	0.28	0.71	1.23				61	15	13	11			
	5	0.05	0.44	0.84	1.34			52	15	13	11	9		
	6	−0.13	0.25	0.61	0.99	1.48		45	15	13	11	9	7	
	7	−0.25	0.13	0.47	0.81	1.18	1.64	40	15	13	11	9	7	5

In M*plus*, one cannot obtain more than one test statistic associated with a particular estimator in one run, and thus analyses had to be done separately for each test statistic studied. For the cat-DWLS estimator, the analysis was done three times for each type of fitted model. The first cat-DWLS analysis set ESTIMATOR = WLSM, to obtain the mean-adjusted statistic $T_{\mathrm{DWLS-M}}$ given by equation (1). The second cat-DWLS analysis set ESTIMATOR = WLSMV, SATTERTHWAITE = ON, to obtain the original mean- and variance-adjusted statistic $T_{\mathrm{DWLS-MV1}}$ given by equation (2). The third cat-DWLS analysis set ESTIMATOR = WLSMV (omitting the second command activates the default, which is equivalent to specifying SATTERTHWAITE = OFF), obtaining the new mean- and variance-adjusted statistic $T_{\mathrm{DWLS-MV2}}$ given by equation (4). Note that the terminology used by the M*plus* syntax is somewhat misleading in that the estimator in all three analyses actually remains the same (diagonally weighted least squares), but what changes is the printed test statistic. For the cat-ULS estimator, the analysis was done only twice for each type of fitted model, because the cat-ULS version of the mean-adjusted statistic that would be analogous to (1) is not available in M*plus*. The first cat-ULS analysis set ESTIMATOR = ULSMV, SATTERTHWAITE = ON, to obtain the original mean- and variance-adjusted statistic $T_{\mathrm{ULS-MV1}}$ given by equation (3). The second cat-ULS analysis set ESTIMATOR = ULSMV, obtaining the new mean- and variance-adjusted statistic $T_{\mathrm{ULS-MV2}}$ given by equation (5).

Two models were fitted to data. The first model was the correct model that generated the data: a two-factor CFA model with free loadings and factor correlation. Rejection rates of the five test statistics for this model provide information about Type I error rates. The second model was a one-factor model with freely estimated loadings. Because this is the wrong model for the data, rejection rates of the five test statistics for this model provide information about power.

5. Results

Findings are summarized with respect to three outcomes: non-convergence/improper solutions rates; Type I error rates; and power. These are discussed in turn.

5.1. Convergence failures and improper solutions

While the focus of this paper is on test statistics, and not on parameter estimates, rates of non-convergence and improper solutions remain relevant. When comparing rejection rates of test statistics, particularly across different estimators, results may depend on how convergence failures and improper solutions are treated during the comparison. Even within the same estimator, different test statistics may 'win' when the comparison is done including improper solutions compared to when excluding them. We first discuss the observed number of convergence failures and rates of improper solutions before addressing the issue of how they should be treated in the test statistics comparison.

Table 2 (left panel) shows the number of convergence failures for model 1. At $N = 600$, there are no convergence failures, and these columns are omitted. Note that convergence rates differ by the type of estimator only (cat-ULS vs. cat-DWLS), and within a particular estimator are not affected by the type of test statistic. Most convergence failures occur when the sample size is small and the data have few categories. Convergence rates for binary data are the worst. However, the number of convergence failures is negligible in the S, MA-I, and MA-II conditions. The highest observed rate of convergence failures

Table 2. Number of convergence failures and convergence failures plus outliers out of 1,000 replications in each cell of the design: model 1. At $N = 600$, no convergence failures occurred

		Convergence failures								Convergence Failures + Improper Solutions							
Threshold condition	Number of categories	$N = 100$		$N = 150$		$N = 100$		N = 150		N = 100		N = 150		N = 100		N = 150	
		DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS
S	2	10	11	0	0	0	0	0	0	148	143	76	74	1	0	0	0
	3	3	4	0	0	0	0	0	0	46	47	23	19	0	0	0	0
	4	0	1	0	0	0	0	0	0	19	19	1	2	0	0	0	0
	5	0	0	1	0	0	0	0	0	20	21	7	7	0	0	0	0
	6	0	0	0	0	0	0	0	0	13	13	1	1	0	0	0	0
	7	0	0	0	0	0	0	0	0	15	20	1	1	0	0	0	0
MA-I	2	13	16	1	1	0	0	0	0	173	171	74	71	5	5	0	0
	3	1	1	0	0	0	0	0	0	44	43	16	21	0	0	0	0
	4	0	2	0	0	0	0	0	0	38	37	9	7	0	0	0	0
	5	0	0	0	0	0	0	0	0	12	13	4	3	0	0	0	0
	6	0	1	0	0	0	0	0	0	20	21	2	3	0	0	0	0
	7	0	0	0	0	0	0	0	0	11	11	2	3	0	0	0	0
MA-II	2	9	6	2	2	0	0	0	0	192	185	77	82	6	6	2	2
	3	1	1	0	0	0	0	0	0	37	31	4	4	0	0	0	0
	4	0	0	0	0	0	0	0	0	39	39	12	12	0	0	0	0
	5	2	1	0	0	0	0	0	0	29	25	4	4	0	0	0	0
	6	0	0	0	0	0	0	0	0	18	18	1	1	0	0	0	0
	7	1	0	0	0	0	0	0	0	16	19	3	2	0	0	0	0
EA-I	2	80	116	48	69	2	1	0	0	463	457	330	316	53	43	9	6
	3	21	18	2	2	0	0	0	0	173	156	82	79	4	3	0	0
	4	7	6	1	0	0	0	0	0	88	76	20	21	0	1	0	0
	5	3	3	0	0	0	0	0	0	43	44	16	12	0	0	0	0
	6	1	1	0	1	0	0	0	0	32	28	4	7	0	0	0	0
	7	1	1	1	0	0	0	0	0	27	23	6	5	0	0	0	0
EA-II	2	78	79	31	66	1	1	0	0	357	281	235	285	84	114	23	27
	3	12	13	2	4	0	0	0	0	255	245	116	113	6	7	0	0
	4	5	4	0	0	0	0	0	0	96	97	41	41	2	1	0	0
	5	1	2	0	0	0	0	0	0	61	57	11	10	0	0	0	0
	6	0	1	0	0	0	0	0	0	36	33	6	7	0	0	0	0
	7	0	0	0	0	0	0	0	0	20	26	6	8	0	0	0	0

is 11.6%, corresponding to the cat-DWLS estimator. ULS almost always produces better convergence rates than DWLS. The highest convergence failure rate for ULS is 8%). Across all conditions, 94 more replications converged via ULS than DWLS. The ULS fit function is simpler and thus may be computationally more stable under difficult conditions.

Somewhat surprisingly, convergence rates in all conditions are much better for the larger model 2 (these data are not presented). It appears that a greater number of indicators per factor (10 rather than 5) increases the stability of estimation. The number of convergence failures is less than 5 out of 1,000 in all but three cells; in these three cells, all corresponding to the DWLS estimator, the number of failures is 7, 7, and 11. These values are too small to make any difference for the rejection rates.

The right panel of Table 2 shows the *total* number of convergence failures and improper solutions for model 1. That is, the numbers in the right panel include the convergence failures in the left panel plus any additional problematic cases. A replication was said to have an improper solution if at least one residual variance parameter took on a negative value (because the polychoric correlation matrix has 1s on the diagonal, this is equivalent to excluding cases where at least one factor loading was estimated to be greater than 1). Additionally, all replications were checked for outlying estimates of standard errors (SEs), namely SEs greater than 1. However, with the exception of a single replication in a single cell, all SE outliers occurred in replications that also contained improper solutions.

The pattern here is similar, in that the intersection of a small size and binary data creates the most troublesome conditions in terms of the number of problematic cases. The most difficult conditions correspond to the two extreme asymmetry threshold conditions, where almost half of all replications produce improper solutions or result in convergence failures in some cells. It is now the case that cat-DWLS leads to slightly lower combined rates of convergence failures and improper solutions than does cat-ULS. A total of 91 more cases are considered acceptable under cat-DWLS than under cat-ULS. This advantage is mostly due to improper solutions in the two extreme asymmetry conditions.

The number of improper solutions is much smaller for the larger model 2 (these data are not presented). The total number of convergence failures and improper solutions across S, MA-I, and MA-II threshold conditions was between 0 and 4 for data with 3–7 categories, and between 0 and 2 for the largest three sample sizes for data with any number of categories. The only conditions with a greater number of problematic cases were at the intersection of 2-category data and $N = 100$, where the greatest number of improper solutions was 24. In the EA-I and EA-II threshold conditions, the greatest number of problematic cases was 129. In general, the number of problematic cases for model 2 was at least three times smaller than the corresponding number for model 1.

One way to summarize the results of Table 2 is as follows: ULS is more likely to produce *any* output, while DWLS is more likely to produce "clean" output. These findings replicate those of Forero *et al*. (2009), who found that cat-DWLS produced more cases that converged without outliers, and of Yang-Wallentin *et al*. (2010), who found that ULS converged more frequently. However, the differences among the methods in the number of acceptable cases, defined either way, is never greater than 6% of all cases, and is typically much smaller. It is not clear that one method should be preferred over the other based on convergence rates and improper solutions alone.

In order to meaningfully compare Type I error rates for the five test statistics, a decision must be made about how to treat convergence failures and improper solutions in the computations of the Type I error rates. There is some disagreement among

methodologists as to the best strategy. From a statistical point of view, Type I error rates are only meaningful if they are computed across *all* replications in a cell, that is, out of 1,000 cases. Conditioning the choice of replications to be kept in the analysis in any way ruins the statistical rationale for expecting a 5% rejection rate at $\alpha = .05$. This is because exclusion criteria are typically correlated with the size of the test statistic itself. Some programs, including M*plus*, do not produce any output when a case fails to converge; it is thus impossible to use the inclusive strategy of evaluating rejection rates across all cases. Because researchers frequently interpret lack of convergence as indicative of poor model fit, another approach is to count non-converged cases as rejections of the model (Yuan & Hayashi, 2003). This strategy has the potential to produce strongly biased rejection rates in difficult conditions (e.g., small N, asymmetric threshold distributions), and it is not a very common strategy in practice. An intermediate strategy would be to simply exclude convergence failures from the analysis. We follow this strategy.[2]

The case of improper solutions is more complicated, and the decision has the potential to skew the results since many such cases were observed. Chen, Bollen, Paxton, Curran, & Kirby (2001) conducted a simulation study investigating the rate of improper solutions as a function of model misspecification and did not find a clear relationship, concluding that "researchers should not use negative error variance estimates as an indicator of model misspecification" (p. 501). Improper solutions are in fact to be expected in small samples and do not represent a statistical anomaly (Savalei & Kolenikov, 2008). Thus, unlike with convergence failures, replications with improper solutions probably should *not* be counted as cases where the model is rejected. In fact, because such cases typically produce full model output, one can simply include them in the study, which is the strategy employed here. We believe it would be statistically unwise to exclude them from the computation of rejection rates, because as much as 46% of all replications in some cells would have to be excluded. However, results were compared with and without the inclusion of improper solutions, and only minor differences were found (see also Chen *et al.*, 2001). The largest of these differences are noted in this text.

5.2. Type I error rates

Tables 3–8 present Type I error rates at $\alpha = .05$ for data with 2 to 7 categories, respectively. Data for both models are included in each table. Rejection rates are based on all converged cases. Rejection rates in these tables are highlighted if they are statistically greater than .05. The 95% confidence interval for rejection rates when the population value is .05 is from .0365 to .0635, based on 1,000 replications. Rejection rates in Tables 3–8 are additionally printed in bold if they fall outside the bounds specified by Bradley's liberal criterion, which are from .025 to .075 (Bradley, 1978). In the few difficult conditions when virtually all cells are highlighted and in bold, test statistics can be compared based on the absolute rejection rates – the extent of inflation still matters in this case, in that a rejection rate of 10% indicates better performance in difficult conditions than a rejection rate of 20%.

Across all numbers of categories (all tables), the original and the new versions of the mean- and variance-adjusted statistics perform very similarly for both estimation methods.

[2]Results treating convergence failures as rejections can easily be obtained by combining the presented results with the data from Table 2. For instance, if convergence failures were counted as rejections in the $N = 100$, EA-II, 2-category condition, the cat-ULS statistics would have Type I rates that are 8% higher, and cat-DWLS statistics would have Type I error rates that are 11.6% higher.

Table 3. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 2. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05 (for 1,000 replications, this interval is from .0365 to .0635). Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, *N*	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.090**	.047	.051	**.021**	**.024**	**.238**	.043	.048	**.012**	**.013**
	150	**.079**	.051	.054	.036	.037	**.131**	.034	.035	**.013**	**.013**
	350	.065	.044	.044	.037	.037	**.096**	.042	.042	.028	.030
	600	.072	.058	.059	.055	.055	**.077**	.046	.047	.037	.039
MA-I	100	**.105**	.063	.064	.026	.027	**.238**	.048	.054	**.016**	**.016**
	150	**.089**	.056	.058	.037	.040	**.175**	.047	.050	**.016**	**.020**
	350	.073	.057	.057	.046	.048	**.101**	.036	.037	**.024**	.025
	600	**.085**	.068	.068	.061	.063	**.095**	.051	.053	.040	.043
MA-II	100	**.099**	.057	.059	.031	.033	**.231**	.047	.059	**.005**	**.008**
	150	**.096**	.058	.062	.037	.040	**.181**	.052	.054	**.016**	**.018**
	350	.055	.041	.041	.035	.035	**.101**	.048	.049	.029	.034
	600	.060	.049	.049	.046	.046	**.087**	.062	.063	.048	.050
EA-I	100	**.390**	**.231**	**.244**	**.010**	**.013**	**.942**	**.709**	**.736**	**.003**	**.003**
	150	**.276**	**.207**	**.218**	.025	.027	**.768**	**.578**	**.587**	**.012**	**.016**
	350	**.075**	.051	.053	.042	.042	**.156**	.044	.045	.027	.033
	600	**.080**	.059	.060	.056	.058	**.106**	.050	.051	.044	.046
EA-II	100	**.457**	**.342**	**.355**	**.008**	**.010**	**.953**	**.835**	**.849**	**.001**	**.001**
	150	**.352**	**.284**	**.287**	.030	.031	**.922**	**.789**	**.796**	**.010**	**.012**
	350	**.108**	**.083**	**.084**	.055	.056	**.328**	**.218**	**.220**	.047	.049
	600	**.078**	.060	.061	.060	.062	**.161**	**.092**	**.092**	.063	.065

The new versions exhibit slightly higher rejection rates. The cat-ULS mean- and variance-adjusted statistics (equations (3) and (5)) are particularly similar, with the maximum difference never exceeding 1% for any pair of cells corresponding to model 1, and with the maximum difference never exceeding 1.5% for any pair of cells corresponding to model 2. In the vast majority of conditions, the differences are much smaller. The cat-DWLS statistics (equations (2) and (4)) are also very similar but the differences are slightly larger. For model 1, the difference between statistics (2) and (4) exceeds 1% only in two cells across all tables. For model 2, the difference between statistics (2) and (4) exceeds 1% in many cells corresponding to the smallest sample size ($N = 100$), but it remains less than 2.5%. The largest differences occur for data with 7 categories. Thus, the original versions of the mean- and variance-adjusted statistics perform uniformly better, but the difference is typically small. The difference between old and new mean- and variance-adjusted statistics is not emphasized in the remainder of this section, and only the behaviour of the original mean- and variance-adjusted statistics (2) and (3) will be discussed.

Table 3 presents the rejection rates for binary data. Test statistics generally do best with symmetric (S) thresholds, followed by moderate asymmetry (MA) conditions,

Table 4. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 3. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05. Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, N	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.103**	.059	.062	.025	.027	**.218**	.029	.031	**.005**	**.007**
	150	**.097**	.054	.054	.034	.035	**.169**	.029	.035	**.017**	**.017**
	350	**.080**	.058	.059	.044	.048	**.098**	.036	.037	.027	.030
	600	.061	.048	.049	.039	.040	**.098**	.044	.045	.032	.033
MA-I	100	**.102**	.057	.059	.038	.041	**.229**	.039	.042	**.010**	**.012**
	150	**.103**	.070	.074	.054	.054	**.168**	.033	.042	**.017**	**.017**
	350	.068	.047	.047	.039	.039	**.100**	.044	.045	.028	.033
	600	.069	.050	.051	.049	.050	**.105**	.054	.055	.040	.043
MA-II	100	**.112**	.056	.063	.032	.035	**.243**	.046	.051	**.021**	.025
	150	**.096**	.067	.069	.048	.052	**.145**	.033	.036	**.014**	**.015**
	350	.066	.046	.046	.039	.039	**.121**	.047	.050	.037	.037
	600	**.082**	.066	.066	.059	.059	**.084**	.047	.050	.039	.042
EA-I	100	**.150**	**.082**	**.086**	.052	.057	**.433**	**.101**	**.112**	.031	.033
	150	**.116**	.069	.071	.044	.045	**.291**	.068	.075	**.024**	.026
	350	**.090**	.067	.068	.050	.054	**.126**	.047	.048	.032	.034
	600	**.076**	.051	.052	.049	.050	**.107**	.052	.055	.042	.044
EA-II	100	**.178**	**.098**	**.106**	.054	.059	**.443**	**.138**	**.145**	.053	.062
	150	**.145**	**.098**	**.101**	.061	.065	**.271**	**.092**	**.095**	.043	.048
	350	.064	.046	.046	.034	.037	**.147**	.072	.074	.052	.054
	600	.065	.057	.058	.052	.052	**.105**	.052	.053	.045	.045

followed by extreme asymmetry (EA) conditions. The cat-DWLS mean-adjusted statistic $T_{\mathrm{DWLS-M}}$ (equation (1)) performs the worst, exhibiting inflated rejection across almost all conditions, particularly in small samples ($N = 100$ and 150) and in the EA conditions, where its rejection rates are abysmal, exceeding 20%. They are worse for model 2. These rejection rates become somewhat smaller (by .013 to .035) when improper solutions are excluded, but this improvement is not very helpful (these data are not presented). The mean- and variance-adjusted statistics $T_{\mathrm{DWLS-MV1}}$ and $T_{\mathrm{ULS-MV1}}$ (equations (2) and (3), respectively) perform well in S and both MA conditions, even in small samples. However, $T_{\mathrm{ULS-MV1}}$ tends to under-reject models somewhat in small samples, particularly for the larger model 2, and $T_{\mathrm{DWLS-MV1}}$ produces better rejection rates. In the EA conditions, however, the performance of $T_{\mathrm{DWLS-MV1}}$ becomes abysmal for small sample sizes ($N = 100$ and 150). These rejection rates are up to 2.3% smaller when improper solutions are excluded, but again, this decrease is inconsequential (these data are not presented). The performance of $T_{\mathrm{ULS-MV1}}$ remains quite good even in the EA conditions, but this statistic continues to under-reject in smaller sample sizes, particularly with model 2. Overall, because under-rejection is typically considered to be less of a

Table 5. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 4. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05. Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, *N*	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.156**	**.085**	**.089**	.051	.054	**.368**	**.081**	**.089**	**.019**	**.021**
	150	**.120**	.068	.069	.046	.049	**.225**	.072	**.082**	.031	.033
	350	**.095**	.059	.060	.045	.046	**.144**	.058	.061	.041	.043
	600	.060	.047	.048	.043	.044	**.109**	.045	.046	.039	.039
MA-I	100	**.155**	**.077**	**.081**	.050	.053	**.422**	**.112**	**.126**	.038	.042
	150	**.134**	**.084**	**.089**	.065	.069	**.281**	**.098**	**.105**	.041	.045
	350	**.094**	.068	.072	.060	.061	**.162**	.063	.068	.054	.056
	600	**.087**	.064	.064	.062	.063	**.111**	.048	.048	.036	.038
MA-II	100	**.173**	**.099**	**.106**	.061	.065	**.418**	**.127**	**.140**	.040	.046
	150	**.117**	.075	**.076**	.054	.056	**.261**	.073	**.083**	.036	.036
	350	**.078**	.057	.059	.043	.047	**.144**	.067	.066	.046	.051
	600	.068	.051	.051	.046	.047	**.121**	.074	.074	.062	.064
EA-I	100	**.156**	**.077**	**.084**	.032	.038	**.366**	**.083**	**.098**	**.022**	.033
	150	**.117**	**.076**	**.078**	.055	.056	**.248**	.074	**.080**	.035	.039
	350	**.080**	.057	.062	.045	.046	**.142**	.050	.051	.037	.040
	600	**.091**	.061	.061	.057	.059	**.087**	.041	.045	.036	.036
EA-II	100	**.175**	**.091**	**.097**	.050	.050	**.377**	**.106**	**.115**	.030	.036
	150	**.121**	.069	.071	.051	.052	**.242**	**.081**	**.087**	.037	.038
	350	**.092**	.066	.069	.053	.056	**.123**	.056	.058	.038	.040
	600	.064	.055	.056	.049	.052	**.103**	.050	.053	.039	.040

problem than over-rejection, it can be concluded that $T_{\text{ULS–MV1}}$ outperforms $T_{\text{DWLS–MV1}}$ with binary data, and $T_{\text{DWLS–M}}$ should not be used.

Table 4 presents the results for data with 3 categories. The patterns of results are generally similar to those for binary data. Test statistics again do best in S and MA conditions. The cat-DWLS mean-adjusted statistic $T_{\text{DWLS–M}}$ again does not do well, particularly in the two smaller sample sizes. This statistic will not be discussed for the rest of this section. The mean- and variance-adjusted statistics $T_{\text{DWLS–MV1}}$ and $T_{\text{ULS–MV1}}$ perform well in S and both MA conditions. In the EA conditions, $T_{\text{DWLS–MV1}}$ again exhibits inflated rejection rates in smaller sample sizes, but the extent of this over-rejection is not nearly as dramatic as it was with binary data. Interestingly, $T_{\text{ULS–MV1}}$ performs best in the EA conditions, but in the S and MA conditions tends to under-reject in the smaller sample sizes. It is difficult to recommend one mean- and variance-adjusted statistic over the other from these data. There are virtually no differences in the results when improper solutions are excluded; only in two cells do the results change by more than 1%, and this change does not affect the conclusions. Removing improper solutions has virtually no effect on data with more than 3 categories, and will not be discussed further.

Table 6. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 5. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05. Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, N	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.205**	**.116**	**.124**	.067	.074	**.444**	**.120**	**.132**	.049	.054
	150	**.139**	**.095**	**.095**	.070	.071	**.292**	**.088**	**.093**	.037	.038
	350	**.104**	**.079**	**.080**	.065	.067	**.147**	.052	.054	.039	.039
	600	**.085**	.062	.063	.056	.056	**.132**	.052	.054	.043	.044
MA-I	100	**.228**	**.135**	**.140**	**.080**	**.081**	**.525**	**.172**	**.187**	**.082**	**.089**
	150	**.158**	**.100**	**.104**	.072	.074	**.360**	**.140**	**.145**	.072	**.082**
	350	**.107**	.073	.074	.057	.061	**.162**	**.078**	**.081**	.054	.055
	600	**.095**	.074	**.077**	.069	.070	**.114**	.040	.043	.036	.037
MA-II	100	**.225**	**.127**	**.134**	.068	.073	**.500**	**.160**	**.176**	.058	.068
	150	**.158**	**.112**	**.114**	**.080**	**.087**	**.348**	**.126**	**.134**	.064	.070
	350	**.077**	.054	.055	.041	.044	**.156**	.071	.074	.051	.053
	600	**.082**	.060	.061	.054	.055	**.125**	.050	.052	.041	.041
EA-I	100	**.137**	**.079**	**.081**	.052	.053	**.378**	**.098**	**.108**	.036	.039
	150	**.120**	**.077**	**.081**	.056	.059	**.272**	.071	.074	.033	.042
	350	**.088**	.062	.062	.058	.058	**.135**	.043	.045	.029	.029
	600	**.085**	.065	.067	.059	.064	**.115**	.048	.051	.037	.039
EA-II	100	**.172**	**.099**	**.103**	.059	.062	**.404**	**.107**	**.112**	.041	.048
	150	**.112**	.070	.075	.053	.055	**.274**	**.090**	**.093**	.049	.054
	350	**.093**	.062	.064	.058	.060	**.135**	.056	.060	.041	.044
	600	.069	.057	.057	.055	.055	**.100**	.053	.056	.039	.039

Table 5 presents the results for data with 4 categories. The main change in the pattern of the results is that, relative to the data with fewer categories, $T_{\text{DWLS}-\text{MV1}}$ now performs worse, exhibiting inflated rejection rates, in S and MA conditions when the sample size is $N = 100$ or 150. However, relative to data with fewer categories, $T_{\text{DWLS}-\text{MV1}}$ performs better in the two EA conditions. $T_{\text{ULS}-\text{MV1}}$ performs better than $T_{\text{DWLS}-\text{MV1}}$ in almost all conditions. It is worth noting that as the number of categories has increased from 2 to 4, the results for all test statistics have become less differentiated as a function of the threshold conditions. Thresholds matter less as the data approach continuity.

Table 6 presents the results for data with 5 categories. The main change in the pattern of results is that the rejection rates in the S and both MA threshold conditions are uniformly higher. Even $T_{\text{ULS}-\text{MV1}}$, which tended to under-reject models with fewer categories, now exhibits slightly inflated rejection rates, particularly in smaller samples. Its performance in the S and MA conditions is still better than that of $T_{\text{DWLS}-\text{MV1}}$, however. Additionally, in the EA conditions, $T_{\text{ULS}-\text{MV1}}$ does very well, while $T_{\text{DWLS}-\text{MV1}}$ does poorly in small samples. Overall, the performance of all statistics is now worse in the MA conditions than in the EA conditions. Table 7, which presents data for 6 categories,

Table 7. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 6. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05. Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, *N*	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.242**	**.138**	**.145**	**.091**	**.096**	**.559**	**.201**	**.220**	**.079**	**.087**
	150	**.166**	**.103**	**.108**	.068	.068	**.358**	**.128**	**.133**	.055	.060
	350	**.109**	.072	.074	.061	.064	**.187**	**.081**	**.085**	.062	.067
	600	**.077**	.064	.064	.055	.056	**.126**	.060	.061	.047	.050
MA-I	100	**.237**	**.155**	**.160**	**.093**	**.101**	**.563**	**.208**	**.224**	**.088**	**.097**
	150	**.172**	**.115**	**.116**	**.085**	**.088**	**.416**	**.182**	**.187**	**.092**	**.096**
	350	**.124**	**.090**	**.093**	.074	**.077**	**.199**	**.092**	**.096**	.067	.069
	600	**.095**	.073	.073	.063	.065	**.132**	.064	.067	.053	.055
MA-II	100	**.239**	**.158**	**.162**	**.082**	**.088**	**.577**	**.236**	**.251**	**.096**	**.101**
	150	**.168**	**.114**	**.115**	**.079**	**.083**	**.348**	**.126**	**.134**	.064	.070
	350	**.112**	.075	**.076**	.057	.062	**.179**	**.086**	**.090**	.065	.065
	600	**.081**	.068	.071	.052	.055	**.140**	.072	.072	.062	.063
EA-I	100	**.183**	**.101**	**.107**	.052	.056	**.435**	**.117**	**.128**	.049	.054
	150	**.141**	**.094**	**.097**	.072	.074	**.284**	**.079**	**.082**	.049	.052
	350	**.093**	.061	.065	.052	.053	**.159**	.063	.065	.054	.056
	600	**.079**	.067	.067	.062	.063	**.116**	.064	.066	.048	.051
EA-II	100	**.177**	**.110**	**.113**	.069	.075	**.428**	**.131**	**.145**	.059	.066
	150	**.139**	**.086**	**.087**	.071	.072	**.287**	**.097**	**.101**	.048	.051
	350	**.104**	**.081**	**.084**	.072	.074	**.146**	.067	.071	.052	.053
	600	**.078**	.061	.062	.055	.056	**.119**	.064	.064	.054	.056

exhibits similar patterns, except that the performance of all statistics deteriorates slightly. This pattern continues in Table 8, which presents data for 7 categories. All test statistics over-reject at the smallest two sample sizes, but $T_{\text{ULS}-\text{MV1}}$ does much better than $T_{\text{DWLS}-\text{MV1}}$. The performance with EA thresholds is slightly better than the performance with MA or S thresholds.

Overall, the two mean- and variance-adjusted statistics followed somewhat different patterns. The cat-DWLS statistic $T_{\text{DWLS}-\text{MV1}}$ performed fairly well in S and the two MA conditions when the number of categories was 2 or 3, and then deteriorated for these conditions when the number of categories was 4–7. The cat-ULS statistic $T_{\text{ULS}-\text{MV1}}$ performed well or under-rejected in the S and the MA conditions when the number of categories was 2–4. In the EA conditions, $T_{\text{DWLS}-\text{MV1}}$ performed very poorly when the number of categories was 2, then showed increasing improvement as the number of categories increased from 3 to 4, then slowly began to deteriorate as the number of categories further increased from 5 to 7. In the EA conditions, $T_{\text{ULS}-\text{MV1}}$ performed well with 3–7 categories, but under-rejected a little with 2 categories.

Table 8. Rejection rates of five test statistics at $\alpha = .05$ when the number of categories is 7. The rates are out of the number of all converged cases. Values are highlighted if they are statistically greater than .05. Values are highlighted and in bold if they additionally fall outside Bradley's liberal criterion (between .025 and .075)

		Model 1					Model 2				
		DWLS			ULS		DWLS			ULS	
Threshold condition	Sample size, N	(1)	(2)	(4)	(3)	(5)	(1)	(2)	(4)	(3)	(5)
S	100	**.291**	**.193**	**.204**	**.126**	**.131**	**.665**	**.290**	**.315**	**.121**	**.134**
	150	**.193**	**.134**	**.138**	**.092**	**.095**	**.463**	**.190**	**.211**	**.095**	**.102**
	350	**.114**	**.079**	**.081**	.061	.063	**.199**	**.096**	**.097**	.070	.073
	600	**.104**	**.078**	**.082**	.073	**.077**	**.152**	**.075**	**.076**	.060	.061
MA-I	100	**.261**	**.172**	**.177**	**.097**	**.100**	**.620**	**.252**	**.271**	**.098**	**.113**
	150	**.179**	**.127**	**.130**	**.090**	**.094**	**.429**	**.160**	**.170**	.071	**.080**
	350	**.114**	**.089**	**.091**	**.078**	**.081**	**.213**	**.090**	**.093**	.065	.068
	600	**.097**	.072	.074	.067	.070	**.149**	**.084**	**.085**	.066	.068
MA-II	100	**.218**	**.154**	**.156**	**.094**	**.099**	**.593**	**.238**	**.261**	**.091**	**.103**
	150	**.174**	**.116**	**.120**	**.078**	**.082**	**.450**	**.185**	**.192**	**.097**	**.102**
	350	**.115**	**.077**	**.078**	.063	.064	**.262**	**.114**	**.121**	**.078**	**.083**
	600	**.090**	.066	.067	.060	.061	**.155**	**.080**	**.082**	.070	.071
EA-I	100	**.208**	**.129**	**.135**	**.081**	**.083**	**.534**	**.172**	**.185**	**.079**	**.083**
	150	**.144**	**.093**	**.093**	.072	.072	**.351**	**.107**	**.117**	.060	.065
	350	**.094**	.070	.072	.061	.061	**.165**	.065	.068	.051	.054
	600	**.080**	.063	.063	.058	.059	**.128**	.051	.052	.039	.043
EA-II	100	**.203**	**.126**	**.131**	**.079**	**.085**	**.521**	**.191**	**.208**	**.087**	**.098**
	150	**.146**	**.099**	**.107**	**.076**	**.077**	**.319**	**.098**	**.103**	.054	.061
	350	**.085**	.061	.063	.056	.056	**.179**	.074	.074	.053	.058
	600	**.080**	.049	.050	.045	.046	**.132**	.069	.071	.062	.063

5.3. Power

Table 9 presents selected power results for $T_{\text{ULS}-\text{MV1}}$ and $T_{\text{DWLS}-\text{MV1}}$. Only the smallest two sample sizes are presented. Power results are not interpretable when Type I error is not controlled, because inflated Type I error will always lead to artificially high power. Similarly, extremely low Type I error rates can lead to artificially low power. Because, in many conditions, $T_{\text{DWLS}-\text{MV1}}$ tended to exhibit inflated rejection rates (e.g., two-category data, EA thresholds, small samples), while $T_{\text{ULS}-\text{MV1}}$ tended to exhibit rejection rates below nominal, the power comparison of the two statistics is not very meaningful. To get around this problem, Table 9 simply highlights any cell that exhibits power less than .9, and additionally shows in bold any cell that exhibits power less than .8. Given that a grossly misspecified model is fitted to data (a one-factor model is fitted to two-factor data with a factor correlation of .3), it is reasonable to wish that power be at least .8 in such a situation.

Table 9 reveals that power is much better for the larger model (model 2) than for the smaller model (model 1). When a one-factor model is fitted to the two-factor data with 10 indicators per factor (model 2), power is always greater than .8 for data with 4–7 categories. For the S and the two MA conditions, power is greater than .9 for data

Table 9. Power of the new mean- and variance-adjusted test statistics (equations (4) and (5)) at $\alpha = .05$ at $N = 100$ and 150. Rejection rates are out of the number of all converged cases. Values less than .9 are highlighted. Values less than .8 are in bold.

		Model 1				Model 2			
		$N = 100$		$N = 150$		$N = 100$		N = 150	
Threshold condition	Number of categories	DWLS	ULS	DWLS	ULS	DWLS	ULS	DWLS	ULS
S	2	**.532**	**.422**	**.763**	**.706**	.881	**.764**	.988	.967
	3	**.730**	**.622**	.938	.897	.978	.922	.999	.997
	4	.883	.827	.981	.969	.997	.992	1.000	.999
	5	.929	.889	.997	.989	1.000	.997	1.000	1.000
	6	.962	.938	.996	.991	1.000	1.000	1.000	1.000
	7	.971	.944	.998	.997	1.000	1.000	1.000	1.000
MA-I	2	**.479**	**.358**	**.693**	**.612**	.857	**.711**	.970	.947
	3	**.790**	**.726**	.948	.928	.988	.961	1.000	.997
	4	.867	.812	.972	.955	.995	.985	1.000	1.000
	5	.919	.882	.989	.982	1.000	.995	1.000	1.000
	6	.955	.916	.995	.987	1.000	1.000	1.000	1.000
	7	.962	.942	.999	.997	1.000	1.000	1.000	1.000
MA-II	2	**.504**	**.396**	**.690**	**.634**	.857	**.723**	.974	.949
	3	**.782**	**.713**	.948	.922	.983	.965	.999	.999
	4	.864	.815	.973	.955	.998	.995	.999	.999
	5	.949	.907	.983	.970	.999	.997	1.000	1.000
	6	.941	.898	.992	.989	1.000	1.000	1.000	1.000
	7	.966	.941	.997	.995	1.000	1.000	1.000	1.000
EA-I	2	**.400**	**.075**	**.444**	**.186**	.917	**.135**	.917	**.498**
	3	**.508**	**.378**	**.729**	**.631**	.884	**.758**	.974	.950
	4	**.713**	**.626**	.889	.846	.980	.931	1.000	.999
	5	.818	**.747**	.952	.932	.986	.970	1.000	.999
	6	.888	.834	.977	.969	1.000	1.000	1.000	1.000
	7	.905	.881	.985	.981	1.000	1.000	1.000	1.000
EA-II	2	**.511**	**.040**	**.606**	**.162**	.956	**.041**	.983	**.388**
	3	**.536**	**.427**	**.720**	**.654**	.921	.826	.981	.955
	4	**.703**	**.621**	.884	.857	.970	.941	.999	.999
	5	.838	**.786**	.948	.927	.994	.984	1.000	.999
	6	.882	.835	.979	.966	1.000	.990	1.000	1.000
	7	.925	.889	.979	.973	1.000	1.000	1.000	1.000

with 3–7 categories, and it is reasonably high even for data with 2 categories, never falling below .7. The problematic conditions are the EA conditions with binary data, particularly when $N = 100$. Here, power is extremely high for $T_{\text{DWLS–MV1}}$ and extremely low for $T_{\text{ULS–MV1}}$. For instance, in the EA-II condition, power is .96 for $T_{\text{DWLS–MV1}}$ and an abysmal .04 for $T_{\text{ULS–MV1}}$. A comparison to Type I error rates is necessary to reveal the uselessness of both statistics in this situation. Type I error rates in this condition are .835 for $T_{\text{DWLS–MV1}}$ and .001 for $T_{\text{ULS–MV1}}$ (see Table 2). Thus, $T_{\text{DWLS–MV1}}$ tends to reject all models regardless of whether or not they are correct, and $T_{\text{ULS–MV1}}$ tends to accept all models regardless of whether or not they are correct. Thus, a combination of

binary data, small sample size, and extreme thresholds creates a situation where model evaluation is not possible using *any* test statistic.

When a one-factor model is fitted to the two-factor data with 5 indicators per factor (model 1), power is generally worse. In the S and the two MA conditions, power is greater than .8 for data with 4–7 categories. Power is worse, falling to .62, for the EA conditions for data with 4–7 categories. Binary and 3-category data present the most problems for power. In S and the two MA conditions, the two statistics have similar power in this situation. In the EA conditions, particularly with binary data, it is again the case that the test statistics diverge, and that both are useless. Power is as high as the Type I error rate for the $T_{\mathrm{DWLS-MV1}}$ statistic, and power is as low as the Type I error rate for the $T_{\mathrm{ULS-MV1}}$ statistic. Overall, one cannot recommend one statistic over another on the basis of power, because either they both perform fairly well, or, in the most difficult conditions, both fail.

Data for $N = 350$ are not presented. For model 2, power is at least .99 in all conditions and for both test statistics. For model 1, power is at least .99 for 3–7 categories across all conditions and for both test statistics. For binary data in the S and the MA conditions, power is at least .99. For binary data in the EA conditions, power is between .74 and .81. Data for $N = 600$ are also not presented. When $N = 600$, power is at least .99 for 3–7 categories, and at least .95 for binary data.

6. Summary and discussion

This paper has summarized the results of a Monte Carlo study conducted to compare the performance of five different categorical data test statistics available in M*plus* 6.11. Three of the statistics are associated with the DWLS estimator, and are the mean-adjusted and two types of mean- and variance-adjusted test statistic. Two of the statistics are associated with the ULS estimator, and are two types of mean- and variance-adjusted test statistic.

While some earlier research (Yang-Wallentin *et al.*, 2010) supports the use of the mean-adjusted DWLS statistic, $T_{\mathrm{DWLS-M}}$ (equation (1)), this statistic was found to perform very poorly, exhibiting extremely inflated Type I error rates in most conditions, particularly for the larger model 2. Its performance is only occasionally acceptable at the largest studied sample size and with the smaller model 1. Thus, while the mean-adjusted statistic is often found to perform well with continuous non-normal data, its categorical data counterpart is not recommended.

This study also examines two different versions of the mean- and variance-adjusted statistics, for both estimators. The original version (statistics $T_{\mathrm{DWLS-MV1}}$ and $T_{\mathrm{ULS-MV1}}$) adjusts the degrees of freedom (Satorra & Bentler, 1994; Muthén *et al.*, 1997; Muthén, 1993) of the reference distribution, which may be theoretically problematic. The new version (statistics $T_{\mathrm{DWLS-MV2}}$ and $T_{\mathrm{ULS-MV2}}$) does not require an adjustment for degrees of freedom, and thus has theoretical advantages (Asparouhov & Muthén, 2010). It was found, however, that the new versions of these statistics had slightly more inflated Type I error rates, although this difference typically did not exceed 1%. Thus, we tentatively recommend the new versions of the mean- and variance-adjusted statistics (which are now the default in M*plus*), although further study is perhaps needed to ensure that the inflation in Type I error rate does not become greater under some other set of conditions.

When comparing Type I error rates across the mean- and variance-adjusted statistics across estimators, it appears to be the case that the cat-ULS statistic did better overall than the cat-DWLS statistic. Its Type I error rates were almost never inflated, but it tended to exhibit very low rejection rates in some conditions, particularly with fewer

categories. The Type I error rates of the cat-DWLS statistic were frequently inflated, particularly with greater number of categories. Inflated Type I error rates are considered problematic. Type I error rates below nominal are not necessarily problematic unless they translate into much lower power. Thus, we recommend the cat-ULS statistic in any condition where its power is considered adequate (by Table 9), which is in the majority of the conditions studied. More generally, because cat-ULS estimates and robust standard errors have been found to be slightly superior to cat-DWLS estimates in previous research (Forero *et al*., 2009), we recommend the use of the cat-ULS estimator over the cat-DWLS estimator with categorical data, particularly in small to moderate samples.

The most problematic conditions for both statistics were created by the intersection of small samples, few categories, and extreme thresholds. This effect was mostly limited to $N = 100$, although sometimes $N = 150$, and to binary (and less frequently, 3-category) data. In these conditions, the cat-DWLS statistic had very high Type I error rates and power rates, so that the statistic would tend to reject any model. The cat-ULS statistic had very low Type I error rates and power rates, so that the statistic would accept any model. There is no remedy for this. We have to accept the fact that categorizing data leads to loss of information, and when this categorization is most severe (binary data), and done in such a way as to be least informative (extreme thresholds), a sample size of $N = 100$ is simply not large enough to provide information about the correctness of any particular model. With continuous data, it is possible to obtain information about the appropriateness of a model at $N = 100$. With severely categorical data, this sample size is just not enough. Thus, we recommend that with binary and 3-category data, samples of at least $N = 150$ be collected to draw any inferences about correctness of the hypothesized model. The only exception is when estimated thresholds appear symmetric; however, even in this case power tends to be low.

References

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

Babakus, E., Ferguson, C. E. J., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24,* 222–228. doi:10.2307/3151512

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203. doi: 10.1207/s15328007sem1302_2

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Chen, F., Bollen, K., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, *29*, 468–508. doi: 10.1177/0049124101029004003

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32. doi: 10.1007/BF02291477

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346. doi: 10.1207/S15328007SEM0903_2

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326. doi: 10.1111/j.2044-8317.1994.tb01039.x

Flora, D., & Curran, P. (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., & 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

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

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–390. doi: 10.1007/BF02296131

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1990). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, *55*, 45–51. doi: 10.1007/BF02294742

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–358. doi: 10.1111/j.2044-8317.1995.tb01067.x

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*, 495–507. doi: 10.1007/s11135-007-9133-z

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*, 209–227. doi: 10.1007/BF02294836

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020. doi: 10.1198/016214504000002069

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560. doi: 10.1007/BF02293813

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132. doi: 10.1007/BF02294210

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

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 manuscript.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analysing multitrait–multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results? *British Journal of Mathematical and Statistical Psychology*, *59*, 195–213. doi: 10.1348/000711005X67490

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, 485–500. doi: 10.1207/s15327906mbr1404_7

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., & Savalei, V. (2012). *When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions*. Manuscript submitted for publication.

Rigdon, E. E., & Ferguson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, *28*, 491–497. doi: 10.2307/3172790

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, *17*, 245–268. doi: 10.1080/10705511003659375

Savalei, V., & Kolenikov, S. (2008). Constrained vs. unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, 150–170. doi: 10.1037/1082-989X.13.2.150

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 variable analysis: Applications to developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Yang-Wallentin, F., Jöreskog, K., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*, 392–423. doi: 10.1080/10705511.2010.489003

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289–309. doi: 10.1111/j.2044-8317.1998.tb00682.x

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*, 93–110. doi: 10.1348/000711003321645368

Received 2 December 2011