Structural Equation Modeling, 16:625–641, 2009

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510903203573

Psychology Press
Taylor & Francis Group

Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation

Carlos G. Forero, Alberto Maydeu-Olivares, and David Gallardo-Pujol
University of Barcelona

Factor analysis models with ordinal indicators are often estimated using a 3-stage procedure where the last stage involves obtaining parameter estimates by least squares from the sample polychoric correlations. A simulation study involving 324 conditions (1,000 replications per condition) was performed to compare the performance of diagonally weighted least squares (DWLS) and unweighted least squares (ULS) in the procedure's third stage. Overall, both methods provided accurate and similar results. However, ULS was found to provide more accurate and less variable parameter estimates, as well as more precise standard errors and better coverage rates. Nevertheless, convergence rates for DWLS are higher. Our recommendation is therefore to use ULS, and, in the case of nonconvergence, to use DWLS, as this method might converge when ULS does not.

Factor analysis is often performed on ordinal indicators, such as Likert-type ratings. In this case, the scores assigned to the ratings can be treated as if they were continuous and a standard factor analysis model is employed. This is unsatisfactory as predicted values for the observed scores can never be exact integers, whereas the scores assigned to the ratings are exact integers (McDonald, 1999). Alternatively, the ordinal nature of such scores can be taken into account by adding a threshold response process to a standard factor analysis model. This ensures that the predicted values are exact integers and therefore it is in principle a better suited model for this kind of data. Henceforth, we refer to a factor analysis model with a threshold response process as the ordinal factor analysis model.

Christoffersson (1975) was the first to propose a sound estimation method for a factor analysis model with binary indicators. He used the matrix of sample cross-products using weighted least squares (WLS) estimation, where the weight matrix is the inverse of a consistent estimate of the covariance matrix of the sample statistics. Parameter estimates are consistent and asymptotically normal and they have minimum variance among the class of estimators using

Correspondence should be addressed to Carlos G. Forero, Faculty of Psychology, University of Barcelona, P. Valle de Hebrón, 171, 08035 Barcelona, Spain. E-mail: carlos.garcia.forero@ub.edu

univariate and bivariate information. This procedure yields asymptotically correct standard errors for the parameter estimates and a chi-square goodness-of-fit test.

However, Christoffersson's (1975) method is computationally demanding because univariate and bivariate integration is needed at each iteration of the estimation process. Muthén (1978) proposed a three-stage estimation procedure with a reduced computational cost using the inverse of a consistent estimate of the covariance matrix of the estimated tetrachoric correlations as weight matrix. Thus, integration is only performed to estimate the tetrachorics, but the number of observed variables that can be analyzed is still limited by the need to store the weight matrix. This procedure was extended so that the model could include indicators of different measurement scales, as well as covariates (Muthén, 1984). The WLS estimator turned out to be impractical, however, as it was found to converge very slowly to its asymptotic properties (Muthén & Kaplan, 1985, 1992). Therefore, very large sample sizes are needed to obtain accurate estimates and standard errors with more than a few indicators.

Christoffersson (1977) and McDonald (1982) devised two-stage estimation procedures for binary data that could handle a larger number of variables. Their proposals differed in the weight matrix used in the second stage: diagonally weighted least squares (DWLS: Christoffersson, 1977) or unweighted least squares (ULS: McDonald, 1982). By using DWLS or ULS, a much larger number of indicators can be handled, but at the price of estimates with larger asymptotical variances than WLS. Additionally, neither standard errors nor goodness-of-fit tests were directly available. A breakthrough took place when Muthén (1993), drawing on work later published by Satorra and Bentler (1994), provided formulae that allowed obtaining asymptotic standard errors and goodness-of-fit tests for a three-stage ULS procedure. This ULS estimator provided more accurate estimates, standard errors, and goodness-of-fit tests than WLS in finite samples (Muthén, 1993). Additionally, these formulae can be adapted for other fitting functions using polychorics, such as DWLS or the maximum likelihood (ML) fitting function (see Jöreskog, Sörbom, du Toit, & du Toit, 1999).

Currently, M*plus* (Muthén & Muthén, 2006), LISREL (Jöreskog & Sörbom, 2005), and EQS (Bentler, 2006) implement three-stage estimation of the ordinal factor analysis model. M*plus* and LISREL use the procedure in Olsson (1979) in the first two stages to estimate the polychoric correlations. Although M*plus* and LISREL use asymptotically equivalent formulae for estimating the asymptotic covariance matrix of the polychorics (Muthén, 1984; see also Jöreskog, 1994; Muthén & Satorra, 1995), they may differ to some extent in finite samples (Maydeu-Olivares, 2006). In the third stage, WLS, DWLS, and ULS are available in M*plus*; LISREL also implements ML. EQS also uses a three-stage procedure that estimates polychoric correlations using formulae given in Lee, Poon, and Bentler (1995) and implements WLS and ULS in the third stage.

In recent years, the use of DWLS as an estimation method for ordinal factor analysis has become popular. One reason for this might be that it is a scale-invariant estimator if continuous indicators are used. This means that if the fitted model is scale invariant, and the indicators are continuous, DWLS yields the same fit function minima and linearly transformed parameters whenever the data are linearly transformed. ULS does not show this property. However, when all the indicators are ordinal, scale invariance is irrelevant. Another reason for its popularity might be that, computationally, DWLS is only slightly more involved than ULS and, asymptotically, the variance of DWLS estimates is smaller than that of ULS estimates. Yet, this is not necessarily the case in finite samples.

A number of studies (e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004) have addressed DWLS performance in finite samples. Other studies have also addressed the performance of ULS in finite samples (Bolt, 2005; Boulet, 1996). Nevertheless, we are only aware of three studies that have compared the relative performance of ULS and DWLS. Rigdon and Fergurson (1991) conducted a simulation study to compare the performance of both methods in estimating a two-factor model with four indicators per factor. Maydeu-Olivares (2001) compared ULS and DWLS estimation of Thurstonian models of paired comparisons and reported comparable performance of both estimators, with slightly better results for ULS. Tate (2003) conducted a comparison of methods with dichotomous indicators with correlated and uncorrelated factors, and found that DWLS and ULS showed comparable parameter recovery across all conditions of number of factors, item difficulty, and factor loading. All three studies have clear limitations. In Rigdon and Fergurson's (1991) study, only a few conditions were investigated. Also, they could not compare the standard errors because their study was performed before the asymptotic standard errors for DWLS and ULS were correctly computed (see their quote on p. 496 of a personal communication by Jöreskog, 1989). The other two studies only considered binary data. Also, Maydeu-Olivares considered nonstandard models, whereas Tate used just one replication per condition.

To fill in this gap, we performed an extensive simulation study to investigate whether ULS or DWLS yields more accurate parameter estimates and standard errors in finite samples. To do so, we considered in ordinal factor analysis models when all the indicators are ordinal. WLS is not considered because it is well established that it performs very poorly unless sample size is large and model size is small (Dolan, 1994; Flora & Curran, 2004; Muthén & Kaplan, 1992).

THE ORDINAL FACTOR ANALYSIS MODEL

Consider a questionnaire consisting of n items y_i, $i = 1, \ldots, n$, to be rated using one of m response alternatives. These alternatives can be scored using the successive integers $k = 0, \ldots, m-1$. The ordinal factor analysis model assumes that a set of n latent response variables $\mathbf{y}^*$ underlies the n observed categorical responses $\mathbf{y}$. The latent response variables are related to the observed categorical responses via a threshold relationship,

$$y_i = k \text{ if } \tau_{i,k} < y_i^* < \tau_{i,k+1}. \tag{1}$$

where $\tau_{i,0} = -\infty$ and $\tau_{i,m-1} = \infty$. That is, an individual will choose response alternative k when his latent response value y_i^* is between thresholds $\tau_{i,k}$ and $\tau_{i,k+1}$. It is also assumed that factors are linked to the latent responses $\mathbf{y}^*$ by means of a standard factor analytic model

$$\mathbf{y}^* = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\eta}$ is a $p \times 1$ vector of factors, $\mathbf{\Lambda}$ is a $n \times p$ matrix of factor loadings, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of measurement errors. By assumption, both the factors $\boldsymbol{\eta}$ and the measurement errors $\boldsymbol{\varepsilon}$ are normally distributed. It is also assumed that the mean of the factors and measurement errors is zero, and that the factors and measurement errors are uncorrelated. Generally, it is

further assumed that measurement errors are mutually uncorrelated, so that their covariance matrix, $\mathbf{\Theta}$, is diagonal. Given these assumptions, latent responses $\mathbf{y}^*$ are normally distributed with mean zero and covariance matrix

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}' + \mathbf{\Theta} \tag{3}$$

where $\mathbf{\Psi}$ is the covariance matrix of the factors. The model is identified by the general identification rules of the factor model (see, e.g., Bollen, 1989; McDonald, 1999), except that the variances of the measurement errors (i.e., the diagonal elements of $\mathbf{\Theta}$) are not identified. The ordinal factor analysis model can be identified by setting

$$\mathbf{\Theta} = \mathbf{I} - \text{diag}(\mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}'), \tag{4}$$

so that the covariance matrix of the latent responses $\mathbf{y}^*$, given in Equation 3, is a correlation matrix, $\mathbf{P}$. Note that because the latent responses $\mathbf{y}^*$ underlying the observed ordinal variables are normally distributed, $\mathbf{P}$ is a matrix of tetrachoric correlations. It turns out that this model is mathematically equivalent to an item response theory (IRT) model proposed by Samejima (1969)—see Takane and de Leeuw (1987) for details—which is referred to in the IRT literature as normal ogive graded response model.

Let $\boldsymbol{\theta}$ denote the vector of mathematically independent parameters in $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{\Theta}$ (i.e., factor loadings, correlations among the factors, and—possibly—covariances among the measurement errors). In this article we consider three-stage estimation methods: In a first stage, thresholds are estimated one variable at a time using ML. In a second stage, each polychoric correlation ρ is estimated separately—also by ML—from each pair of variables using the thresholds from the previous stage. In the third stage, the thresholds and polychoric correlations estimated in the previous two stages are collected in the vector $\hat{\boldsymbol{\kappa}} = (\hat{\boldsymbol{\tau}}', \hat{\boldsymbol{\rho}}')'$, and model parameters are estimated by minimizing a least square function. If no restrictions are imposed on the thresholds, then a least squares function based on the polychoric correlations alone can be employed (Muthén, 1978)

$$F = (\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta}))'\hat{\mathbf{W}}(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})) \tag{5}$$

where $\boldsymbol{\rho}(\boldsymbol{\theta})$ denotes the restrictions imposed on the population polychoric correlation matrix.

The minimization procedure depends on the weight matrix $\hat{\mathbf{W}}$ used in Equation 5. Let $\hat{\mathbf{\Gamma}}$ be an estimate of the asymptotic covariance matrix of estimated polychoric correlations. We can choose $\hat{\mathbf{W}} = \hat{\mathbf{\Gamma}}^{-1}$, an alternative known as WLS (Muthén, 1978, 1984). A second choice is $\hat{\mathbf{W}} = (diag(\hat{\mathbf{\Gamma}}))^{-1/2}$, which uses as weights only the estimated variances of the estimated polychoric correlations. This method is known as DWLS (Muthén, du Toit, & Spisic, 1997). A third, widely used choice is $\hat{\mathbf{W}} = \mathbf{I}$, a method known as ULS (Muthén, 1993).

All three choices of weight matrices yield consistent and asymptotically normal estimates. Also, asymptotically correct standard errors can be obtained for all four estimators. The formulae are provided, for instance, in Jöreskog et al. (1999). Asymptotically, the best estimator is WLS, as it is the one that provides estimates with smallest variance. It is also the only estimator for which $N\hat{F}$ (the minimum of the fit function multiplied by sample size) is

asymptotically chi-square. For all other estimators, $N\hat{F}$ can be adjusted by its asymptotic mean (or by its asymptotic mean and variance) as suggested by Muthén (1993; see also Satorra & Bentler, 1994) to obtain a goodness-of-fit test. Because previous research has shown that in finite samples the asymptotically optimal WLS performs the worst, in this research we investigate by means of a simulation study which of the two most widely used remaining estimators, ULS and DWLS, yields better results in finite samples.

DESCRIPTION OF THE SIMULATION STUDY

A simulation study was conducted to compare DWLS and ULS in estimating a factor analysis model with categorical ordered indicators under different settings of dimensionality, factor loading, sample size, number of items per factor, number of response alternatives per item, and item skewness. A total of 324 conditions per estimation method were investigated, using 1,000 replications for each setting. A full factorial design was used by crossing three sample sizes (200, 500, and 2,000 respondents); two levels of factor dimensionality (one and three factors); three test lengths (9, 21, and 42 items); three levels of factor loadings λ: low ($\lambda = .4$), medium ($\lambda = .60$), and high ($\lambda = .8$); and six item types (three types consist of items with two categories, and another three of items with five categories) that varied in skewness, kurtosis, or both.

Sample sizes were chosen to range from small to large in typical applications. Two hundred observations was the smallest sample size. Flora and Curran (2004) found that DWLS began to show parameter overestimation bias when using small sample sizes ($N \leq 200$). Small to medium test lengths were chosen because prior results suggest that the performance of DWLS improves with increasing test length (Finger, 2002; Oranje, 2003).

The item distributions used in the study are depicted in Figure 1. Note that Types I to III are dichotomous. The threshold of Type III items was chosen such that only 10% of the respondents endorse the correct category. Type II items are endorsed by 15% of the respondents and, consequently, present smaller amounts of skewness and kurtosis. Forty percent of the respondents endorse Type I items. Indicators of Types IV through VI have five response categories. Type IV skewness and kurtosis match those of a standard normal distribution. Type V and Type VI items are considerably skewed. In these items, the probability of endorsing a certain category decreases as the category label increases.

Preliminary simulation runs with both methods revealed (see the discussion section) that decreasing the correlation among the factors decreases convergence rates and decreases parameter estimation and standard error estimation accuracy. To test the methods under the most stringent conditions—and highlight differences among them—three-dimensional models were set up to be orthogonal. As a result, there are six levels of number of indicators per factor (3, 7, 9, 14, 21, and 42 items).

Finally, we include items with low (.4) to high (.8) factor loadings in typical applications (Briggs & MacCallum, 2003; Ximénez, 2006). Factor loading levels were chosen to represent weak, medium, and strong factors. Although equal factor loadings across indicators were used for data generation to facilitate the reporting of findings, they were not constrained to be equal during estimation.

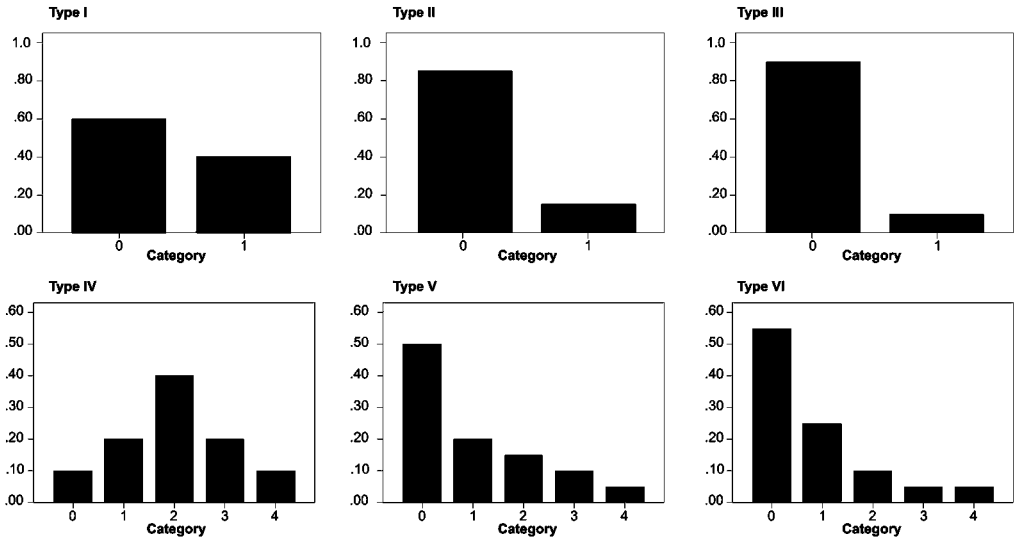

FIGURE 1 Bar graphs of the different types of items employed in the simulation study. Type I, II, and III had skewness equal to 0.40, 2.00, 2.65, and kurtosis of 1.17, 4.84, 8.11, respectively. Type IV, V, and VI items had skewness equal to 0.00, 0.98, 1.5 and kurtosis of 2.50, 2.80, and 4.31, respectively.

All simulations were performed using M*plus* version 4.2 (Muthén & Muthén, 2006), and default convergence criteria were used for both methods. The following outcomes were investigated: (a) proportion of proper solutions per condition, (b) relative bias of parameter estimates, (c) relative bias of standard errors, and (d) coverage rates.

Two repeated-measures analyses of variance (ANOVAs) were performed to test the effects of the simulation conditions on estimated factor loadings (λ) and their standard errors. Estimation method (ULS vs. DWLS) was used on identical data sets, and therefore it was used as a within-subjects factor. Because data sets were generated independently, the remaining simulation conditions (sample size, dimensionality, test length, factor loading, number of item categories and item skewness) were entered as between-subject factors. The large number of replications ($n = 648{,}000$) might cause even negligible effects to reach statistical significance. Actually, all effects in the analysis were found to be significant at $p < .0001$, and for the sake of readability we are not providing p values in the results section. Instead, results were assessed using partial η^2 to evaluate the importance of each effect.

RESULTS OF THE SIMULATION STUDY

Convergence Rates

Convergence rate was defined as the proportion of replications per condition that converged with the M*plus* default values, excluding improper solutions. A solution was deemed improper when at least one estimated factor loading was larger than or equal to 1 in absolute value.

As in Flora and Curran (2004), these solutions were considered invalid and removed from subsequent analyses. Notice that improper or nonconvergent solutions are of no use to the applied researcher (Chen, Bollen, Paxton, Curran, & Kirby, 2001).

On average, convergence rates across the 324 conditions were 97.4% for DWLS and 96.4% for ULS. However, convergence rates differed depending on the number of indicators per dimension, item skewness, and sample size. Both estimators showed smaller convergence rates for models with only three indicators per dimension. In this setting, convergence rates were better for DWLS: Average convergence was 90.6% for DWLS versus 85.4% for ULS. When the number of indicators per dimension was seven or more, average convergence rates were similar (roughly 99%). Increasing skewness worsened convergence: When item skewness was greater than or equal to 1.5, average convergence was 96.4 for DWLS and 94.7 for ULS. When item skewness was below 1.5, convergence performance was, on average, similar across the methods (98%). Finally, sample size improved convergence rates.

Perhaps the most shocking result obtained is that these estimation methods have difficulties in estimating models where sample size is 200, the number of indicators per factor is only 3, and item skewness is large (≥ 1.5). When these conditions are simultaneously met, the average convergence rate is 57% for ULS and 71% for DWLS. However, in one of the conditions convergence rate is as low as 19.5% for ULS and 27.9% for DWLS. Yet, a sample size of 500 observations and at least seven indicators per dimension were sufficient to produce acceptable convergence rates (i.e., at least 80% on average) for both methods, regardless of item skewness.

Relative Bias of Parameter Estimates

Relative bias of parameter estimates was computed as a proportion using $(\overline{\theta} - \theta)/\theta$, where $\overline{\theta}$ is the average parameter estimate across valid replications and θ denotes the true parameter value. A relative bias below 10% was considered acceptable. Values from 10% to 20% indicated substantial bias, whereas those above 20% were deemed unacceptable.

Figure 2 depicts graphically the relative bias for the factor loading parameters for all 324 conditions. As can be seen in this table, ULS is acceptably biased (relative bias $< 10\%$) for all conditions, whereas DWLS showed a few biased conditions. When the DWLS bias was positive, models involved models with three indicators per factor. When the bias was negative, models involved more than seven indicators. More specifically, DWLS yielded negatively biased conditions for models with three dichotomous indicators per factor, low factor loadings ($\lambda = 0.4$) and $N < 2{,}000$. DWLS yielded negatively biased estimates for conditions where skewness was 2.65, $N = 200$, and the number of indicators was larger than seven.

The repeated-measures ANOVA showed that main effects and interactions up to the third order between method and between-subject factors explained very little variance of the factor loading estimates ($R^2 = .37$). When considering interactions between method and single between-subject factors, the most important sources of variance were produced by Method $\times$ Sample Size, $F(2, 323784) = 52519.36$, partial $\eta^2 = .247$; Method $\times$ Test Length, $F(2, 323784) = 63497.73$, partial $\eta^2 = .282$; and Method $\times$ Factor Loading. The most important third-order interaction was Method $\times$ Test Length $\times$ Factor Loading, $F(4, 323784) = 73154.72$, partial $\eta^2 = .475$. Interactions between Method $\times$ Dimensionality $\times$ Test Length, $F(2, 323784) = 75743.23$, partial $\eta^2 = .319$, and interactions between Method $\times$

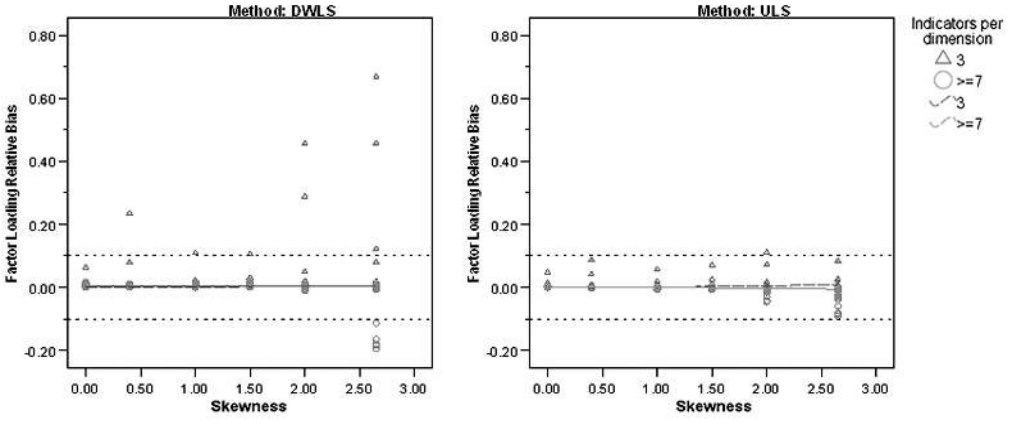

FIGURE 2 Relative bias of factor loading estimates. Dashed reference lines indicate ±10% bias. A nonparametric regression procedure was used to model the relationship between relative bias and item skewness by number of indicators per dimension.

Dimensionality × Factor Loading, $F(2, 323784) = 73662.88$, partial $\eta^2 = .313$, were also of great importance among the third-order effects. Fourth-order interactions increased the percentage of explained variance to 68.9%. The most important effects among fourth-order effects were Method × Dimensionality × Test Length × Factor Loading, $F(4, 323784) = 73469.94$, partial $\eta^2 = .476$; Method × Dimensionality × Sample Size × Test Length × Factor Loading, $F(8, 323784) = 22745.58$, partial $\eta^2 = .360$; and Method × Test Length × Factor Loading × Skewness, $F(4, 323784) = 34076.86$, partial $\eta^2 = .296$.

Including fifth-order interactions increased R^2 to .89. Again, the most important interactions included the effects of dimensionality, test length, sample size, skewness, and factor loading. The most relevant fifth-order effects were Method × Dimensionality × Test Length × Sample Size × Factor Loading, $F(8, 323784) = 21631.57$, partial $\eta^2 = .348$; and Method × Dimensionality × Test Length × Factor Loading × Skewness, $F(4, 323784) = 31677.54$, partial $\eta^2 = .281$. The R^2 value reached .94 when taking into account sixth-order interactions. Among these interactions, Method × Dimensionality × Test Length × Sample Size × Factor Loading × Skewness, $F(8, 323784) = 14800.69$, partial $\eta^2 = .268$. The seventh-order interaction among all factors did not explain a substantial amount of variance.

The ANOVA model provided evidence of the importance of sample size, dimensionality, test length, factor loading, and skewness, but effects from these factors appeared in the form of high-order interactions. To shed additional light on the relative performance of ULS versus DWLS estimates, Table 1 shows the average relative bias of factor loading by method, skewness level, number of indicators per factor, and true parameter value. As can be seen in this table, ULS relative bias is always equal to or smaller than DWLS relative bias. Also, in only one entry of this table is relative bias larger than our cutoff criterion for good performance, namely when the number of indicators per factor is three, factor loadings are .4, item skewness is equal or larger than 1.5, and DWLS estimation is used. In this entry of the table, DWLS relative bias is 23%. In contrast, for ULS it is only 5%. Table 1 also shows that accuracy increases for

TABLE 1
Average Relative Bias of λ Parameter Estimates for Each Method by Number of Indicator Per Factor, Item Skewness, and True λ Parameter

		Method					
		ULS			*DWLS*		
Indicators Per Factor	*Skewness*	λ *0.40*	*0.60*	*0.80*	λ *0.40*	*0.60*	*0.80*
3	<1.5	.03	.00	.00	.06	.00	.00
	≥1.5	.05	.01	.00	.23	.02	.00
7	<1.5	.00	.00	.00	.00	.00	.00
	≥1.5	−.02	−.01	.00	−.01	.00	.00
9	<1.5	.00	.00	.00	.01	.00	.00
	≥1.5	−.02	−.01	.00	−.02	.00	.00
14	<1.5	.00	.00	.00	.00	.00	.00
	≥1.5	−.02	−.01	.00	−.02	.00	.00
21	<1.5	.00	.00	.00	.01	.01	.00
	≥1.5	−.02	−.01	.00	−.02	.01	.01
42	<1.5	.00	.00	.00	.01	.00	.00
	≥1.5	−.02	−.01	.00	−.02	.01	.01

Note. ULS = unweighted least squares; DWLS = diagonally weighted least squares.

both methods with increasing true factor loading so that when $\lambda \geq .6$ most conditions showed below 1% in absolute value.

Relative Bias of Standard Errors

Standard error relative bias was computed using $(\overline{SE}_\theta - sd_\theta)/sd_\theta$, where $\overline{SE}_\theta$ was the average standard error of a parameter estimate across valid replications and sd_θ the standard deviation of the parameter estimates across valid replications.

The ANOVA results for standard errors revealed a complex pattern of high-order interactions. Main effects from method and second-order interactions between method and between-subject effects explained very little variance of the factor loading standard errors ($R^2 = .12$). Adding third-order interactions to the model increased R^2 to .43. The most important third-order effects were Method × Test Length × Factor Loading, $F(4, 323784) = 415743.61$, partial $\eta^2 = .837$, and Method × Dimensionality × Test Length, $F(2, 323784) = 525720.40$, partial $\eta^2 = .765$. Also of great importance were the effects Method × Number of Categories × Test Length, $F(2, 323784) = 422020.42$, partial $\eta^2 = .723$, and Method × Number of Categories × Test Length, $F(2, 323784) = 418394.58$, partial $\eta^2 = .721$).

Taking into account fourth-order interactions notably enlarged the proportion of explained variance ($R^2 = .77$). Many fourth-order terms yielded important effects (partial $\eta^2 > .30$), but the largest effect came from the Method × Dimensionality × Test Length × Factor Loading interaction, $F(4, 323784) = 409984.52$, partial $\eta^2 = .835$. Other substantial effects came from the interactions Method × Number of Categories × Test Length × Factor Loading, $F(4, 323784) = 319132.63$, partial $\eta^2 = .798$, and Method × Number of Observations × Test Length × Factor Loading, $F(8, 323784) = 123859.08$, partial $\eta^2 = .754$.

Fifth-order interactions increased the proportion of explained variance to $R^2 = .95$. There were important effects from Method × Dimensionality × Test Length × Factor Loading × Number of Categories, $F(4, 323784) = 315403.75$, partial $\eta^2 = .796$, and Method × Dimensionality × Test Length × Factor Loading × Sample Size, $F(8, 323784) = 121316.40$, partial $\eta^2 = .750$. Sixth-order interactions still improved the amount of explained variance by the model ($R^2 = .99$), specifically from the Method × Dimensionality × Test Length × Sample Size × Factor Loading × Skewness interaction, $F(8, 323784) = 21163.22$, partial $\eta^2 = .343$. Again, the seventh-order interaction did not contribute substantially to the full model.

Figure 3 depicts graphically the relative bias of the standard errors for all 324 conditions investigated. The figure shows that for most conditions standard errors relative bias was acceptable ($< |10\%|$) for both ULS and DWLS. However, when the bias was not acceptable, the bias' magnitude was often very large. As a result, the Y-axes in Figure 3 are on a logarithmic scale, so that very different bias values could be represented while providing an illustrative overall picture of performance. The results shown in Figure 3 reveal that the magnitude of the relative bias for the failed conditions (i.e., bias $> |10\%|$) is larger for DWLS than for ULS. Also, we see that among the failed conditions bias is positive when the number of indicators per factor is three, but negative when the number of indicators per factor is seven or larger. Although not readily apparent in Figure 3, it is worth noting that DWLS failed to yield acceptable standard errors for all conditions where skewness = 2.65.

To shed additional light on the relative performance of ULS versus DWLS standard errors, Table 2 provides the average bias for the factor loadings' standard errors by method, skewness level, true parameter value, and number of indicators per factor. In most cells in this table, ULS provides more accurate standard errors than DWLS. Unacceptable standard errors appear when item skewness ≥ 1.5, true factor loading is low or medium ($\lambda \leq 0.60$) or the number of indicators per factor is three. The more of these three conditions are involved, the more

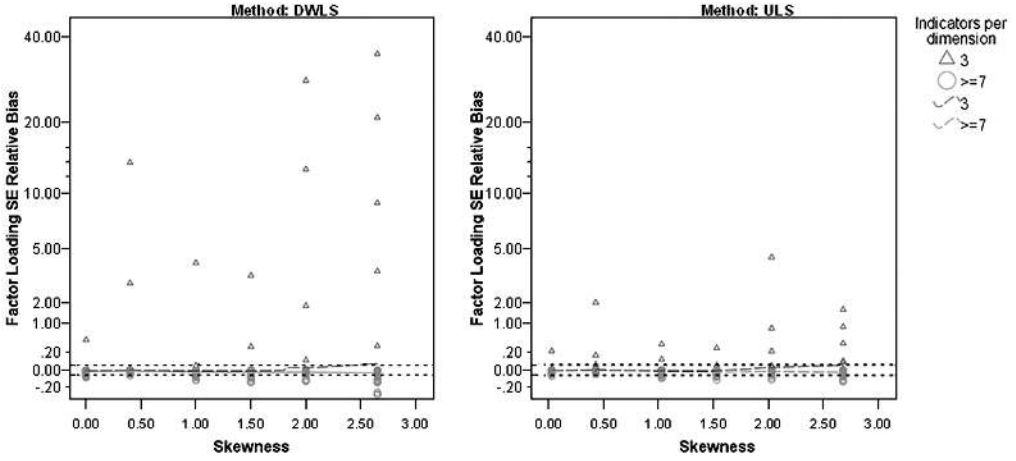

FIGURE 3 Standard error estimates of factor loadings. The Y-axis is in a logarithmic scale. Dashed reference lines indicate ±10% bias. A nonparametric regression procedure was used to model the relationship between relative bias and item skewness by number of indicators per dimension.

TABLE 2
Average Relative Bias of the Standard Errors for λ Parameter Estimates for Each Method by Number of Indicators Per Factor, Item Skewness, and True λ Parameter

		Method					
		ULS			*DWLS*		
Indicators Per Factor	*Skewness*	*λ 0.40*	*λ 0.60*	*λ 0.80*	*λ 0.40*	*λ 0.60*	*λ 0.80*
3	<1.5	.38	.00	.00	2.35	−.01	−.01
	≥1.5	.99	.13	.02	11.61	1.23	−.01
7	<1.5	−.04	−.02	−.01	−.05	−.02	−.02
	≥1.5	−.09	−.03	−.02	−.10	−.05	−.03
9	<1.5	−.05	−.03	−.02	−.06	−.04	−.03
	≥1.5	−.10	−.04	−.03	−.13	−.07	−.04
14	<1.5	−.02	−.01	−.01	−.03	−.02	−.02
	≥1.5	−.06	−.03	−.03	−.12	−.05	−.04
21	<1.5	−.04	−.02	−.02	−.05	−.04	−.03
	≥1.5	−.06	−.03	−.03	−.14	−.06	−.04
42	<1.5	−.03	−.02	−.02	−.04	−.03	−.03
	≥1.5	−.05	−.04	−.03	−.12	−.06	−.05

Note. ULS = unweighted least squares; DWLS = diagonally weighted least squares.

likely it is standard errors are unacceptable. In contrast, when none of these conditions is met, standard errors are acceptable. Also, when skewness ≥ 1.5 and the true factor loadings are .4, standard errors are unacceptable unless the number of indicators is larger than nine for ULS, but it is unacceptable for DWLS, regardless of the number of indicators.

In sum, in most conditions standard errors were acceptable for both methods. However, standard error inaccuracies were quite large when they did appear. DWLS was especially prone to produce extremely biased standard errors, often showing more than 100% relative bias for the biased conditions.

Parameter Coverage

Figure 4 shows the coverage of 95% confidence intervals for parameter estimates for all 324 conditions investigated. We see in this figure that ULS and DWLS factor loadings' coverage was generally good for both methods (between 92% and 98% for 95% intervals). Figure 4 reveals that coverage variability increased with increasing skewness, and problems appeared at large skewness levels (≥ 2.5). In general, ULS offered more precise coverage than did DWLS, due to more accurate factor loading estimates and smaller standard errors.

DISCUSSION

This simulation study sought to investigate the relative performance of two asymptotically suboptimal least squares methods (DWLS and ULS) in estimating factor analysis models with

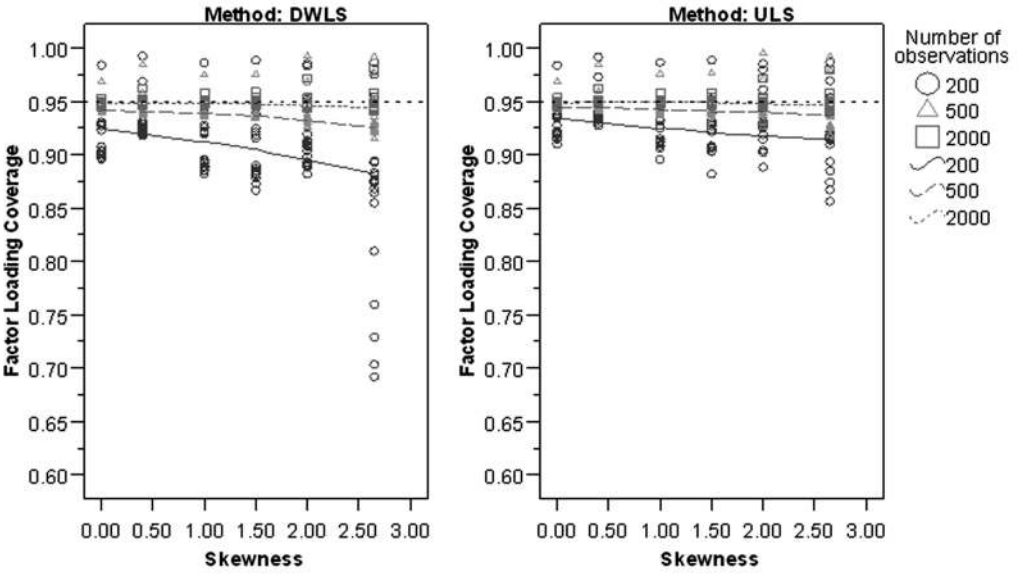

FIGURE 4 Proportion of conditions (coverage) where 95% confidence intervals for parameter estimates include true parameter. Coverage should be close to nominal rate (95%). A nonparametric procedure has been used to model the relationship between coverage and item skewness by sample size.

ordinal categorical indicators. We manipulated a comprehensive set of factors that could influence their performance, resulting in 324 conditions. The conditions were chosen to highlight the differences among the estimators, as well as to cover a wide variety of conditions. Results demonstrate that the pattern of relative biases, both from the estimates and the standard errors, depended on a complex pattern of high-order interactions. In general, the factors involved in such interactions are combinations of sample size, number of indicators per factor, and skewness, which had a slightly different effect on each method. However, estimates and standard errors had a tendency to fail under similar conditions.

DWLS failed in 62 of the 324 conditions, where parameter or standard error bias was larger than 10% (our cutoff criteria for "good" performance). Specifically, it was more likely to fail in conditions involving the combination of (a) low sample sizes ($N = 200$), (b) low factor loadings ($\lambda = .4$), or (c) high skewness (≥ 1.5). The more of these factors that were involved the greater the likelihood that DWLS yielded biased parameter estimates, standard errors, or both. Few indicators per factor (seven or fewer) posed an even more challenging setting in these conditions: DWLS failed in all conditions involving three or seven indicators per factor when sample size was 200 observations and the factor loading was low. It also failed under the preceding conditions when sample size was 500 and the skewness was over 2.00.

Overall, the performance of ULS was excellent and it failed in just 37 conditions. Even in these settings, coverage was close to the nominal 95%, and in 34 of these conditions DWLS failed, too. The other three conditions in which ULS (but not DWLS) failed had three indicators per factor, but no other systematic pattern was evident among them. Among the conditions where both ULS and DWLS failed, 78.6% involved models with nine or fewer indicators, 73%

TABLE 3
Summary Table Comparing DWLS Versus ULS Performance

		ULS		
		Succeeds		
		Performance Better Than DWLS	*Performance Worse Than DWLS*	*Fails*
λ	DWLS succeeds	56.5%	23.5%	0.9%
	DWLS fails	4.9%	3.7%	10.5%
λSE	DWLS succeeds	68.2%	11.7%	0.9%
	DWLS fails	8.6%	0.0%	10.5%

Note. ULS = unweighted least squares; DWLS = diagonally weighted least squares. A total of 324 conditions were investigated. A successful condition is defined as a condition in which the relative biases of λ estimates, and λ standard errors are smaller than 10% (in absolute value). A failed condition is defined as a condition in which at least one of these two criteria is not met. For those conditions where at least one method succeeds, the table provides the percentage of all conditions where ULS outperforms DWLS in each of the two criteria.

involved low factor loadings (λ = 0.4) and 75% involved skewness above 1.5. Interestingly, both ULS and DWLS failed in estimating the hardest model (nine dichotomous indicators, three factors, lowest factor loading), even at the highest sample size employed (2,000) and for the two highest skewness levels (1.96 and 2.67).

To gain further insight into which method (DWLS or ULS) is more accurate, the percentage of conditions in which ULS is more precise than DWLS across two criteria (factor loading relative bias, and factor loading standard error relative bias) was computed and is summarized in Table 3. As can be seen, ULS clearly outperforms DWLS in estimating factor loadings and their standard errors whenever both methods succeed (less than 10% bias).

ULS yields somewhat smaller biases and smaller standard deviations for factor loading estimates. This is shown in Table 4, which provides the standard deviation of factor loading estimates by method, skewness level, true parameter value, and number of indicators per factor. Factor loadings standard deviations were, on average, smaller for ULS. This is especially evident when using models with just three indicators per factor and true factor loadings smaller than .8. In this case, the standard deviation of the estimates may be reduced by almost half when using ULS with low factor loadings and skewness < 1.5. The reduction might be threefold in the same setting with skewness ≥ 1.5. To some extent this is one of the reasons why the bias of DWLS is larger than that for ULS in conditions in which unacceptable results are obtained. Also partially as a result of this, coverage for factor loadings is better when estimation is based on ULS.

CONCLUSIONS

Previous research on the finite sample behavior of ordinal factor analysis estimators had found that asymptotically suboptimal estimators (ULS and DWLS) provide, in most cases,

TABLE 4
Average Standard Deviation of λ Parameter Estimates for Each Method by Number of Indicators Per Factor, Item Skewness, and True λ Parameter

		Method					
		ULS			*DWLS*		
Indicators Per Factor	*Skewness*	λ *0.40*	*0.60*	*0.80*	λ *0.40*	*0.60*	*0.80*
3	<1.5	.12	.07	.04	.22	.08	.04
	≥1.5	.18	.11	.06	.58	.20	.06
7	<1.5	.07	.05	.03	.07	.05	.03
	≥1.5	.12	.08	.04	.13	.08	.04
9	<1.5	.08	.05	.03	.08	.05	.03
	≥1.5	.12	.07	.04	.13	.08	.04
14	<1.5	.06	.04	.02	.06	.04	.02
	≥1.5	.10	.06	.04	.11	.06	.04
21	<1.5	.06	.05	.03	.06	.05	.03
	≥1.5	.09	.06	.04	.10	.06	.04
42	<1.5	.06	.04	.03	.06	.04	.03
	≥1.5	.08	.06	.04	.09	.06	.04

Note. ULS = unweighted least squares; DWLS = diagonally weighted least squares.

more accurate parameter estimates and standard errors than the asymptotically optimal WLS estimator. However, not enough research had been conducted to guide applied researchers in choosing between both methods. Our study has provided some useful insights into the behavior of these estimators in a large number of conditions.

One of the most relevant results for applied researchers is the existence of conditions for which neither estimator yields adequate results. Whenever possible, these conditions should be avoided in applications: (a) a small number of indicators per dimension, (b) binary items, (c) low factor loadings (around $\lambda = .4$), (d) high item skewness (≥ 1.5), and (e) small sample size (around 200 observations). The more of these factors that are involved, the greater the likelihood that the estimators yield inadequate estimates and standard errors. Not surprisingly, increasing sample size protects against estimation problems. In most settings, a sample of 2,000 observations guarantees that both estimators will perform well.

Another relevant result for applied researchers concerns convergence rates. An improper or nonconvergent solution is of no use to the applied researcher: A model is useless if its estimation does not converge. Evidence suggests that improper solutions mainly affect the problematic cases (i.e., models involving one or more of the factors listed earlier), and both estimators had convergence problems in these cases.

Yet, the main conclusion of this research is that except in a small number of conditions (around 10%) both estimators provide adequate parameter estimates, standard errors, and parameter coverage. This is excellent news for applied researchers as some of the conditions investigated were truly difficult. The focus of this study, however, was the comparative behavior of DWLS in relation to ULS. In this regard, the results are clear: DWLS generally outperforms ULS in convergence rates, but ULS outperforms DWLS in estimation accuracy. In problematic

cases DWLS is more likely to converge than is ULS, but whenever ULS converges it yields more accurate parameter estimates and standard errors than does DWLS. In nonproblematic cases, the behavior of the two estimators is similar, with ULS being slightly more precise. Also, and in contrast to asymptotic results, ULS estimates have smaller standard deviations for the sample sizes considered here.

The implications of this study are limited by the specification of the conditions employed. For instance, we did not experimentally manipulate the correlations among the factors. We run additional simulations where we experimentally manipulated the correlations among the factors. As in Flora and Curran (2004), we found higher convergence rates and more precise parameter estimates and standard errors when factors were correlated. Also, as in the results reported in this article, ULS yielded slightly more accurate estimates for the parameters, interfactor correlations, and standard errors than DWLS, whereas both methods showed similar convergence rates when the factors were correlated. As for the case of models with higher order factors, the accuracy of the estimates and standard errors depend on the estimation accuracy in first-level factors. Preliminary simulations indicate that convergence and estimation accuracy problems are aggravated in the presence of second-order factors or multidimensional indicators. Applied researchers should be aware that estimates in second-order factors should be greatly affected by the harshest conditions in this study.

Another issue is that only factor analysis models were considered here. However, a small simulation study involving Thurstonian models for paired comparisons (Maydeu-Olivares, 2001) also found that ULS outperformed DWLS. A further topic for future research is the behavior of both estimators when the factor analysis model is misspecified. Once again, previous studies (e.g., Flora & Curran, 2004; Maydeu-Olivares, 2006) suggest that the ordinal factor analysis estimates are robust to mild model misspecification.

Future research should also address other estimators. For instance, a fourth alternative to the weight matrix $\hat{\mathbf{W}}$ used in Equation 5 is to employ $\hat{\mathbf{W}} = \mathbf{D}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \hat{\boldsymbol{\Sigma}}^{-1})\mathbf{D}$, where $\otimes$ denotes a Kronecker product and $\mathbf{D}$ is the duplication matrix described in Magnus and Neudecker (1988). This iteratively reweighted least squares function has the minimum at the same point that the ML fitting function used to estimate the model parameters from polychorics. Another alternative is the use of the polychoric instrumental variable estimator recently proposed by Bollen and Maydeu-Olivares (2007).

In conclusion, the results of our simulation study enable us to offer clear advice to applied researchers regarding the use of ULS versus DWLS in estimating an ordinal factor analysis model. In general, use ULS as it provides more accurate and less variable parameter estimates, as well as more precise standard errors. However, in the case of nonconvergence of ULS, use DWLS as this method might converge when ULS does not.

ACKNOWLEDGMENTS

This study was partially supported by Grant SEJ2006-08204/PSIC (Albert Maydeu-Olivares, Principal Investigator) from the Spanish Ministry of Education and a 2006 Dissertation Support Award from the Society of Multivariate Experimental Psychology.

REFERENCES

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

Bentler, P. M. (2006). *EQS structural equation modeling software.* Encino, CA: Multivariate Software.

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

Bollen, K. A., & Maydeu-Olivares, A. (2007). Polychoric instrumental variable (PIV) estimator for structural equations with categorical variables. *Psychometrika, 3,* 309–326.

Bolt, D. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics. A festschrift for Roderick P. McDonald* (pp. 27–71). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Boulet, J. R. (1996). *The effect of nonnormal ability distribution on IRT parameter estimation using full-information methods.* Unpublished doctoral dissertation, University of Ottawa, Ottawa, Canada.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38,* 25–56.

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

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika, 40,* 433–438.

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.

Finger, M. S. (2002, April). *A comparison of full information and unweighted least squares limited-information item parameter estimation methods used with the two-parameter normal ogive model.* Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.

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.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381–389.

Jöreskog, K. G., & Sörbom, D. (2005). *LISREL 8.72*. Chicago, IL: Scientific Software.

Jöreskog, K. G., Sörbom, D., du Toit, S., & du Toit, M. (1999). *LISREL 8: New statistical features.* Chicago, IL: Scientific Software.

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.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics.* New York: Wiley.

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

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika, 71,* 57–77.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Muthén, B. (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., 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, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

Muthén, B., & Muthén, B. (2006). *M*plus (Version 4.1). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60,* 489–503.

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

Oranje, A. (2003, April). *Comparison of estimation methods in factor analysis with categorized variables: Applications to NAEP data.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Rigdon, E. E., & Fergurson, 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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4).

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.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27,* 159–203.

Ximénez, M. C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling, 13,* 587–614.