CHAPTER 10

# Estimation in Structural Equation Modeling

**Pui-Wa Lei**
**Qiong Wu**

Model estimation and evaluation are integral parts of any application of structural equation modeling (SEM). Quality of model parameter estimates, their associated standard error estimates, and overall model fit statistics depend on the choice of appropriate estimation methods. Desirable properties of estimators include asymptotic consistency, unbiasedness, and efficiency. An estimator is consistent if it approaches the true parameter as sample size increases toward infinity, unbiased if its expected value equals the parameter that it estimates (i.e., the average of estimates from an infinite number of independent samples from the same population will equal the population parameter), and efficient if its variability is the smallest among consistent estimators. Some estimators are also asymptotically normally distributed, allowing adequate significance testing of individual parameters using the *z*-test.

## SAMPLE DATA AND MODEL PARAMETERS

The basic elements of data for SEM analyses are sample variances and covariances of observed variables. Given a hypothesized SEM model, individual observed variables can be written as a function of unknown parameters (i.e., path coefficients or factor loadings) and other observed or latent variables in the model. These functions describe structural relations (or causal hypotheses) among the variables and are referred to as "structural equations." From the set of structural equations, variances and covariances of observed variables can be expressed in terms of unknown parameters in the model, such as path coefficients, factor loadings, and variances and covariances of latent variables. These variances and covariances are model-specific and are called model-implied variances and covariances.

As an example, suppose that a simple four-variable two-factor confirmatory factor analysis (CFA) model, as depicted in Figure 10.1, is being estimated from a sample variance–covariance matrix (see upper half of Table 10.1). The factors of interest are reading ($F_1$) and mathematics ($F_2$). Reading is indicated by basic word reading ($X_1$) and reading comprehension ($X_2$) scores. Mathematics is indicated by calculation ($X_3$) and reasoning ($X_4$) scores. There are four structural equations (ones that stipulate the causal relationship among variables) for the model in Figure 10.1, one for each observed variable: $X_1 = 1 * F_1 + \delta_1$, $X_2 = \lambda_{21} * F_1 + \delta_2$, $X_3 = 1 * F_2 + \delta_3$, and $X_4 = \lambda_{42} * F_2 + \delta_4$. In the equations, the $\lambda$'s represent factor loadings, and the $\delta$'s represent measurement errors. The equations suggest that each observed variable (outcome) is a function of a common factor (cause) and a unique measurement error (other unspecified causes). Two loadings are fixed to 1 to set

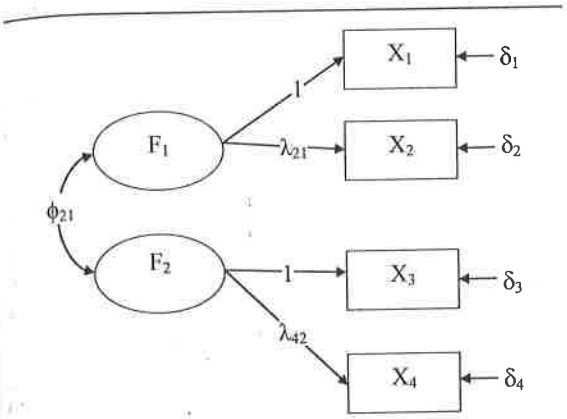


**FIGURE 10.1.** A CFA model with unknown model parameters.

the scale for the factors (i.e., $F_1$ takes the scale of $X_1$; $F_2$ takes the scale of $X_3$).

Assuming independence between latent factors and measurement errors, variances (V) and covariances (C) among the observed variables can be expressed in terms of model parameters based on the structural equations using covariance algebra—for example, $V(X_1) = C(X_1, X_1) = C(1 * F_1 + \delta_1, 1 * F_1 + \delta_1) = V(F_1) + V(\delta_1)$. $C(X_1, X_2) = C(1 * F_1 + \delta_1, \lambda_{21} * F_1 + \delta_2) = \lambda_{21} * V(F_1)$ because the $\delta$'s are uncorrelated in the model.

The complete set of model-implied variances and covariances for this example are arranged in a matrix form, parallel to the sample counterpart, and shown in the lower half of Table 10.1. The model parameters to be estimated for this CFA model include factor loadings, factor variances, covariance between the two factors, and variances of measurement errors. Readers who wish to learn more about covariance algebra can consult Bollen (1989, pp. 21–23). However, standard SEM software programs will calculate the model-implied variance–covariance matrix given a user-supplied path diagram or a set of structural equations.

## GENERAL ESTIMATION PROCEDURES

Model parameters ($\theta$, which is a generic notation for all unknown parameters in the model) are estimated by minimizing some form of discrepancy between a sample variance–covariance matrix (S) and model-implied variance–covariance matrix [$\Sigma(\theta)$]. This is similar to the ordinary least squares (OLS) estimation in regression in which the sum of squared differences between observed and predicted values for individual observations is minimized. The difference is that observations in regression are individual scores, while observations in SEM are sample variances and covariances.

Moreover, a closed-form solution (one that can be obtained analytically) is available for regression, but it is often unavailable for SEM. SEM generally relies

**TABLE 10.1. Sample and Model-Implied Variance–Covariance Matrices for the Model in Figure 10.1**

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| | | Sample variance–covariance matrix (S) | | |
| $X_1$ | 211.60 | — | — | — |
| $X_2$ | 178.78 | 247.32 | — | — |
| $X_3$ | 125.12 | 158.87 | 236.12 | |
| $X_4$ | 116.85 | 163.21 | 181.20 | 227.78 |
| | | Model-implied variance–covariance matrix [$\Sigma(\theta)$] | | |
| $X_1$ | $V(F_1) + V(\delta_1)$ | — | — | — |
| $X_2$ | $\lambda_{21} * V(F_1)$ | $\lambda_{21} * \lambda_{21} * V(F_1) + V(\delta_2)$ | — | — |
| $X_3$ | $C(F_1, F_2) = \phi_{21}$ | $\lambda_{21} * \phi_{21}$ | $V(F_2) + V(\delta_3)$ | |
| $X_4$ | $\lambda_{42} * \phi_{21}$ | $\lambda_{21} * \lambda_{42} * \phi_{21}$ | $\lambda_{42} * V(F_2)$ | $\lambda_{42} * \lambda_{42} * V(F_2) + V(\delta_4)$ |

*Note.* — indicates redundant symmetric values.

on iterative procedures to solve for unknown parameters in the model. Educated guesses of parameter values, known as starting or initial values ($\hat{\theta}_0$), which are usually supplied by standard SEM programs, are used to start the iterative process. An intermediate model-implied variance–covariance matrix can then be calculated by substituting the unknown parameters with their initial values [$\Sigma(\hat{\theta}_0)$]. Minimizing the discrepancy between the sample and the intermediate model-implied variance–covariance matrices will provide a new set of parameter estimates. This new set of parameter estimates will replace the old set of initial values, and the process repeats until changes in parameter estimates between adjacent iterations become acceptably small. The rules used to stop the iterative process are called "convergence criteria." When convergence criteria are met, the model is said to have converged, and the last set of parameter estimates is taken to be the final solution for the unknown parameters. Parameter estimates for the above example are shown in Table 10.2.

Different estimation methods or estimators minimize different functions of the discrepancy between S and $\Sigma(\theta)$, called fit or discrepancy functions ($F$). A model fit statistic is $T = (N - 1) * F$, where $N$ is sample size, and $F$ is the minimum of the fit function when the model converges. When assumptions of estimators are met, $T$ often approximately follows a $\chi^2$ distribution with degrees of freedom ($df$) equal to the number of unique variances and covariances minus the number of estimated model parameters. Therefore, overall fit of the model to data can be assessed using a $\chi^2$ test. The asymptotic variance–covariance matrix of an estimator, if it exists, provides standard error estimates of individual parameter estimates for significance tests.

In the following, we introduce some popular estimation methods used in SEM, including the maximum likelihood family, least squares family, and Bayesian method. We also briefly review empirical findings about these methods and illustrate the importance of choosing estimation methods in consideration of commonly encountered conditions in practice. For concern of space, we limit our discussion to basic covariance structure models and estimation methods that are relatively more widely used and studied. Our selection of estimation methods is by no means inclusive or comprehensive.

**TABLE 10.2. Parameter Estimates for the Model in Figure 10.1**

| | Unstandardized parameter estimates | Standard errors | Standardized parameter estimates |
|---|---|---|---|
| Factor loadings | | | |
| $X_1$ | 1[a] | | .79 |
| $X_2$ ($\lambda_{21}$) | 1.34* | .15 | .99 |
| $X_3$ | 1[a] | | .87 |
| $X_4$ ($\lambda_{42}$) | 1.02* | .10 | .90 |
| Factor variances and covariance | | | |
| V($F_1$) | 133.21* | 29.73 | 1 |
| V($F_2$) | 177.23* | 34.78 | 1 |
| C($F_1$, $F_2$) or $\phi_{21}$ | 118.71* | 24.92 | .77 |
| Error variances | | | |
| V($\delta_1$) | 78.39* | 15.09 | .37 |
| V($\delta_2$) | 7.39 | 18.35 | .03 |
| V($\delta_3$) | 58.89* | 14.95 | .25 |
| V($\delta_4$) | 42.52* | 14.28 | .19 |

[a]Values are fixed to 1 to set the scale of latent factors.
*Estimates are significantly different from 0 at the .05 level.

## MAXIMUM LIKELIHOOD ESTIMATORS

Contemporary SEM appears to originate from econometrics, psychometrics, and mathematical statistics (Bielby & Hauser, 1977). Karl G. Jöreskog is often credited for unifying path analysis from econometrics and factor analysis from psychometrics into a general framework of covariance structure analysis (e.g., Bielby & Hauser, 1977). Lawley (1940) was the first to apply maximum likelihood (ML) estimation in factor analysis, but the iterative numerical procedure used by Lawley and several others (e.g., Bargmann, 1957; Emmett, 1949) was not very successful. A major breakthrough came in the 1970s when Jöreskog and his colleagues (Jöreskog, 1977; Jöreskog & Lawley, 1968) introduced a numerical solution for ML factor analysis and developed the LISREL (LInear Structural RELations) software program (Jöreskog & van Thillo, 1973). The availability of the LISREL program has played a significant role in popularizing SEM. ML remains the most well known and widely used estimator to date. The fit function for ML given by Bollen (1989) is shown in Equation 10.1.

$$F_{ML} = \log|\Sigma(\theta)| + \mathrm{tr}(S\Sigma^{-1}(\theta)) - \log|S| - p \quad (10.1)$$

In Equation 10.1, log(.) is the natural logarithm function, tr(.) is the trace function, and $p$ is the number of observed variables. Under the assumption of multivariate normality of observed variables and a correct model specification, the ML estimator is asymptotically consistent, unbiased, efficient, and normally distributed, and the model fit statistic ($T_{ML}$) is asymptotically distributed as $\chi^2$ with $df = p(p+1)/2 - t$, where $t$ is the number of model parameters estimated.

The ML estimator tends to produce relatively unbiased parameter estimates (provided that the proper covariance matrix is analyzed and that the model is correctly specified) but inflate model $\chi^2$ and deflate standard error estimates under non-normality (e.g., Bollen, 1989, pp. 417–418; Chou, Bentler, & Satorra, 1991; Finch, West, & MacKinnon, 1997). Fortunately, a robust asymptotic covariance matrix for the estimated parameters, from which robust standard errors can be obtained, and corrections to the model fit statistic ($T_{ML}$) to better approximate the reference $\chi^2$ distribution are documented in Satorra and Bentler (1988, 1994). The $\chi^2$ correction entails adjusting the mean of the test statistic alone, resulting in the Satorra–Bentler scaled statistic, or adjusting both the mean and variance resulting in the Satorra–Bentler adjusted statistic (Fouladi, 2000). The Satorra–Bentler scaled $\chi^2$ statistic is available in most specialized SEM software programs. Both versions of adjustment are available in Mplus (Muthén & Muthén, 1998–2010) with the mean-adjusted version labeled as MLM and mean- and variance-adjusted version as MLMV. Little is known about the performance of MLMV, perhaps because it is not available in most SEM programs other than Mplus.

More recently, Asparouhov and Muthén (2005) developed another more general robust ML estimator (called MLR in Mplus) based on Skinner's (1989) pseudo-ML method and by using adjustments similar to those of Satorra and Bentler (1988) to deal with complex sampling designs. Initial evidence provided by the authors seems to support its use for survey data. Additionally, MLR can be used for non-normal data (Muthén & Muthén, 1998–2010, p. 484).

Another variant of ML developed by Finkbeiner (1979) to handle missing data is referred to as the full-information maximum likelihood (FIML) estimator. Assuming data are missing at random (MAR), FIML maximizes a modified log-likelihood function that makes use of all available individual observations (e.g., Enders, 2006, pp. 322–323). Therefore, raw individual data are required for FIML. This approach to handling missing data is logically appealing because there is no need to make additional assumptions for imputation and no loss of observations. It is regarded as one of the state-of-the-art treatments of missing data (Schafer & Graham, 2002). It has also been found to work better than listwise deletion in simulation studies (Kline, 2010, p. 59).

## LEAST SQUARES ESTIMATORS

Before the availability of LISREL, OLS estimation (i.e., multiple regression technique) had been used to derive estimates for recursive path models (e.g., Blalock, 1964; Duncan, 1966). However, the capability of OLS is very limited because it cannot deal with nonrecursive path models or measurement models (Kline, 2005, p. 159). Therefore, the use of OLS in SEM is very rare nowadays.

The two-stage least squares (2SLS) estimation method can be considered an extension of the OLS method for handling nonrecursive models and models with la-

tent variables. The 2SLS method often involves the use of instrumental variables. An instrumental variable has a direct effect on a "problematic" causal variable (i.e., one that is correlated with the equation's disturbance), but no direct effect on the endogenous variable (Kline, 2010, p. 156). See Figure 10.2 for a conceptual illustration of an instrumental variable ($I$) for a "problematic" causal variable ($X$) in predicting an endogenous variable ($Y$). In the enclosed part of Figure 10.2, $X$ is a "problematic" predictor because it is correlated with the disturbance of $Y$ (D). The variable $I$ is an appropriate instrumental variable for $X$ because it has a direct effect on $X$ but not on $Y$. The 2SLS estimator applies multiple regressions in two stages. In the first stage, the "problematic" causal variable is regressed on the instrumental variable(s) (i.e., $X = \beta_0 + \beta_1 I + e$ for the example, where the $\beta$'s are regression coefficients, and $e$ is the error term for $X$). In the second stage, the endogenous variable is regressed on the predicted value of the "problematic" causal variable from the first stage (i.e., $Y = \pi_0 + \pi_1 \hat{X} + u$ for the example, where the $\pi$'s are regression coefficients, and $u$ is the error term for $Y$). The purpose of this two-stage process is to replace the "problematic" causal variable with its predicted value (predicted by the instrumental variables) that is uncorrelated with the equation's disturbance. Since not all parameters are estimated simultaneously, 2SLS is a limited-information method. Limited-information methods may be less susceptible to spreading of model misspecification to other equations than full-information estimation methods such as ML (Bollen, Kirby, Curran, Paxton, & Chen, 2007).

Variants of the 2SLS estimator have been developed since the 1950s in econometrics in the context of simultaneous equation models (e.g., Basmann, 1957). Jöreskog (1983) also proposed a 2SLS estimator to estimate starting values for the LISREL program (Jöreskog & Sörbom, 1993). The 2SLS estimator developed by Bollen (1996a, 1996b, 2001) is probably the most general in that "it permits correlated errors across equations, . . . estimates intercepts, and provides asymptotic covariance matrix of the estimator for significance testing" and that it is equivalent to other versions of 2SLS under certain conditions (Bollen et al., 2007, p. 54). According to Bollen and colleagues (2007), the 2SLS estimator is consistent, asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient among limited information estimators, and the version proposed by Bollen (1996a, 1996b, 2001) provides an accurate asymptotic covariance matrix without assuming normality of observed variables.

One advantage is that the 2SLS estimator does not require a specialized SEM program to implement it. Researchers can use any software programs that have 2SLS procedures or that perform OLS regression; however, researchers may need to make proper adjustments to the estimates of the standard errors manually (Bollen, 1996a). Moreover, Bollen and Bauer (2004) developed an automatic algorithm to help select model-implied instrumental variables and provided a Statistical Analysis Software (SAS)/interactive matrix language (IML) macro to implement it.

Both OLS and 2SLS are noniterative, limited-information estimation methods, and the computation of parameter estimates does not require any starting values. Full-information least squares estimation methods that simultaneously estimate all parameters are generally iterative and require starting values to successively minimize a particular fit function of the difference between the vector of elements in the sample variance and covariance matrix ($s$) and the vector of elements in the model-implied variance and covariance matrix [$\sigma(\theta)$]. The $s$ and $\sigma(\theta)$ vectors are of order $p(p + 1)/2$ containing unique elements from $\mathbf{S}$ and $\Sigma(\theta)$, respectively. The fit function that the least squares family minimizes has the general form of Equation 10.2.

$$F_{LS}(\theta) = [s - \sigma(\theta)]' W^{-1} [s - \sigma(\theta)] \quad (10.2)$$

Equation 10.2 defines a family of estimation methods sometimes known as generalized least squares (e.g., Anderson & Gerbing, 1988) or weighted least squares (e.g., Bollen, 1989, p. 425; Kline, 2005, p. 159). However, for the purpose of this chapter, we use these terms for specific methods discussed later rather than as a family of methods.

Different least squares estimation methods employ different weight matrices, $\mathbf{W}$'s, which are $p(p + 1)/2$

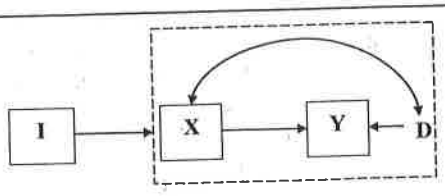


**FIGURE 10.2.** A conceptual illustration of an instrumental variable ($I$) for a "problematic" causal variable ($X$).

by $p(p + 1)/2$ square matrices. If an identity matrix (**I**) is used as **W**, for example, the estimation method is reduced to the unweighted least squares method (ULS). ULS is a consistent estimator, and it makes no distributional assumption about the observed variables (Bollen, 1989, p. 112). However, it requires all observed variables to be measured on the same scale, and it is generally less efficient than the ML estimator (Kline, 2010, p. 176).

Expositions of the normal theory–based generalized least squares (GLS) method appeared in the 1970s (e.g., Anderson, 1973; Browne, 1974; Jöreskog & Goldberger, 1972). One form of the fit function for GLS is $F_{GLS}(\theta) = \frac{1}{2}\text{tr}(\{[\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\mathbf{V}^{-1}\}^2)$, where $\mathbf{V} = \mathbf{S}$ is a $p$ by $p$ square matrix. Note that when **V** is chosen to be $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$, minimizing this form of the fit function leads to the normal theory–based ML (Lee & Jennrich, 1979). According to Bollen (1989, pp. 428–429), $F_{GLS}$ has been shown to be equal to Equation 10.2 by Browne (1974) and can be considered a special case of Equation 10.2. Using the more general form of Equation 10.2, the weight matrix, **W**, for GLS contains elements that are functions of the second-order product moments around the mean, $[\mathbf{W}_{GLS}]_{ij,gh} = s_{ig}s_{jh} + s_{ih}s_{jg}$, $i \geq j$, $g \geq h$, where $s_{ig}$ represents the covariance between observed variables $i$ and $g$. Like ML, GLS assumes multivariate normality or no excessive kurtosis (Browne, 1974) and is consistent, asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient among full-information estimators.

The weighted least squares (WLS) estimator proposed by Browne (1984) relaxes the distributional assumption and is referred to as Browne's asymptotically distribution-free (ADF) method. In spite of this name, it should be noted that Browne's ADF method is not the only one that makes no distributional assumptions. A number of other methods, including the 2SLS and the ULS methods mentioned earlier, also do not make distributional assumptions. The WLS estimator uses the asymptotic covariance matrix of sample variances and covariances as its weight matrix with a typical element consisting of estimates of the second- and fourth-order product moments around the mean, $[\mathbf{W}_{WLS}]_{ij,gh} = s_{ijgh} - s_{ij}s_{gh}$, $i \geq j$, $g \geq h$, where

$$s_{ijgh} = \frac{\sum_{n=1}^{N}(X_{ni} - \bar{X}_i)(X_{nj} - \bar{X}_j)(X_{ng} - \bar{X}_g)(X_{nh} - \bar{X}_h)}{N}$$

and $N$ is sample size. Under correct model specification and multivariate normality, the individual elements of $\mathbf{W}_{GLS}$ and $\mathbf{W}_{WLS}$ will converge in probability to $\sigma_{ig}\sigma_{jh} + \sigma_{ih}\sigma_{jg}$ (Hu, Bentler, & Kano, 1992). Research has shown that WLS has a strict requirement on sample size and may produce large amounts of bias with small samples (e.g., Hoogland & Boomsma, 1998).

Because inverting the full-weight matrix of WLS is computationally demanding, and it is likely the culprit for the poor performance of WLS with less than large samples, diagonally weighted least squares methods, in which only diagonal elements of the WLS weight matrix are kept in the weight matrix, are often used to lessen the computational burden. A version of the diagonally weighted least squares (DWLS) estimator is available in LISREL, beginning in version 7 (Jöreskog & Sörbom, 1988). Two other versions known as mean-adjusted WLS (WLSM) and mean- and variance-adjusted WLS (WLSMV; Muthén, 1993; Muthén, du Toit, & Spisic, 1997) are available in Mplus (Muthén & Muthén, 1998–2010).

DWLS estimators can be and are often used when some or all observed endogenous variables are non-normal and categorical. They are commonly used with alternative correlations that estimate association between latent response variables for ordinal data (i.e., polychoric correlation between categorical variables, or polyserial correlation between categorical and continuous variables). WLSM and WLSMV were specifically developed for categorical variable methodology and were referred to as robust WLS estimators by Muthén and colleagues (1997). A robust asymptotic covariance matrix for the estimated parameters and corrected model $\chi^2$ test statistics similar to those of Satorra and Bentler (1994) are available for these estimators. WLSM differs from WLSMV in the adjustment to the model $\chi^2$ test statistic and in their degrees of freedom. Degrees of freedom for WLSMV are estimated from the sample, and they can vary from sample to sample for the same model.

## BAYESIAN ESTIMATION

According to Dunson, Palomo, and Bollen (2005), there is a long history of Bayesian methods in factor analysis and general SEM models. The earliest work on factor analysis models (Martin & McDonald, 1975) was cited by the authors in the 1970s, and that on general SEM models (Bauwens, 1984), in the 1980s. Recent developments of Bayesian estimation in SEM have focused on

the use of Markov Chain Monte Carlo (MCMC) methods to handle complex cases, including nonlinear structural models and multilevel data (Dunson et al., 2005).

Bayesian methods take a completely different perspective on estimation, in that they assume true model parameters are random and unknown, whereas in ML, true parameters are considered fixed and unknown but their estimates are random (Arbuckle, 1995–2009). In the Bayesian framework, parameters' prior distributions are combined with empirical data likelihood based on Bayes's theorem to form posterior distributions for parameter estimates. Since analytical approaches to obtaining posterior distribution are most often impractical due to the difficulty of estimating high-dimensional integrals, empirical approximations are adopted by simulating values based on Monte Carlo procedures. Regular Monte Carlo procedures that draw independent samples may not be feasible because posterior distributions are often of an unknown form. Under such circumstances, MCMC can be used to draw dependent samples from a series of distributions that is in the limit equal to drawing from the stationary posterior distribution (Gilks, Richardson, & Spiegelhalter, 1996). More details about Bayesian estimation in SEM can be found in Kaplan and Depaoli, Chapter 38, this volume.

Compared to the frequentist approach (i.e., ML and LS), Bayesian estimation has a few advantages. First, ML and LS confidence intervals assume that parameter estimates are asymptotically normal, whereas Bayesian credibility intervals are based on percentiles of the posterior distribution, which is not restricted to any fixed form. Second, when models are computationally intractable for ML, the Bayesian method can come to the rescue. Hence, the Bayesian method can simply be viewed as a computational tool to obtain parameter estimates (Muthén & Muthén, 1998–2010). A disadvantage of the Bayesian method is that it is computationally intensive and may take a long time to obtain a solution with an acceptably low level of Monte Carlo error (Dunson et al., 2005).

## SOFTWARE PROGRAMS THAT PERFORM THE DIFFERENT ESTIMATION METHODS

There are a number of software programs that estimate SEM models. We chose the four most frequently used programs (Amos 18: Arbuckle, 2009; EQS 6.1: Bentler, 2005; LISREL 8.8: Jöreskog & Sörbom, 2006; Mplus 6: Muthén & Muthén, 2010) and briefly describe their similarities and differences in terms of their estimation capacities.

Estimators that are available in all four software programs include ML, FIML, GLS, ULS, and WLS or ADF. ML and GLS assume either multivariate normality or no excessive kurtosis of observed variables. ML is the default estimator for all four programs when observed endogenous variables are continuous. When raw data are analyzed, missing code is provided, and when ML is requested, these programs produce FIML solutions. When data distributions deviate from multivariate normality, different estimation options are available in different programs.

The robust ML approach, that is, using regular ML for model estimation, along with robust standard errors and scaled model $\chi^2$ of Satorra and Bentler (1988, 1994) to evaluate model fit, is a popular choice for non-normal continuous data. It is available in LISREL 8.8 (by including an asymptotic covariance matrix of the variances and covariances estimated by PRELIS), EQS 6.1 (with METHOD = ML, ROBUST), and Mplus 6 (with ESTIMATOR = MLM), but not in Amos 18. Bootstrapping is available in Amos 18, so that one can estimate standard errors and model fit statistics empirically by applying this resampling technique on the sample data at hand. Moreover, when the keyword ROBUST is used in EQS 6.1, the program provides three additional residual-based test statistics, including Browne's (1984) original residual-based $\chi^2$, the Yuan–Bentler extension of Browne's residual-based $\chi^2$ test for small samples (Bentler & Yuan, 1999; Yuan & Bentler, 1998), and an even more radical modification of Browne's test called the Yuan–Bentler residual-based $F$-statistic (Yuan & Bentler, 1998). Another robust ML estimator for non-normal or cluster data, called MLR, is currently available only in the Mplus 6 program.

The ADF estimator can also be considered for model estimation with non-normal continuous data when the sample size is very large (i.e., in the thousands). ADF or WLS is available in all four programs. LISREL 8.8 provides ADF solutions when an asymptotic covariance matrix estimated by PRELIS is submitted as input and the WLS option is requested. Mplus 6 also uses the keyword WLS, while EQS 6.1 uses AGLS and Amos 18 uses ADF for Browne's (1984) ADF estimator. Moreover, EQS 6.1 provides two different adjustments to the model test statistic called the Yuan–Bentler corrected arbitrary distribution generalized least squares (AGLS) test statistic (Yuan & Bentler, 1997b) and the Yuan–Bentler AGLS $F$-statistic (Yuan & Bentler, 1999). In

addition, it provides corrected standard error estimates (Yuan & Bentler, 1997a) for small samples.

When categorical variables are declared in LISREL 8.8 (PRELIS), Mplus 6, or EQS 6.1, all three programs estimate thresholds for categorical variables, polychoric/polyserial correlations among observed variables, and an asymptotic covariance matrix before estimating structural model parameters. The programs differ in the ways these matrices are computed. LISREL 8.8 estimates polychoric/polyserial correlations using Olsson's (1979) procedure. EQS 6.1 uses a partition maximum likelihood approach (PML; Lee, Poon, & Bentler, 1995; Poon & Lee, 1987), while Mplus 6 employs the limited-information likelihood approach of Muthén (1984). The default estimator for categorical variables is WLSMV in Mplus 6; robust ML with robust statistics based on the Lee and colleagues' (1995) optimal weight matrix is the default estimator in EQS 6.1. LISREL 8.8 provides a robust ML solution with Satorra–Bentler scaled statistics by default when a polychoric matrix and asymptotic covariance matrix calculated by PRELIS are submitted as input. Instead of using ML or ADF estimation, Amos 18 only allows Bayesian estimation when non-numeric data (i.e., ordered categorical) are declared.

Bayesian estimation is currently available only in Mplus 6 and Amos 18. Both programs use MCMC as the sampling algorithm. To request Bayesian estimation in Mplus 6, "ESTIMATOR = BAYES" should be specified. Mplus 6 allows different types of Bayes point estimates; users can request mean, median, or mode of the posterior distribution by specifying "POINT = MEAN," "POINT = MEDIAN," or "POINT = MODE," respectively. By default, Amos 18 provides the posterior mean and median for point estimates. The statements used by the four software programs to request the different estimation methods discussed above are summarized in Appendix 10.1.

## EMPIRICAL FINDINGS ABOUT DIFFERENT ESTIMATION METHODS

Although Equation 10.2 can be considered the general fit function for normal theory ML, normal theory GLS, and Browne's ADF or WLS, these estimators are not necessarily equivalent. When multivariate normality of observed variables does not hold or sample size is small, $F_{WLS}$ will not be equivalent to $F_{GLS}$ or $F_{ML}$ (Yuan & Chan, 2005). Unfortunately, multivariate normality of observed variables, assumed by ML and GLS, is rarely satisfied (e.g., Micceri, 1989), and sample sizes required by the theoretically correct ADF estimator under non-normality are often unavailable in practice. It is unlikely that these estimators will be equivalent in realistic conditions. The choice of estimation method becomes essential because it will affect evaluation of model fit and parameter estimates (Bentler & Dudgeon, 1996). Users have to rely on empirical findings about different estimators in various practical conditions to inform their choice in applications of SEM.

### Estimators for Continuous Variables

When continuous observed variables are not normally distributed, robust ML estimators (i.e., ML with Satorra–Bentler scaled statistics) appear to work better than Browne's ADF estimator, especially when sample sizes are not large (e.g., Chou et al., 1991; Curran, West, & Finch, 1996; Hu et al., 1992). Sample sizes required for ADF to work satisfactorily increase for larger models, increasingly non-normal distribution of observed variables, or both. The required sample size is often unrealistic for many social and behavioral fields. Hu and colleagues (1992), for example, found that the sample size required for ADF to work well for a 15-variable, three-factor CFA model was 5,000 under a symmetrical but highly kurtotic distribution. The Satorra–Bentler scaled $\chi^2$, however, has been shown repeatedly to work fairly well relative to the ML and ADF $\chi^2$ across various levels of non-normality and sample size (e.g., Chou et al., 1991; Curran et al., 1996; Hu et al., 1992).

Among the modified model fit test statistics for ML, Yuan and Bentler (1998) found that both the Satorra–Bentler scaled $\chi^2$ and the Yuan–Bentler residual-based $\chi^2$ performed well under a variety of conditions at sample sizes of 200 or above. When sample size was 120 or smaller, however, Bentler and Yuan (1999) found the Type I error rate of the Satorra–Bentler scaled $\chi^2$ to be inflated two to four times the nominal level, while that of the Yuan–Bentler residual-based $\chi^2$ was too conservative (underrejected correct models). The Yuan–Bentler residual-based $F$-test was found to perform better than the Satorra–Bentler scaled $\chi^2$ test, Browne's residual test, and the Yuan–Bentler residual-based $\chi^2$ test under various levels of non-normality and very small sample conditions ($N \leq p(p + 1)/2$ or $N{:}t \leq 3.64{:}1$, where $t$ is the number of estimated model parameters; Bentler & Yuan, 1999). Because the theoretical null distribution of the Satorra–Bentler scaled statistic is unknown,

Bentler and Yuan recommend using the Yuan–Bentler residual-based $\chi^2$ or the Yuan–Bentler residual-based $F$-test for non-normal data when the sample size is medium to large [$N > p(p + 1)/2$]. The Yuan–Bentler residual-based $F$-test is recommended when the sample size is very small [$df \leq N \leq p(p + 1)/2$] regardless of sampling distribution (Bentler & Yuan, 1999).

Fouladi (1999) suggested that small-sample performance of Satorra and Bentler's (1988, 1994) test statistics could be improved by incorporating a Bartlett (1950) correction. In a large simulation study, Nevitt and Hancock (2004) examined the performance of a number of modified model fit test statistics for non-normal data under various ratios of sample size to number of estimated model parameters ($N$:$t$ ranged from 1:1 to 10:1, with $N$'s ranging from 35 to 1,290) and sampling distributions (skewness ≤ 3, kurtosis ≤ 21). The authors found that the Satorra–Bentler scaled statistic exhibited inflated Type I error rate at small sample sizes ($N$:$t$ ≤ 10:1) but the Bartlett-corrected version provided good Type I error control and superior power compared to other test statistics in nearly all sample size and distribution conditions they investigated. However, the Satorra–Bentler adjusted statistic and its Bartlett-corrected version tended to provide low rejection rates for both correctly and incorrectly specified models (Nevitt & Hancock, 2004). Regarding the residual-based test statistics, the authors found the performance of the Yuan–Bentler $\chi^2$ and $F$-test statistics to be "erratic, controlling Type I error rates under some conditions and failing under others" (Nevitt & Hancock, 2004, p. 468). As a result, the authors recommended the Bartlett-corrected Satorra–Bentler scaled statistic for evaluating model fit with small samples.

Besides model fit test statistics ($\chi^2$ or $F$-test), other fit indices that are functionally related to model $\chi^2$ are often used to evaluate model fit, and standard error estimates are used to test parameter estimates for statistical significance. Yu and Muthén (2002) showed that model fit indices (root mean square error of approximation [RMSEA], Tucker–Lewis index [TLI], and comparative fit index [CFI]) based on Satorra–Bentler scaled $\chi^2$ performed better than those based on regular ML and ADF under moderate to severe non-normality and small sample size conditions. Nevitt and Hancock (2000) also reported improved performance of adjusted RMSEA based on Satorra–Bentler scaled $\chi^2$ when models were properly specified for non-normal data. Similarly, robust standard errors of Satorra and Bentler (1988, 1994) have been found to show less negative bias than those of regular ML and ADF (Chou & Bentler, 1995; Chou et al., 1991).

Compared to the robust ML approach for continuous non-normal data, Nevitt and Hancock (2001) found that the bootstrapping approach (with Amos) did not work well with small samples ($N$ = 100 or $N$:$t$ = 5:1). When sample size was at least 200, bootstrapping provided better Type I error control but lower power than did the Satorra–Bentler scaled statistic under severe non-normality (skewness = 3, kurtosis = 21). Moreover, bootstrapped standard errors were somewhat less biased but more variable than Satorra–Bentler robust standard errors when sample size was not small (Nevitt & Hancock, 2001).

## Estimators for Ordered Categorical Variables

When approximately normal ordinal variables with at least five categories are treated as continuous and analyzed with normal theory ML, research has shown that model $\chi^2$ and fit indices are not greatly misleading, but parameter estimates are slightly underestimated, and standard error estimates are negatively biased (e.g., Babakus, Ferguson, & Jöreskog, 1987; Muthén & Kaplan, 1985). When non-normal ordinal variables are treated as continuous and analyzed with normal theory ML, model–data fit is underestimated (e.g., Green, Akey, Fleming, Hershberger, & Marquis, 1997; West, Finch, & Curran, 1995), and negative bias in parameter and standard error estimates is more severe (e.g., Babakus et al., 1987; Muthén & Kaplan, 1985). The performance of the normal theory ML estimator is worse as the number of categories decreases, level of non-normality increases, and sample size decreases (e.g., Babakus et al., 1987; Bollen, 1989; Dolan, 1994).

Because analyzing a polychoric/polyserial matrix leads to consistent estimators of parameters with any fit functions (Bollen, 1989), alternative estimation methods for categorical data in SEM usually begin with estimating polychoric/polyserial correlations, assuming underlying normality of latent response variables. Dolan (1994) found that ML with a polychoric matrix provided better parameter estimates than ML with Pearson matrix for symmetrically distributed variables with up to seven categories. Coenders, Satorra, and Saris (1997) also found that analyzing a polychoric matrix worked better than analyzing a Pearson matrix when normality of latent response variables held, even though intervals between categories of ordinal vari-

ables were unequal (the authors called this "high transformation error," p. 273).

However, a polychoric or polyserial correlation may provide a biased estimate of the latent association when normality of latent response variables is violated or when the cell size of the bivariate frequency table is sparse (e.g., Olsson, 1979). Although researchers have found polychoric/polyserial correlations to be fairly robust to moderate non-normality of latent response variables ($|\text{skewness}| \leq 1.25$ and $|\text{kurtosis}| \leq 3.75$; e.g., Coenders et al., 1997; Flora & Curran, 2004), they are not robust to severe non-normality (e.g., skewness = 5, kurtosis = 50; Flora & Curran, 2004).

Among estimation methods used on polychoric/polyserial matrices, WLSMV has been found to perform better than the full WLS (e.g., Flora & Curran, 2004, who used Mplus for WLS; Berkovits & Hancock, 2000, who used EQS for WLS). Moreover, Muthén and colleagues (1997) found that WLSMV outperforms WLSM in Type I error control. However, WLSMV appeared to be slightly more sensitive to small samples than robust ML with Satorra–Bentler scaled statistics (Lei, 2009). Lei (2009) found that when sample size was at least 250 ($N{:}t > 10{:}1$), robust ML and WLSMV performed similarly regardless of level of score skewness. When sample size was 100 ($N{:}t < 10{:}1$) and ordinal variables were moderately skewed (skewness = 2.3, kurtosis = 5.3), however, WLSMV provided a slightly higher percentage of invalid rejections, a lower percentage of valid rejections, and more negatively biased standard error estimates than did ML with Satorra–Bentler scaled statistics. Lei also found the Satorra–Bentler scaled $\chi^2$ test to be generally more powerful than the Yuan–Bentler residual-based $F$-test (at $N > p(p + 1)/2$), although they provided similar Type I error rates.

## SUMMARY OF EMPIRICAL FINDINGS

In summary, when simultaneously considering both model test statistics, and parameter and standard error estimates, the literature suggests that normal theory ML with the Satorra and Bentler (1988, 1994) scaled statistics would work better than ML without scaling for continuous variables when normality of variables is violated; it would also work better than the ADF estimator when sample size is not very large. Among methods for improving the model fit test statistic, the Satorra–Bentler scaled $\chi^2$ appears to perform comparably to the Yuan–Bentler residual-based tests and bootstrapping when sample size is not too small ($N \geq 200$ or $N{:}t \geq 10{:}1$). When sample size is very small ($N < 200$ or $N{:}t < 10{:}1$), incorporating a Bartlett correction to the Satorra–Bentler scaled $\chi^2$ statistic may improve the accuracy of model evaluation, and it appears to work more consistently than the Yuan–Bentler residual-based $F$-test.

For categorical observed variables, polychoric/polyserial correlation is recommended for analysis unless normality of the underlying scales is severely violated (e.g., skewness = 5, kurtosis = 50; examined by Flora & Curran, 2004) or the expected bivariate frequencies are small (Olsson, 1979). WLSMV or robust ML with a Satorra–Bentler scaled statistic on polychoric/polyserial correlations would work better than full WLS for ordinal variables at realistic sample sizes. The robust ML approach appears to perform similarly to WLSMV when sample size is not too small (e.g., $N{:}t \geq 10{:}1$) but may perform better than WLSMV when the sample size is small (e.g., $N = 100$ or $N{:}t < 10{:}1$). In general, full WLS estimators are not recommended for either continuous or ordinal variables, unless the sample size is extremely large.

## AN ILLUSTRATION

To illustrate the use and performance of various estimators for non-normal variables, we created several simulated data sets with the desired data characteristics (continuous vs. ordinal non-normal data, and small vs. large sample sizes). A simple 18-variable, two-factor CFA model with simple structure is used for illustration purposes. Population parameters, factor loadings, and correlation between the two factors were taken from estimates of a large, actual data set. The two constructs of interest are generalized anxiety (measured with eight items) and depression (measured with 10 items). Original item responses were on a 5-point scale. Respondents were asked to rate brief descriptions of symptoms (e.g., "I have sleep difficulties," "I feel helpless") from 0 (*not at all like me*) to 4 (*extremely like me*).

The number of unique sample variances and covariances for 18 observed variables is 171 [i.e., 18 * (18 + 1)/2]. The 18-variable, two-factor model has 134 *df*, with 18 factor loadings, 18 error variances, and an interfactor correlation (factor variances were fixed to 1 to scale the latent variables), resulting in a total of 37 parameters to be estimated. Two levels of sample size, 200 and 1,000, were included. Both sample sizes were

greater than the number of unique sample variances and covariances, but the ratio of sample size to number of estimated parameters was less than 10 to 1 for a sample size of 200 (about 5.4:1) and greater than 10 to 1 for sample size of 1,000 (about 27:1). A sample size of 1,000 is expected to work well for robust ML with Satorra–Bentler corrected statistics or with a Yuan–Bentler residual-based $\chi^2$ or $F$-test when observed variables are continuous non-normal, and for robust ML and WLSMV analysis with polychoric matrix when observed variables are ordinal non-normal. A sample size of 200 (less than 10 observations per estimated parameter), however, may pose a challenge for these estimators. Nevertheless, a sample size of 1,000 might still be too small for normal theory ML and full WLS $\chi^2$ to behave well when observed variables are non-normal in either continuous or ordinal form.

A large level of non-normality (all indicators had univariate skewness = 3, kurtosis = 21) was simulated for continuous observed variables to challenge the robust ML approach. For ordinal variables, either *non-normality* of latent response variables (univariate skewness = 3, kurtosis = 21) with low transformation error (i.e., equal interval between categories with cut points for categorization at –.75, –.25, .25, and .75) or *multivariate normal* latent response variables with high transformation error (unequal interval and asymmetric categorization at cut points .67, 1.28, 1.645, and 2.05) was simulated. The former resulted in univariate negative skewness (average = –.64) and negative kurtosis (average = –.46) of the observed ordinal variables and the latter univariate positive skewness (average = 2.34) and positive kurtosis (average = 5.16). If normality of latent response variables were important for using a polychoric matrix, then parameter estimates produced by analyzing a polychoric matrix (holding estimators constant) would be expected to be worse in the first case than in the second case despite the more severe level of non-normality of the observed ordinal variables in the second case.

The normal theory ML estimator, the full WLS (or ADF) estimator, and the MLM with Satorra–Bentler scaled $\chi^2$ and fit indices based on scaled statistics were used to estimate the model with Mplus 6 for continuous observed variables. Residual-based model test statistics based on ML were obtained from EQS 6.1 by fitting the same model to the same continuous data sets. For ordinal observed variables, solutions from normal theory ML and MLM were requested from Mplus 6 by treating the 5-point ordinal variables as continuous to see if ignoring the metric problem was robust to source (from non-normal latent response variables or transformation error) or level (small skewness and kurtosis vs. moderate skewness and kurtosis) of observed non-normality. A polychoric matrix of ordinal variables (PC) was analyzed with ML (referred to as ML + PC in Table 10.3) and robust ML (labeled as MLM + PC in Table 10.3) using EQS 6.1, as well as with full WLS and WLSMV using Mplus 6.

Model fit statistics and bias of parameter estimates (Average % bias and Average % absolute bias of model parameters) based on one replication for the various combinations of non-normality, sample size, and estimation methods are provided in Table 10.3. Bias of standard error estimates could not be evaluated because only one replication for each cell was simulated for illustration purposes.

As can be seen in Table 10.3, all ML $\chi^2$ without adjustment rejected the true model incorrectly, as did some fit indices based on unadjusted ML $\chi^2$ at the small sample size. As expected, Satorra–Bentler scaled $\chi^2$ and fit indices based on the scaled $\chi^2$ performed well for continuous variables despite the large departure from normality. The Yuan–Bentler residual-based $\chi^2$ and $F$-test also performed well at both sample sizes for continuous non-normal variables, while Browne's residual-based $\chi^2$ did not do well in the small-sample condition. Furthermore, average % bias and average % absolute bias of the ML parameter estimates across model parameters were small for continuous non-normal data. The performance of the robust ML approach with adjusted model test statistics (with Satorra–Bentler scaled $\chi^2$ or Yuan–Bentler $\chi^2$ or $F$-test) appeared to hold up well under continuous non-normality even in the small-sample condition ($N$ = 200), with the ratio of sample size to number of estimated parameters being just over 5 to 1. Consistent with findings from the literature, the theoretically correct estimator under non-normality, full WLS or ADF, performed worse than the robust ML method in estimation of both model fit and model parameters at either sample size.

For ordinal variables, the performance of the estimation methods in parameter estimation appeared to depend on the source or magnitude of non-normality. Treating ordinal variables as continuous and estimating the model with normal theory ML produced a substantial amount of negative bias (about 14% on average) in parameter estimates, regardless of sample size, when transformation error was high (asymmetrical categorization with unequal intervals between adjacent cat-

**TABLE 10.3. Fit Statistics for Selected Estimation Methods under Different Non-Normal Variable and Sample Size Conditions**

| | $\chi^2$ ($df$ = 134) | $p$ | CFI | TLI | RMSEA | Average % bias of parameter estimates | Average % absolute bias of parameter estimates |
|---|---|---|---|---|---|---|---|
| **Severely non-normal continuous variables** | | | | | | | |
| $N$ = 200 | | | | | | | |
| ML | 337.52 | **.000** | **.911** | **.898** | **.087** | .08 | 5.88 |
| MLM[a] | 149.35 | .173 | .982 | .980 | .024 | — | — |
| WLS | 398.44 | **.000** | **.797** | **.769** | **.099** | 5.45 | 9.65 |
| $N$ = 1,000 | | | | | | | |
| ML | 395.39 | **.000** | .975 | .972 | .044 | .17 | 1.69 |
| MLM[b] | 150.55 | .156 | .995 | .995 | .011 | — | — |
| WLS | 149.46 | .171 | **.949** | **.942** | .011 | –7.18 | **10.21** |
| **Ordinal variables with moderate observed non-normality (MVN latent response variables + high transformation error)** | | | | | | | |
| $N$ = 200 | | | | | | | |
| ML | 197.12 | **.000** | **.948** | **.941** | .049 | **–14.11** | **14.11** |
| MLM | 115.99 | .867 | 1 | 1 | .000 | — | — |
| ML+PC | 478.67 | **.000** | **.854** | **.834** | NA | .06 | 4.60 |
| MLM+PC[c] | 130.74 | .564 | 1 | 1 | NA | — | — |
| WLSMV | 139.53 | .354 | .997 | .997 | .014 | .70 | 4.40 |
| WLS | NA | NA | NA | NA | NA | NA | NA |
| $N$ = 1,000 | | | | | | | |
| ML | 238.39 | **.000** | .983 | .980 | .028 | **–13.74** | **13.74** |
| MLM | 122.33 | .756 | 1 | 1 | .000 | — | — |
| ML+PC | 497.79 | **.000** | .966 | .961 | NA | .40 | 2.26 |
| MLM+PC[b] | 135.28 | .453 | 1 | 1 | NA | — | — |
| WLSMV | 128.38 | .621 | 1 | 1 | .000 | .59 | 2.36 |
| WLS | 149.27 | .174 | .996 | .995 | .011 | 5.09 | 6.32 |
| **Ordinal variables with mild observed non-normality (severely non-normal latent response variables + low transformation error)** | | | | | | | |
| $N$ = 200 | | | | | | | |
| ML | 177.34 | **.007** | .980 | .977 | .040 | 1.67 | 3.30 |
| MLM | 164.97 | **.036** | .985 | .983 | .034 | — | — |
| ML+PC | 276.52 | **.000** | **.949** | **.942** | NA | 7.49 | 7.90 |
| MLM+PC[a] | 158.69 | .072 | .994 | .993 | NA | — | — |
| WLSMV | 165.68 | **.033** | .994 | .993 | .034 | 8.13 | 8.23 |
| WLS | NA | NA | NA | NA | NA | NA | NA |
| $N$ = 1,000 | | | | | | | |
| ML | 163.22 | **.044** | .997 | .997 | .015 | 1.84 | 2.77 |
| MLM | 145.59 | .233 | .999 | .999 | .009 | — | — |
| ML+PC | 273.94 | **.000** | .990 | .988 | NA | 7.62 | 7.62 |
| MLM+PC[b] | 136.31 | .428 | 1 | 1 | NA | — | — |
| WLSMV | 154.32 | .111 | .999 | .999 | .012 | 7.67 | 7.67 |
| WLS | 182.35 | **.004** | .994 | .994 | .019 | **11.61** | **11.61** |

*Note.* NA, not available; — indicates same as above. **Bolded** fit statistics indicate rejection of the model at their respective recommended criteria (i.e., $p < .05$, CFI < .95, TLI < .95, RMSEA > .06); **bolded** average bias and average absolute bias of parameter estimates are > 10%, a level considered unacceptable by Muthén et al. (1997).

[a]Yuan–Bentler $\chi^2$ and $F$-test were not significant, but Browne's residual-based $\chi^2$ was at $p < .05$.

[b]All residual-based test statistics were not significant at $p > .05$.

[c] All residual-based test statistics were significant at $p < .05$.

egories) or when observed skewness and kurtosis were large (2.34 and 5.16, respectively). However, the same approach produced fairly unbiased parameter estimates (< 5% on average) when transformation error was low (symmetrical categorization with equal intervals between adjacent categories on severely non-normal latent response variables) or when observed skewness and kurtosis were low (–.64 and –.46, respectively), regardless of sample size. Note that levels of transformation error and of observed non-normality were confounded in this example. That is, high transformation error happened to result in large observed non-normality, and low transformation error happened to result in a low level of observed non-normality. Analyzing PC instead of a Pearson matrix with ML or WLSMV produced acceptable parameter estimates with a small amount of bias (< 10% on average) in both levels of observed non-normality. However, the amount of bias was notably lower when normality of latent response variables held, despite the larger level of observed skewness and kurtosis (average absolute bias < 5%) than when normality of latent response variables was violated (average absolute bias around 8%). Normality of latent response variables appeared to be important when PC was analyzed. More severe departure from normality than the magnitude simulated here may bias parameter estimates to an unacceptable level.

Holding sample size and source of non-normality for observed ordinal variables constant, robust ML analyzing a polychoric matrix with Satorra–Bentler scaled statistics performed similarly to WLSMV in their estimation of model fit and model parameters, both yielding mostly accurate fit statistics about the true model and an acceptable amount of bias in parameter estimates. As expected, full WLS was inferior to WLSMV in all cases. When the sample size was small, full WLS failed to yield a solution; when it did produce a solution at the larger sample size, its parameter estimates were more biased than WLSMV estimates. A sample size of 1,000 did not appear to be large enough for full WLS to outperform WLSMV.

Regardless of the type of matrix analyzed, for ordinal variables or sample size (200 or 1,000), the Satorra–Bentler scaling appeared to successfully reduce the inflation in model $\chi^2$ of the ML estimator and would correctly retain the true model at the .05 significance level in most cases by itself. Moreover, the performance of Satorra–Bentler scaled $\chi^2$ for ML with PC was more consistent than that of the residual-based test statistics.

Although no definitive conclusions can be made from this single-replication simulation, it illustrates the importance of choosing appropriate estimation methods based on considerations of data characteristics, including sample size or sample size per estimated model parameter, metric of measured variables, and distribution of variables. Our simulation results showed that the default normal theory ML estimator in most standard SEM programs did not work well when observed variables were non-normally distributed; it misinformed model–data fit regardless of the metric of the observed variables and underestimated model parameters for ordinal variables with high transformation error or moderate level of observed non-normality. The ML estimator with Satorra–Bentler scaled statistic performed much better in informing model–data fit than the regular ML model fit statistic regardless of the form or magnitude of non-normality and sample size. Bias of parameter estimates provided by the ML estimator was small under continuous non-normality but large when the metric of ordinal variables with high transformation error (or moderate observed non-normality) was disregarded.

The metric of ordinal variables can be taken into account by analyzing a polychoric/polyserial matrix. Robust ML or WLSMV analysis of a polychoric matrix performed well in informing model–data fit and provided parameter estimates with a tolerably small amount of bias, albeit bias of their parameter estimates might be somewhat higher when normality of latent response variables was severely violated (skewness = 3, kurtosis = 21). Moreover, the theoretically correct WLS or ADF estimator for non-normal data did not perform well at realistic sample sizes compared to robust ML or WLSMV, as has been shown in previous studies.

In conclusion, results of this simple simulation appear to be consistent with findings reported in other studies. It demonstrates differential performance of different estimation methods that have been proposed, recommended at some point, or used in applied research under some common practical conditions. It elucidates the importance of choosing estimation methods based on careful considerations of data characteristics such as sample size or sample size per estimated model parameter, metric of measured variables, and distribution of variables. These considerations are imperative because in real-life application in areas of social and behavioral sciences, non-normality of observed variables is common (e.g., Micceri, 1989) and sample size is typically not large.



## REFERENCES

Anderson, J. C
modeling i
step approa
Anderson, T.
of covarian
*Statistics, I*
Arbuckle, J. L
fordville, Fl
Arbuckle, J. L
cago: Small
Asparouhov, T
*ate statistic*
at the 2005
(FCSM) Re
Babakus, E., I
The sensiti
tor analysis
tributional
*24*, 222–22
Bargmann, R.
*dence in mu*
186). Chap
of Statistics
Bartlett, M. S
sis. *British*
77–85.
Basmann, R.
linear estin
*Econometr*
Bauwens, L. (
*multaneous*
*Carlo*. New
Bentler, P. M.
ware]. Enci
Bentler, P. M.
analysis: S
*nual Revie*
Bentler, P. M.
modeling w
*Behavioral*
Berkovits, I., &
*of methods*
*continuous*
ing of the
New Orlea
Bielby, W., &
*Annual Re*
Blalock, H. M
*mental res*
Carolina.
Bollen, K. A.
*ables*. New
Bollen, K. A.

## REFERENCES


Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411–423.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Annals of Statistics, 1*, 135–141.

Arbuckle, J. L. (1995–2009). *Amos 18 user's guide*. Crawfordville, FL: Amos Development Corporation.

Arbuckle, J. L. (2009). *Amos 18* [Computer software]. Chicago: SmallWaters.

Asparouhov, T., & Muthén, B. (2005, November). *Multivariate statistical modeling with survey data*. Paper presented at the 2005 Federal Committee on Statistical Methodology (FCSM) Research Conference, Arlington, VA.

Babakus, E., Ferguson, C. E., Jr., & 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.

Bargmann, R. (1957). *A study of independence and dependence in multivariate normal analysis* (Mimeo Series No. 186). Chapel Hill: University of North Carolina, Institute of Statistics.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section, 3*, 77–85.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica, 25*, 77–83.

Bauwens, L. (1984). *Bayesian full information analysis of simultaneous equation models using integration by Monte Carlo*. New York: Springer-Verlag.

Bentler, P. M. (2005). *EQS 6.1 for Windows* [Computer software]. Encino, CA: Multivariate Software.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*, 563–592.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34*, 181–197.

Berkovits, I., & Hancock, G. R. (2000, April). *A comparison of methods for structural modeling with polytomous and continuous variables*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bielby, W., & Hauser, R. (1977). Structural equation models. *Annual Review of Sociology, 3*, 137–161.

Blalock, H. M., Jr. (1964). *Causal inferences in nonexperimental research*. Chapel Hill: University Press of North Carolina.

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

Bollen, K. A. (1996a). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika, 61*, 109–121.

Bollen, K. A. (1996b). A limited-information estimator for LISREL models with and without heteroscedastic errors. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 227–241). Mahwah, NJ: Erlbaum.

Bollen, K. A. (2001). Two-stage least squares and latent variable models: Simultaneous estimation and robustness to misspecifications. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 119–138). Lincolnwood, IL: Scientific Software.

Bollen, K. A., & Bauer, D. J. (2004). Automating the selection of model-implied instrumental variables. *Sociological Methods and Research, 32*, 425–452.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research, 36*, 48–86.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8*, 1–24. Reprinted in 1977 in D. J. Aigner & A. S. Goldberger (Eds.), *Latent variables in socioeconomic models* (pp. 205–226). Amsterdam: North Holland.

Browne, M. W. (1984). Asymptotically distribution-free methods for analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Chou, C., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–54). Thousand Oaks, CA: Sage.

Chou, C., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347–357.

Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling, 4*, 261–282.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29.

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.

Duncan, O. (1966). Path analysis: Sociological examples. *American Journal of Sociology, 72*, 1–16.

Dunson, D. B., Palomo, J., & Bollen, K. A. (2005). *Bayesian structural equation modeling* (Technical Report No.

2005-5). Research Triangle Park, NC: Statistical and Applied Mathematical Sciences Institute.

Emmett, W. G. (1949). Factor analysis by Lawley's method of maximum likelihood. *British Journal of Mathematical and Statistical Psychology, 2*, 90–97.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 313–342). Greenwich, CT: Information Age.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and non-normality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4*, 87–107.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika, 44*, 409–420.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491.

Fouladi, R. T. (1999, April). *Model fit in covariance structure analysis under small sample conditions—Modified maximum likelihood and asymptotically distribution free generalized least squares procedures.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling, 7*, 356–410.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In *Markov Chain Monte Carlo in practice* (pp. 1–19). London: Chapman & Hall.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4*, 108–120.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329–367.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351–362.

Jöreskog, K. G. (1977). Factor analysis by least square and maximum likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (Vol. III, pp. 125–165). New York: Wiley.

Jöreskog, K. G. (1983). Factor analysis as an error-in-variables model. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 185–196). Hillsdale, NJ: Erlbaum.

Jöreskog, K. G., & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika, 37*, 243–260.

Jöreskog, K. G., & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology, 21*, 85–96.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7* [Computer software]. Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8* [Computer software]. Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software.

Jöreskog, K. G., & van Thillo, M. (1973). *LISREL–A general computer program for estimating a linear structural equation system involving unmeasured variables* (Research Report No. 73-5). Uppsala, Sweden: University of Uppsala.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh* (A), *60*, 64–82.

Lee, S.-Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44*, 99–113.

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.

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity, 43*, 495–507.

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika, 40*, 505–517.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.

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

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Muthén, B. O., & 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. O., & Muthén, L. K. (2010). *Mplus 6* [Computer software]. Los Angeles: Authors.

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 techni-

cal report. Retrieved from *http://gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf.*

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Authors.

Nevitt, J., & Hancock, G. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education, 68*, 251–268.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*, 353–377.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*, 439–478.

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

Poon, W.-Y., & Lee, S.-Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika, 52*, 409–430.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the Business and Economic Section of the American Statistical Association* (pp. 308–313). Alexandria, VA: American Statistical Association.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–87). New York: Wiley.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yuan, K.-H., & Bentler, P. M. (1997a). Improving parameter tests in covariance structure analysis. *Computational Statistics and Data Analysis, 26*, 177–198.

Yuan, K.-H., & Bentler, P. M. (1997b). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*, 767–774.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289–309.

Yuan, K.-H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 24*, 225–243.

Yuan, K.-H., & Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika, 70*, 791–798.