Finney, S.J. & DiStefano, C. (2006). *Non-normal and Categorical data in structural equation modeling. In G. r. Hancock & R. O. Mueller (Hrsg.).* Structural equation modeling: a second course (S. 269–314). Greenwich, Connecticut: Information Age Publishing

CHAPTER 9

NON-NORMAL AND CATEGORICAL DATA IN STRUCTURAL EQUATION MODELING

Sara J. Finney and Christine DiStefano

Structural equation modeling (SEM) has become an extremely popular data analytic technique in education, psychology, business, and other disciplines (Austin & Calderón, 1996; MacCallum & Austin, 2000; Tremblay & Gardner, 1996). Given the frequency of its use, it is important to recognize the assumptions associated with different estimation methods, demonstrate the conditions under which results are robust to violations of these assumptions, and specify the procedures that should be employed when assumptions are not met. The importance of attending to assumptions and, consequently, selecting appropriate analysis strategies based on the characteristics of the data and the study's design cannot be understated. Put simply, violating assumptions can produce biased results in terms of model fit as well as parameter estimates and their associated significance tests. Biased results may, in turn, result in incorrect decisions about the theory being tested.

While there are several assumptions underlying the popular normal theory (NT) estimators used in SEM, the two assumptions that we focus

Structural Equation Modeling: A Second Course, 269–314

on in this chapter concern the metric and distribution of the data. Specifically, the data required by NT estimators are assumed to be continuous and multivariate normally distributed in the population. We focus on these two assumptions because often the data modeled in the social sciences do not follow a multivariate normal distribution. For example, Micceri (1989) noted that much of the data gathered from achievement and other measures are not normally distributed. This is disconcerting given that Gierl and Mulvenon (1995) found that most researchers do not examine the distribution of their data, but instead simply assume normality. In addition to the pervasiveness of non-normal data, the applied literature is thick with examples of categorical data collected using ordinal measures (e.g., Likert-type scales).

Because of the prevalence of both non-normal and categorical data in empirical research, this chapter focuses on issues surrounding modeling data with these characteristics using SEM. First, we review the assumptions underlying NT estimators. We next describe non-normal and categorical data and review robustness studies of the most popular NT estimator, maximum likelihood (ML), in order to understand the consequences of violating these assumptions. We then discuss four popular strategies that have been used to accommodate non-normal and/or categorical data:

1. Asymptotically distribution-free (ADF) estimation
2. Satorra-Bentler scaled χ^2 and standard errors
3. Robust weighted least squares (WLS) estimation methods implemented in the software program Mplus (e.g, WLSM, WLSMV)
4. Bootstrapping.

For each strategy we present the following: (a) a description of the strategy; (b) a summary of research concerning the robustness of the χ^2 statistic, fit indices, parameter estimates, and standard errors; and (c) a description of implementation across three software programs.

NORMAL THEORY ESTIMATORS

Assumptions of Normal Theory Estimators

As with most statistical techniques, SEM is based on assumptions that should be met in order for researchers to trust the obtained results. Central to SEM is the choice of an estimation method that is used to obtain parameter values, standard errors, and fit indices. The two common NT

estimators are maximum likelihood (ML) and generalized least squares (GLS), and require the following set of assumptions (e.g., Bentler & Dudgeon, 1996; Bollen, 1989):

- Independent observations: Observations for different subjects are independent. This can be achieved through simple random sampling.
- Large sample size: All statistics estimated in SEM are based on an assumption that the sample is sufficiently large.
- Correctly specified model is estimated: The model being estimated reflects the true structure in the population.
- Multivariate normal data: The observed scores have a (conditionally) multivariate normal distribution.
- Continuous data: The assumption of multivariate normality implies that the data are continuous in nature. Categorical data, such as dichotomies or even Likert-type data, cannot by definition be normally distributed because they are discrete in nature (Kaplan, 2000). Therefore, it is often noted that NT estimators require *continuous normally distributed* endogenous variables.

If NT estimators are applied when the above conditions are satisfied, the parameter estimates have three desirable properties: asymptotic unbiasedness (they neither over- nor underestimate the true population parameters in large samples), asymptotic efficiency (variability of the parameter estimate is at a minimum in large samples), and consistency (parameter estimates converge to population parameters as sample size increases).

Defining Normal Theory Estimators

For both ML and GLS estimation methods, model parameters are estimated using an iterative process. The final set of parameters minimizes the discrepancy between the observed sample covariance matrix (**S**) and the model-implied covariance matrix calculated from the estimated model parameters [$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$]. The fit function that is minimized, $F = F[\mathbf{S}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]$, will equal zero if the model perfectly predicts the elements in the sample covariance matrix. If the assumptions noted above are met, the overall fit between the model and the data can be expressed as $T = F(N - 1)$, which follows a central χ^2 distribution.

The fit function for both ML and GLS estimators can be written in the same general form:

$$F = \frac{1}{2} tr[([\mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]\mathbf{W}^{-1})^2] \quad (1)$$

where *tr* is the trace of a matrix (i.e., the sum of the diagonal elements), and $[\mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]$ represents the discrepancy between the elements in the sample covariance matrix and the elements in the model-implied covariance matrix. These residuals $[\mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]$ are weighted by a weight matrix, **W**. The weight matrix differs between the two NT procedures; GLS employs the observed sample covariance matrix **S** as the weight matrix, whereas ML employs the model-implied covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.[1] If all assumptions are met, the two weight matrices will be equivalent at the last iteration and the estimators will produce convergent results (Olsson, Troye, & Howell, 1999). However, if the model is misspecified, **W** at the last iteration will differ between the two techniques, even if all other assumptions are met. This difference in **W** results in different parameter estimates and fit indices across the estimators. Specifically, GLS has been found to produce overly optimistic fit indices and more biased parameter estimates than ML if the estimated model is misspecified. Seeing that most applied researchers are interested in the plausibility of a specified model and would, therefore, prefer fit indices sensitive to model misspecification, ML has been recommended over GLS (Olsson et al., 1999; Olsson, Foss, Troye, & Howell, 2000). We, therefore, limit subsequent discussion of NT estimators to ML.

NON-NORMAL DATA

Assessing Non-Normality

In general, the effects of non-normality on ML-based results depend on its extent; the greater the non-normality, the greater the impact on results. Therefore, researchers should assess the distribution of the observed variables prior to analyses in order to make an informed decision concerning estimation method. Three indices of non-normality are typically used to evaluate the distribution: univariate skew, univariate kurtosis, and multivariate kurtosis. Unfortunately, there is no clear consensus regarding an "acceptable" degree of non-normality. Studies examining the impact of univariate normality on ML-based results suggest that problems may occur when univariate skewness and univariate kurtosis approach values of 2 and 7, respectively (e.g., Chou & Bentler, 1995; Curran, West, & Finch, 1996; Muthén & Kaplan, 1985). In addition, there is no generally accepted cutoff value of multivariate kurtosis that indicates non-normality. A guideline offered through the EQS software program

(Bentler, 2004) suggests that data associated with a value of Mardia's normalized multivariate kurtosis (see Bollen, 1989, p. 424, equation 4, for formula) greater than 3 could produce inaccurate results when used with ML estimation (Bentler & Wu, 2002). This guideline is consistent with discussions by many applied and methodological researchers regarding this issue found on SEMNET (structural equation modeling listserv). Future research should investigate the utility of such a cutoff value and the conditions under which it is relevant (e.g., size of model).

Effects of Analyzing Non-Normal Continuous Data: Empirical Results

Given the abundance of non-normal and categorical data analyzed in the social sciences, a question of significant interest concerns the robustness of ML to these conditions. Research examining the effects of non-normality has typically focused on (a) the χ^2 statistic, (b) other model fit indices, (c) parameter estimates, and (d) standard errors. As detailed below, ML has been found to produce relatively accurate parameter estimates under conditions of non-normality (e.g., Finch, West, & MacKinnon, 1997); however, both the χ^2 statistic and standard errors of the parameter estimates tend to exhibit bias as non-normality increases (e.g., Bollen, 1989; Chou, Bentler, & Satorra, 1991; Finch et al., 1997).

Chi-Square Statistic and Fit Indices

When estimating a correctly specified model, the ML-based χ^2 does not follow the expected central χ^2 distribution if the multivariate normality assumption is violated. More specifically, research has shown that χ^2 is inflated under conditions of moderate non-normality with values becoming more inflated as non-normality increased (e.g., Chou et al., 1991; Curran et al., 1996; Hu, Bentler, & Kano, 1992; Yu & Muthén, 2002). Kurtotic distributions, especially leptokurtic distributions (positive kurtosis), seem to have the greatest effect on χ^2 (e.g., Browne, 1984; Chou et al., 1991). The inflation of the χ^2 statistic may lead to an increased Type I error rate, which is a greater rate of rejecting a correctly specified model than expected by chance.

In addition to the χ^2 statistic, the performance of other fit indices is important to understand given that most researchers are interested in the approximate fit of the model to the data instead of an exact fit evaluation determined solely by the χ^2 test (Bentler, 1990). As Hu and Bentler (1998) explained, "A fit index will perform better when its corresponding chi-square test performs well" (p. 427), meaning that because many fit indices (e.g., comparative fit index [CFI]) are a function of the obtained

χ^2, these too can be affected by the same factors that influence χ^2. Research has shown that if moderately to severely non-normal data are coupled with a small sample size ($N \leq 250$), the ML-based Tucker-Lewis Index (TLI), CFI, and root mean square error of approximation (RMSEA) tend to overreject correctly specified models (Hu & Bentler, 1999; Yu & Muthén, 2002).

Parameter Estimates and Standard Errors

Whereas parameter estimates are unaffected by non-normality, their associated significance tests are incorrect if ML estimation is applied to non-normal data. Specifically, the ML-based standard errors underestimate the true variation of the parameter estimates (e.g., Chou et al., 1991; Finch et al., 1997; Olsson et al., 2000), which results in increased Type I error rates associated with statistical significance tests of the parameter estimates. This would imply that estimates of truly zero parameters could be deemed significantly different than zero, and thus, important, to include in the model. Similar to the χ^2, it appears that kurtotic distributions, specifically leptokurtic distributions, have the greatest impact on standard errors (Hoogland & Boomsma, 1998).

ORDERED CATEGORICAL DATA

Defining Ordered Categorical Data

As stated previously, ML estimation assumes that the observed data are a sample drawn from a continuous and multivariate normally distributed population. In the social sciences, data with these characteristics are not always collected. Frequently, researchers collect and analyze ordinal data, such as data obtained from the use of a Likert scale. While researchers often treat ordinal data as continuous, the ordinal measurements are, as Bollen (1989, p. 433) noted, "coarse" and "crude." Even if the data appear to be approximately normally distributed (e.g., indices of skewness and kurtosis approach zero or plots of the observed data appear to be normal), ordered categorical data are discrete in nature and, therefore, cannot be normally distributed by definition. The crude nature of the measurement will induce some level of non-normality in the data (Kaplan, 2000).

One way in which observed ordered categorical data are thought to occur is when a continuous latent response variable (y^*) is divided into distinct categories (e.g., Bollen, 1989; Muthén, 1993). This has been referred to as the latent response variable formulation (e.g., Muthén & Muthén, 2001). The points that divide the continuous latent response

variable (y^*) into a set number of categories (c) are termed thresholds (τ), where the total number of thresholds is equal to the number of categories less one ($c - 1$). For example, if a Likert scale has five response choices, four threshold values are needed to divide y^* into five ordered categories. The observed ordinal data (y) are thought to be produced as follows:

$$y = \begin{cases} 1 & \text{if } y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < y^* \leq \tau_3 \\ 4 & \text{if } \tau_3 < y^* \leq \tau_4 \\ 5 & y^* > \tau_4 \end{cases} \tag{2}$$

As a result, $y^* \neq y$. Specifically, because subjects respond to the five-point Likert scale, the observed ordinal-level data can only be reported as discrete values from 1 to 5. However, subjects' "true" levels of the latent response variable (y^*) are much more precise than allowed by the five-point response scale. Figure 9.1 illustrates the relation between the continuous latent response variable (y^*), observed level data (y), and the four ($c - 1$) threshold values for a variable with five ordered categories. This figure illustrates how the observed ordinal data provide an approximation of the underlying continuous latent response variable.

This difference between y and y^* has two important consequences when modeling the data. First, unlike y^*, the standard linear measurement model ($y^* = bF + E$) does not hold when modeling y ($y \neq bF + E$).[2] Second, the assumption that the model estimated reflects the true structure in the population ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$) does not hold when ordinal data are present

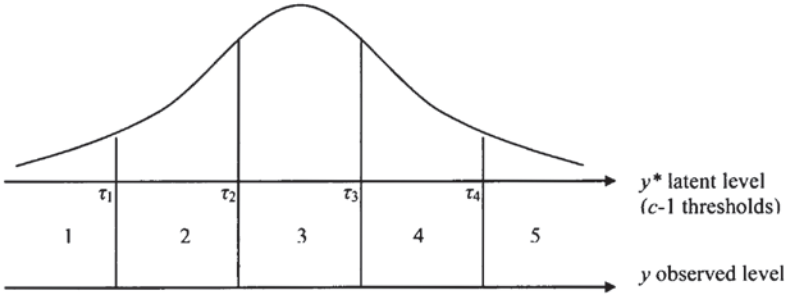

Figure 9.1. Relation between y^*, y, and thresholds.

(Bollen, 1989). Therefore, many studies have been conducted to examine the extent of bias present when applying the standard linear measurement model to ordinal data.

Effects of Analyzing Approximately Normally Distributed Ordered Categorical Data: Empirical Results

When modeling ordinal data, researchers often ignore the categorical nature of the data and apply ML estimation. By creating a covariance matrix based on Pearson product–moment (PPM) correlational techniques and estimating relations at the observed level (y), one is treating the ordinal data as if they were continuous. As the number of ordered categories increases, data more closely *approximate* continuous-level data and, in turn, the obtained correlations are closer to their true values (Bollen, 1989). The fewer categories present, the more severe the attenuation in the PPM correlations and the greater the discrepancy between true and obtained values. As discussed below, if few categories are used and ML estimation is employed, the model fit indices, parameter estimates, and standard errors can be biased.

Chi-Square and Fit Indices

In general, fit indices have been found to perform well if approximately normally distributed five-category ordinal data are treated as continuous (Babakus, Ferguson, & Jöreskog, 1987; Hutchinson & Olmos, 1998). While the χ^2 was found to be robust when modeling ordinal data was collected using four ordered categories, inflation was present if fewer than four categories were used (Green, Akey, Fleming, Hershberger, & Marquis, 1997). In addition, slight underestimation of the goodness-of-fit index (GFI), adjusted GFI (AGFI), and root mean square residual (RMR) has been found if sample sizes are small and ordered categorical data with five categories are analyzed as continuous (Babakus et al., 1987). Researchers generally agree that when ordinal data are approximately normal and have at least five ordered categories that the ordered categorical data may be treated as if they were continuous without great distortion in the fit indices (e.g., Bollen, 1989; Dolan, 1994; Muthén & Kaplan, 1985).

Parameter Estimates and Standard Errors

Previous research has shown that when ordinal data have at least five categories and are approximately normal, treating data as continuous and applying the ML estimator produces slight underestimation in parameter estimates and factor correlations (Babakus et al., 1987; Muthén & Kaplan,

1985). Standard errors have shown a greater sensitivity to categorization than the parameter estimates, exhibiting negative bias (Babakus et al., 1987; Muthén & Kaplan, 1985; West, Finch, & Curran, 1995). If standard errors are too small, tests of parameter significance may be inflated, resulting in Type I errors. As the number of ordered categories decreases, the underestimation in both the parameter estimates and standard errors becomes more severe, even if ordinal data are symmetric.

Effects of Analyzing Non-Normal Ordered Categorical Data: Empirical Results

Ordinal data are considered by some researchers as inherently non-normal (e.g., Muthén & Kaplan, 1985). However, as just described, if the observed data have many categories (e.g., at least five ordered categories) and are approximately normal, use of ML estimation techniques does not result in severe levels of bias in fit indices, parameter estimates, or standard errors. Problems begin to emerge as the number of response options decreases or the observed item distributions diverge widely from a normal distribution. As the number of ordered categories is reduced, there are fewer response choices available for subjects to choose. The fewer the number of categories, the greater the amount of attenuation in PPM estimates. Also, as the number of categories decreases, it becomes less likely that observed data could approximate a normal distribution.

Chi-Square and Fit Indices

Similar to results found when modeling continuous non-normal data, fit indices are adversely affected when ordinal data follow non-normal distributions. When modeling non-normal ordered categorical data, ML-based χ^2 values (Green et al., 1997: West et al., 1995) and RMR values were inflated, and values of the non-normed fit index (NNFI), GFI, and CFI were underestimated (Babakus et al, 1987; Hutchinson & Olmos, 1998). This may suggest that a correctly specified model does not fit the data well and could lead a researcher to discard a plausible model.

Parameters and Standard Errors

As univariate skewness and univariate kurtosis levels of the observed ordered categorical data increase, the negative bias observed with the ML-based parameter estimates and standard errors becomes more pronounced (Babakus et al., 1987; Muthén & Kaplan, 1985). Bias levels increased with lower sample sizes, fewer categories, weaker relations between factors and indicators, or higher levels of non-normality (e.g., Babakus et al., 1987; Bollen, 1989; Dolan, 1994).

TECHNIQUES TO ADDRESS NON-NORMAL AND ORDERED CATEGORICAL DATA

In this section we describe four methods that have been developed to address problems encountered when modeling non-normal and/or categorical data. The first method involves an alternative method of estimation that does not make the distributional assumptions of ML estimation (ADF; Browne, 1984). The second method involves adjusting the ML-based χ^2 and standard errors by a factor based on the level of multivariate kurtosis displayed in the observed data (Satorra-Bentler scaled χ^2 and standard errors). This method also involves adjusting any fit indices used to assess model fit (e.g., CFI, RMSEA). The third method involves employing robust WLS estimation methods (e.g., WLSM, WLSMV) available in the software package Mplus (Muthén & Muthén, 2004). These estimators can be conceptualized as combining an alternative estimation method with an adjustment method. Finally, a fourth method involves bootstrapping empirical distributions of each parameter estimate and the χ^2 statistic in order to produce more accurate standard errors and probability values associated with χ^2.

Asymptotically Distribution-Free (ADF) Estimator

Given the unrealistic assumption of multivariate normality and the lack of robustness of ML to non-normal data, Browne (1984) developed the ADF estimator. Unlike ML, ADF makes no assumption of normality; therefore, variables that are kurtotic have no detrimental effect on the ADF χ^2 or standard errors. In addition, input matrices that take the metric of the variables into consideration can be employed to handle the problems of parameter estimate attenuation associated with a small number of ordered categories (Muthén, 1984). For these reasons, it would seem as though non-normally distributed and/or ordered categorical data could be accommodated by the ADF estimation technique, thus avoiding the problems encountered by NT estimators.

ADF Estimator with Non-normal Continuous Data: Description

In order to understand the ADF estimator and some of its practical limitations, it is important to understand the form of the fit function, which is typically written as

$$F = (\mathbf{s} - \hat{\boldsymbol{\sigma}})' \mathbf{W}^{-1} (\mathbf{s} - \hat{\boldsymbol{\sigma}}) , \tag{3}$$

where $\mathbf{s}$ represents a vector of the nonduplicated elements in the sample covariance matrix ($\mathbf{S}$), $\hat{\boldsymbol{\sigma}}$ represents a vector of the nonduplicated ele-

ments in the model-implied covariance matrix [$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$], and ($\mathbf{s} - \hat{\boldsymbol{\sigma}}$) represents the discrepancy between the sample values and the model-implied values. These residuals ($\mathbf{s} - \hat{\boldsymbol{\sigma}}$) are weighted by a weight matrix, **W**. The weight matrix utilized with the ADF estimator is the asymptotic covariance matrix, a matrix of the covariances of the observed sample variances and covariances (Bollen, 1989). Elements of the asymptotic covariance matrix, $W_{ij,\,kl}$, are calculated using the covariances among the elements in the sample covariance matrix along with the fourth-order moments (Bentler & Dudgeon, 1996),

$$W_{ij,\,kl} = s_{ijkl} - s_{ij}s_{kl}, \tag{4}$$

where s_{ijkl}, a quantity related to multivariate kurtosis, is defined as

$$s_{ijkl} = \frac{\sum_{a=1}^{N}(x_{ai}-\bar{x}_i)(x_{aj}-\bar{x}_j)(x_{ak}-\bar{x}_k)(x_{al}-\bar{x}_l)}{N}, \tag{5}$$

and s_{ij} and s_{kl} are the covariances of observed variables x_i with x_j and x_k with x_l, respectively. This estimator is often called weighted least squares (WLS) or ADF (Bollen, 1989).

There are practical problems in implementing ADF estimation that are related to the weight matrix. Specifically, because the inverse of the weight matrix ($\mathbf{W}^{-1}$) needs to be calculated, a large weight matrix can make ADF estimation computationally intensive. Dimensions of **W** matrix can be calculated as ½ ($p + q$) ($p + q + 1$), where p is the number of observed exogenous variables and q is the number of observed endogenous variables (Bollen, 1989). For example, if a researcher has responses from a 10-item scale and wishes to employ confirmatory factor analysis, the dimensions of **W** would be 55 × 55, resulting in 3,025 elements in **W**. As the number of observed variables increases, the number of elements in **W** increases rapidly. For example, if 10 items were added to the original 10-item measure, the dimensions of the weight matrix would be 210 × 210, resulting in 44,100 elements in **W**. Due to the computational intensity of the ADF technique, it requires very large sample sizes for results to converge to stable estimates. A minimum sample size of $1.5(p + q)(p + q + 1)$ has been suggested (Jöreskog & Sörbom, 1996), but much larger sample sizes may be needed to alleviate estimation and convergence problems.

ADF Estimator with Continuous Non-normal Data: Empirical Results

When modeling non-normal data, theoretically, the ADF estimator should produce parameter estimates with desirable properties and fit sta-

tistics that perform as expected (Browne, 1984). However, empirical research has shown otherwise. ADF tends to break down under common situations of moderate to large models (more than two factors, eight items) and/or small to moderate sample sizes ($N < 500$). As discussed in detail below, the poor performance of the ADF estimator under many conditions makes it an unattractive option when modeling non-normal continuous data.

Chi-square and fit indices. With respect to model fit, ADF yields misleading results unless sample size is extremely large (e.g., Olsson et al., 2000). For example, when modeling non-normal continuous data, Hu and colleagues (1992) found that ADF estimation produced acceptable Type I error rates only when sample sizes reached 5,000. Similarly, Curran and colleagues (1996) found that the ADF-based χ^2 increased as sample size decreased and/or non-normality increased, resulting in correctly specified models being rejected too frequently. Even more problematic is the ADF estimator's lack of sensitivity to model misspecification. Research has shown that ADF estimation produces overly optimistic fit values when models are misspecified, which in turn could lead researchers to fail to reject an incorrectly specified model. The lack of sensitivity to specification errors becomes worse with increasing departures from normality (e.g., Curran et al., 1996; Olsson et al., 2000).

Parameter estimates and standard errors. Empirical results concerning ADF-based parameter estimates and standard errors are also discouraging. Parameters tend to be negatively biased unless the sample is large, with bias levels becoming more pronounced as kurtosis increases. In addition, ADF-based standard errors estimated from a correctly specified model under conditions of non-normality have been found to be superior to ML-based standard errors only when the observed variables have an average univariate kurtosis larger than three and the sample size is greater than 400 (Hoogland & Boomsma, 1998).

ADF Estimator with Continuous Data: Software Implementation

Researchers who wish to utilize ADF as an estimator will find it easy to employ using LISREL, EQS, or Mplus. In LISREL (Jöreskog & Sörbom, 2004), ADF estimation is called WLS and implementation requires the use of two programs (LISREL and PRELIS). As noted, WLS/ADF employs the asymptotic covariance matrix as the weight matrix. PRELIS (preprocessor for LISREL; Jöreskog & Sörbom, 1996) is used to produce both the asymptotic and observed covariance matrices from the raw data. These matrices are input into the LISREL program to estimate the model. Appendix A provides an example of the SIMPLIS command language (user-friendly language employed in LISREL program) that specifies WLS/ADF estimation. Notice that WLS must be specified on the options

line or else ML estimation will be employed by default. The combination of ML estimation and an asymptotic covariance matrix will produce the Satorra-Bentler scaling procedure (discussed below).

Similar to LISREL, the raw data file is necessary in order to construct the asymptotic covariance matrix in EQS (Bentler, 2004) and Mplus (Muthén & Muthén, 2004). Unlike LISREL, a preprocessor is not needed to construct this matrix in either EQS or Mplus; it is constructed and employed by specifying the estimator. Arbitrary GLS (AGLS) estimation is requested as the estimation method for EQS while WLS is requested for Mplus. AGLS follows the same general form as GLS, with the choice of weight matrix based upon fourth-order moments to allow distribution-free requirements of the variables (Bentler, 1995). While a different name is used, it is equivalent to ADF/WLS. All three programs tend to produce similar parameter estimates, standard errors, and fit indices.

ADF Estimator with Ordered Categorical Data: Description of Categorical Variable Methodology (CVM)

As previously discussed, if researchers ignore the ordinal metric of the data (i.e., treating ordinal data as if continuous and employing ML estimation), data–model fit and parameter estimates may be underestimated. Alternative strategies consider the metric of the ordinal data by including this information in the estimation procedures. Specifically, categorical variable methodology (CVM) incorporates the metric of the data into analyses by considering two components: (1) input for analyses that recognizes the ordered categorical indicators and (2) the use of the correct weight matrix when employing ADF/WLS estimation (Muthén, 1984; Muthén & Kaplan, 1985). Therefore, CVM is basically ADF/WLS estimation with specific input to accommodate ordered categorical variables.

Employing CVM. With regard to metric, if data are continuous then data at the observed level are considered equivalent to the underlying latent response variable, that is, $y = y^*$. On the other hand, if data are ordinal, $y \neq y^*$ (e.g., Jöreskog & Sörbom, 1996; Muthén & Kaplan, 1985; Muthén & Muthén, 2001). In order to avoid the consequences of modeling y using the standard linear confirmatory factor analysis model (e.g., $y \neq bF + E$ and $\Sigma \neq \Sigma(\theta)$), y^* can be modeled (Bollen, 1989). Applying the linear factor model to the underlying latent response variable is illustrated in Figure 9.2. Notice that the factor does not directly affect y, but instead, directly affects y^*. Because y^* is a continuous variable, the standard linear model ($y^* = bF + E$) can be used to estimate the relation between y^* and the factor.

Modeling the relation between F and y^* entails computing thresholds and latent correlations. Because thresholds are used in the computation of latent correlations, these will be discussed first. Threshold

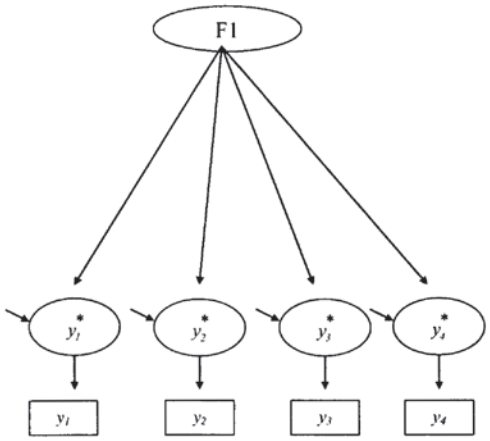

Figure 9.2. Latent response variable formulation.

values, which cut the underlying continuum into ordered categories (see Figure 9.1), may be estimated if the number of subjects who chose a certain category is known (Bollen, 1989). Threshold values are important to estimate not only because they are thought of as critical points that "move" a subject from one category to another but also because they are used to create marginal distributions of ordinal variables that assist with estimation procedures. Because the metric of ordered categorical data is arbitrary, the mean and standard deviation are often set to 0 and 1, respectively. Using a mean of 0 and standard deviation of 1, item thresholds may be estimated by considering the cumulative area under the normal curve up to a given point (Bollen, 1989; Jöreskog & Sörbom, 1996) by:

$$\tau_i = \Phi^{-1}\left[\sum_{k=1}^{i}\frac{N_k}{N}\right], \qquad i = 1, 2, \ldots, c-1 \tag{6}$$

where τ_i is a particular threshold, Φ^{-1} is the inverse of the normal distribution function, N_k is the number of subjects who selected category k, N is the total sample size, and c is the total number of categories. Equation 6 shows that thresholds are calculated using the proportions of subjects within each ordered category (N_k/N). The resulting threshold values (z-values) divide the underlying normal distribution into the c categories and relate the y values to the y^* values. By incorporating thresholds into the standard measurement model, the linear model estimates the relation between the fac-

tor and the continuous underlying latent response variable, thus avoiding the problems associated with modeling y (Bollen, 1989).

Recall that in addition to the linear model not applying to ordinal data, the assumption of a correctly specified model does not hold when ordinal variables are modeled. In brief, the population covariance matrix of the ordinal variables ($\boldsymbol{\Sigma}$) will not equal the population covariance matrix of the continuous underlying latent response variable ($\boldsymbol{\Sigma}^*$), for which the model does hold ($\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}(\boldsymbol{\theta})$). In order to avoid this problem, correlations representing the relations among the y^* variables can be estimated (Bollen, 1989).

The latent correlations represent the theoretical relations between the underlying continuous latent response variables (y^*). For each pair of variables, a latent correlation can be estimated. If both variables are dichotomous, a tetrachoric correlation represents the relation between the y^* variables. If both variables are ordinal, a polychoric correlation represents the relation between the y^* variables. If one variable is ordinal and the other variable is continuous, a polyserial correlation represents the relation between the y^* variables.

The WLS estimator can then be employed using the thresholds and latent correlations:

$$F_{\text{WLS}} = (\mathbf{r} - \hat{\boldsymbol{\rho}})' \mathbf{W}^{-1} (\mathbf{r} - \hat{\boldsymbol{\rho}}) , \tag{7}$$

where $\mathbf{r}$ is a vector containing the sample latent correlations and thresholds, $\hat{\boldsymbol{\rho}}$ is the corresponding vector from the implied matrix, and $\mathbf{W}$ is the asymptotic covariance matrix of $\mathbf{r}$ (e.g., Muthén, 1984). This asymptotic covariance matrix, along with the appropriate correlational input, is required to correctly implement CVM. It is important to realize that both pieces (appropriate correlations and correct $\mathbf{W}$) are necessary. If either piece is missing, the estimation technique is not CVM.

An example to illustrate the calculation of a polychoric correlation. To illustrate how ordinal variables may be accommodated, consider an example employing Rosenberg's (1989) self-esteem scale. A sample of 120 college freshmen responded to questions concerning their self-esteem using a scale with four ordered categories anchored at "Strongly Disagree" to "Strongly Agree." Responses to two of the questions, "On the whole, I am satisfied with myself" (SATISFIED) and "I feel that I have a number of good qualities" (QUALITIES), are provided in Table 9.1.

To illustrate how threshold values are obtained, we can compute these values directly from the sample data using two pieces of information: (1) the proportion of students who selected a certain category and (2) the area under the normal curve. Consider the first threshold for the variable, SATISFIED. This threshold divides the underlying latent response vari-

Table 9.1. Frequency of Responses to Self-Esteem Items

SATISFIED		*QUALITIES*	
Category	*Frequency*	*Category*	*Frequency*
1 (SD)	4	1 (SD)	4
2 (D)	16	2 (D)	27
3 (A)	50	3 (A)	42
4 (SA)	50	4 (SA)	47
N	120	*N*	120
Skew	−.812	Skew	−.514
Kurtosis	.154	Kurtosis	−.726

Note: SD = Strongly Disagree; D = Disagree; A = Agree; SA = Strongly Agree.

able into the two categories of "Strongly Disagree" (category 1) and "Disagree" (category 2). Students below this threshold will have responded "Strongly Disagree" to the statement. There are four students responding "Strongly Disagree" to the SATISFIED item. Using Equation 6, the cumulative probability of cases through category 1 is (4/120), or .033. Considering this as representative of cumulative area under the normal curve, the threshold value is the z-value associated with .033, or a z-value of −1.83. The remaining thresholds may be calculated in a similar manner, as shown in Table 9.2.

To determine the polychoric correlation between two ordinal variables, a contingency table of students' responses to each pair of the variables is needed. The frequency of responses to each option can be tabulated across the pair of variables. Table 9.3 shows the contingency table for the responses to the SATISFIED and QUALITIES variables. From the table, one can see a relation between the responses to the items. For example, those students who agreed or strongly agreed that they were satisfied with themselves generally agreed that they had a number of good qualities. However, some discrepancies are noticed. For example, six students who agreed that they possessed good qualities disagreed with the statement about having self-satisfaction.

Table 9.2. Threshold Values and Cumulative Area

	SATISFIED			*QUALITIES*		
Threshold	1	2	3	1	2	3
Cumulative *N*	4	20	70	4	31	73
Cumulative Area	.033	.167	.583	.033	.258	.608
Threshold Value	−1.834	−0.967	0.210	−1.834	−0.648	0.275

Table 9.3. Contingency Table Between SATISFIED and QUALITIES Variables

	QUALITIES				
SATISFIED	*1 (SD)*	*2 (D)*	*3 (A)*	*4 (SA)*	*Total*
1 (SD)	2	1	1	0	4
2 (D)	0	8	6	2	16
3 (A)	1	15	26	8	50
4 (SA)	1	3	9	37	50
Total	4	27	42	47	120

Note: SD = Strongly Disagree; D = Disagree; A = Agree; SA = Strongly Agree.

The contingency table information reported in Table 9.3 and item thresholds are used to calculate latent correlations among the ordinal variables (see Olsson, 1979, for formula). The estimated polychoric correlation represents the value with the greatest likelihood of yielding the observed contingency table, given the estimated thresholds. Here, the polychoric correlation is .649, reflecting the positive relation, while acknowledging some inconsistency in student responses. Note that the polychoric correlation estimate is higher than the PPM correlation estimate (.551) because the polychoric correlation is disattenuated for the error associated with the ordinal variable's coarse categorization of the underlying latent variable's continuum. To illustrate the polychoric correlation graphically, the relation between the observed ordinal variables and the underlying latent responses variables are plotted in Figure 9.3.

ADF/CVM with Ordered Categorical Data: Empirical Results

Although differences may exist in how software programs employ CVM, the information presented here is made without specific reference to software packages (details concerning the implementation of CVM using LISREL, EQS, and Mplus are described below). In general, empirical studies have found that CVM has both desirable and undesirable characteristics.

Chi-square and fit indices. When approximately normally distributed ordinal data were analyzed using CVM, values of χ^2 were close to expected values when small to moderate models were specified (15 parameters or less) or sample sizes were large (N =1,000; Muthén & Kaplan, 1985; Potthast, 1993). The amount of inflation of the CVM-based χ^2 increased as sample size decreased, model size increased, or non-normality of the data increased (DiStefano, 2002; Muthén & Kaplan, 1992; Potthast, 1993). RMSEA has been found to be somewhat robust to ordinal data analyzed

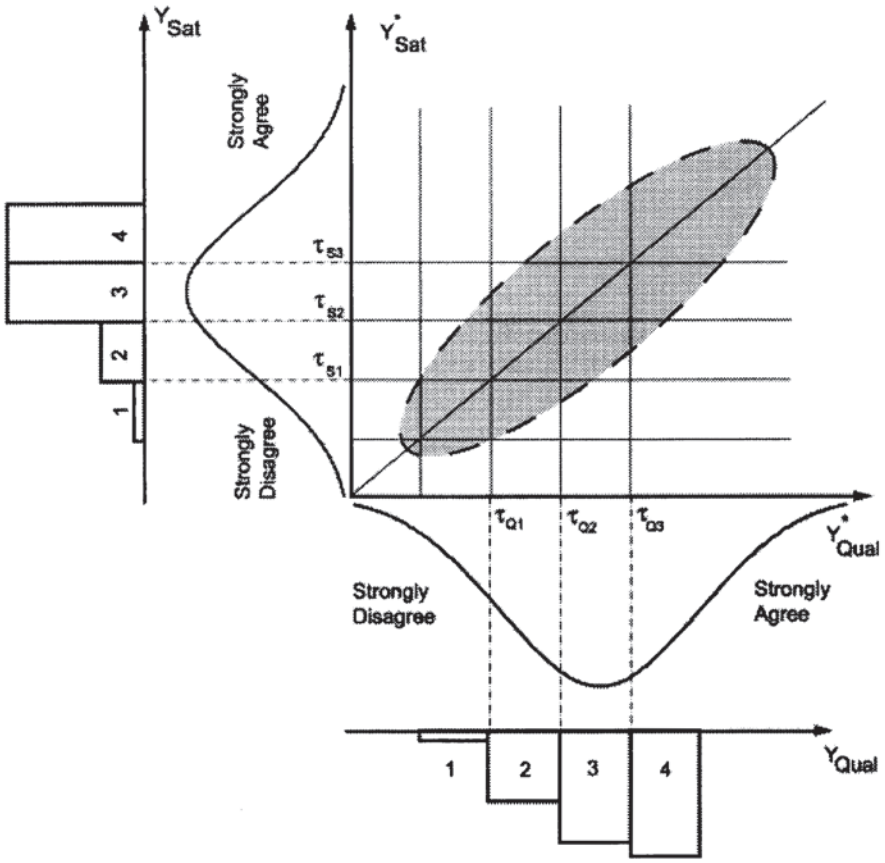

Figure 9.3. Illustration of polychoric correlation between two observed variables. Notes: Sat = SATISFIED; Qual = QUALITIES. Threshold values are specific to the variable in question. For example, S1, S2, S3, respectively, refer to thresholds 1 through 3 for the variable SATISFIED. Similarly, thresholds Q1, Q2, Q3, respectively, refer to thresholds 1 through 3 for the variable QUALITIES.

by CVM, and is not sensitive to sample size or model size when correctly specified models were estimated (Hutchinson & Olmos, 1998).

Parameter estimates and standard errors. A major strength of CVM is that parameter estimates appear unbiased when modeling non-normally distributed ordered categorical data. When estimating correctly specified models, parameter estimates were found to have little bias, regardless of whether dichotomous variables (Dolan, 1994; Muthén & Kaplan, 1985) or non-normal ordered categorical data (approximate values: skewness = 2.5, kurtosis = 6) were analyzed (DiStefano, 2002; Dolan, 1994; Potthast,

no exogenous variable and CVM is conceptualized as discussed above (latent response variable formulation). Specifically, it is assumed that continuous normally distributed latent response variables (y^*) underlie the ordinal variables (y). The thresholds, latent correlations (e.g., polychorics), and the asymptotic covariance matrix are estimated. WLS can then be employed as the estimator.

In case B, there is an exogenous variable influencing the factor (e.g., MIMIC models). In this situation, a different estimation process is employed. As an alternative to the latent response variable formulation there exists an equivalent formulation, termed the conditional probability curve formulation (e.g., Muthén & Asparouhov, 2002; Muthén & Muthén, 2001). Instead of estimating the linear relation between y^* and the factor (as with the latent response variable formulation), the nonlinear relation between y and the factor may be modeled. Specifically, a probit model can be used to estimate the probability (ranging from 0 to 1) that a specific category (k) is selected or exceeded by modeling the nonlinear relation between y and the factor (F):

$$P(y \geq k|\mathrm{F}) = \Phi\,(\alpha_k + \beta\mathrm{F}), \tag{8}$$

where Φ is the standard normal distribution function.[3] The α (intercept) and β (slope) parameters from this conditional probability formulation can be derived from the parameters estimated using the latent response variable formulation, which illustrates the similarity of the two formulations (see Muthén & Asparouhov, 2002, for formula). In fact, the two formulations produce equivalent results in terms of the probability of being in or exceeding a category given the factor, $P(y \leq k|\mathrm{F})$ (Muthén & Asparouhov, 2002). Because these two formulations produce equivalent results, Muthén and Asparouhov explained that the assumption of an underlying continuous latent response variable is not necessary but rather a convenience: "It is shown that the two formulations give equivalent results. The discussion clarifies that the latent response variables are a convenient conceptualization, but that it is not necessary that the data have been generated by categorizing latent response variables" (Muthén & Asparouhov, 2002, p. 2).

When the conditional probability formulation is employed, the first step involves computing the sample statistics: probit thresholds, probit regression coefficients, and probit residual correlations. In the second step, the asymptotic covariance matrix of these sample statistics is constructed. In the final step, estimates of model parameters, standard errors, and model fit information can be computed using a WLS estimator (Muthén & Muthén, 2001).

Satorra-Bentler Scaled Chi-Squares and Standard Errors

Description of Method

Another strategy employed to accommodate non-normal continuous and/or categorical data that has become popular in the last several years involves adjusting the χ^2, fit indices, and standard errors by a factor based on the amount of non-normality in the data. With normally distributed data the expected value of the χ^2 is equal to the model's degrees of freedom when the model is correctly specified. Therefore, if a correctly specified model with 60 degrees of freedom was estimated using multivariate normal data, a χ^2 value of approximately 60 would be expected. However, if data are moderately non-normal, the ML-based χ^2 will be biased even though the model is correctly specified. A correction, typically called the Satorra-Bentler (S-B) scaling procedure, uses the observed data's distributional characteristics to adjust the ML-χ^2 in order to better approximate the theoretical χ^2 reference distribution:

$$\text{S-B } \chi^2 = d^{-1}(\text{ML-based } \chi^2), \tag{9}$$

where d is a scaling factor that incorporates the kurtosis of the variables (Chou & Bentler, 1995; Satorra & Bentler, 1994). If no multivariate kurtosis exists, then ML-based χ^2 = S-B χ^2. However, as the level of multivariate kurtosis increases, the S-B χ^2 becomes more discrepant from the ML-based χ^2.

The S-B scaling method is typically applied with ML estimation. Because ML is employed, computation problems experienced with the ADF estimator are avoided. Recall that ADF requires the inversion of a large asymptotic covariance matrix (**W**). The S-B scaling method does not require inversion of this **W** matrix. Instead, the matrices to be inverted in order to compute the scaling factor are of the smaller dimensions of $df \times df$, where df represents the degrees of freedom associated with the model (Satorra & Bentler, 1994).

A similar scaling process is used to correct the standard errors, alleviating some of the attenuation present when modeling non-normal data using ML estimation. Specifically, the scaled standard errors are adjusted upward to approximate those that would have been obtained if the data were normally distributed.[4] Recall that when non-normal data are analyzed using ML, the parameter estimates are not affected. Therefore, the ML-based parameter estimates are not adjusted in any way when this method is employed.

It must be noted that the typical χ^2 difference test employed for nested model comparisons should not be calculated using the S-B scaled χ^2 values (i.e., simply subtracting χ^2 of the less parsimonious model from the χ^2

of the more parsimonious model). The difference between two S-B χ^2 values is not distributed as a χ^2. Fortunately, fairly simple calculations can be employed to correct the difference test in order to make nested model comparisons using the S-B χ^2 values (see Satorra & Bentler, 2001, for calculations).

S-B Scaling Methods with Continuous Non-normal Data: Empirical Results

Chi-square and fit indices. Studies using correctly specified models and continuous non-normal data have shown that the S-B χ^2 outperforms the ML-based χ^2, particularly as the degree of non-normality increases (e.g., Chou et al., 1991; Curran et al., 1996; Hu et al., 1992; Yu & Muthén, 2002). In addition, it performs better than the ADF-based χ^2 when estimating correctly specified models at all but the largest sample sizes (e.g., N = 1,000, Curran et al., 1996; N = 5,000, Hu et al., 1992).

If the S-B χ^2 is employed to handle non-normal data, it follows that the S-B χ^2 would be incorporated into the calculation of fit indices (e.g., TLI, CFI) in order to gain benefits of the scaling procedure and, in turn, provide more accurate reflections of model–data fit. Few studies have investigated the performance of these S-B scaled indices. Nevitt and Hancock (2000) found that the S-B scaled RMSEA outperformed the unadjusted index. Yu and Muthén (2002) also examined this index in addition to the S-B scaled TLI and CFI and found that, under conditions of moderate to severe non-normality coupled with small sample size ($N \leq 250$), the S-B scaled versions of these three indices are preferred over the ML-based estimates. Yu and Muthén suggested that values at or below .05 for the S-B scaled RMSEA and at or above .95 for the S-B scaled CFI indicate adequate fit, which is quite similar to the cutoff values recommended by Hu and Bentler (1999) for the unadjusted indices.

Parameter estimates and standard errors. The scaled standard errors have also been found to outperform ML-based and ADF-based standard errors under conditions of non-normality (Chou & Bentler, 1995; Chou et al., 1991). Similar to the ML-based and ADF-based standard errors, the scaled standard errors showed some negative bias. However, they tended to be much closer to the expected values of the standard errors than those obtained from either ADF or ML methods. Recall that the ML-based parameter estimates are not adjusted as they are not affected by non-normality.

S-B Scaling Methods with Non-Normal Ordered Categorical Data: Empirical Results

While the functioning of the S-B correction has been generally examined with continuous non-normal data, a few studies have evaluated how the correction performs with ordered categorical data. It is recognized

that this approach treats the categorical data as continuous, ignoring the metric level of the data. A caveat to this method is that ML estimation is known to be sensitive to the number of ordered response categories in addition to non-normal distributions.

Chi-square and fit indices. With regard to model fit, Green and colleagues (1997) found that S-B scaling produced χ^2 values very close to the expected χ^2 values when modeling two-, four- or six-ordered category data that displayed symmetric, uniform, and negatively skewed distributions. The S-B χ^2 did show positive bias when modeling data that exhibited differential skew, with bias being the greatest in the two-category condition. In all conditions, the S-B χ^2 outperformed the ML-based χ^2.

Parameter estimates and standard errors. As noted above, the S-B correction simply scales the standard errors and the χ^2. Thus, S-B–based parameter estimates will be equivalent to ML-based estimates. This implies no correction for the attenuation of the parameter estimates due to the categorical nature of the data. With respect to standard errors, research has found that scaled standard errors exhibited greater precision than ML-based standard errors when non-normally distributed ordered categorical data were analyzed (skewness ≈2, kurtosis ≈6; DiStefano, 2002). The benefit of the S-B scaling method was present even with non-normally distributed data having as few as three ordered categories (DiStefano, 2003).

S-B Scaling Methods: Software Implementation

As discussed above, the S-B scaling method has been applied to ordered categorical data by treating them as continuous (i.e., calculation of PPM covariances instead of latent correlations for ordered categorical data). This implies that the implementation of this method is the same across the two data types. All three software programs calculate the S-B–scaled χ^2 and standard errors. Appendix B provides the syntax needed to employ the S-B scaling method.

Similar to ADF estimation with continuous data, LISREL calculates the S-B χ^2 and scaled standard errors in two steps. The first step involves computing the asymptotic and observed covariance matrix from the raw data file using PRELIS. The second step involves specifying the model and estimation technique in the SIMPLIS program file. Recall that ML is used as the estimator and the obtained χ^2 and standard errors are adjusted for the level of non-normality. It is important to note that LISREL 8.54 does not adjust all fit indices for non-normality. In fact, no incremental indices are adjusted (e.g., CFI, NNFI). Adjustments to these indices must be calculated by hand. This is done by first specifying and estimating the independence model and then using the independence model's S-B χ^2 along with the hypothesized model's S-B χ^2 in the corre-

sponding fit index formula (see, e.g., Hu & Bentler, 1998, for fit index formulas).

Both EQS and Mplus read the raw data file directly into the program in order to construct the necessary matrices used to scale the χ^2 and standard errors. In addition to specifying ML as the estimator in EQS, the word "robust" is included to request the scaling method. In Mplus one simply requests "MLM" as the estimator, which refers to the ML-mean-adjusted χ^2 (equivalent to the S-B χ^2) and scaled standard errors. Unlike LISREL 8.54, both EQS 6.1 and Mplus 3.01 adjust all the fit indices reported when the S-B scaling method is used.

It must be noted that in addition to the MLM estimator, Mplus also provides the MLMV estimator, which produces a mean- and variance-adjusted chi-square. The scaled standard errors are equivalent across the two estimators. Research has shown that the MLM and MLMV chi-squares perform similarly, indicating that the additional adjustment provided by MLMV may not be needed (Muthén, 1999).

Robust WLS Estimation Procedures

Description of Estimators

Because the S-B scaling method does not adjust parameters for the metric of the data when modeling ordered categorical data, it may seem as though CVM is a more attractive option for data of this type. However, the computational demands of the ADF/WLS estimator make CVM an implausible option for dealing with ordered categorical data unless an extremely large sample size is available. Muthén (1993) developed and implemented two robust WLS estimators (WLSM and WLSMV) that avoid the necessity of a large sample size by decreasing the computational intensity found with the traditional ADF/WLS estimator. In addition, these estimators incorporate scaling similar to the S-B scaling methods.

Concerning the details of the estimators, WLSM and WLSMV differ from the conventional ADF/WLS estimator in the use of the asymptotic covariance matrix. Although WLS, WLSM, and WLSMV all use the same asymptotic covariance matrix, they differ in what elements of the weight matrix are used and how they are used. ADF/WLS employs and inverts the full weight matrix in order to estimate parameters, standard errors, and χ^2. Instead of using the full matrix when estimating parameters, WLSM and WLSMV use only the diagonal elements of the weight matrix (e.g., asymptotic variances of the thresholds and latent correlation estimates). While WLSM and WLSMV do utilize the entire weight matrix to compute standard errors for the parameters, they employ a method that avoids its inversion (Muthén, 1993; Muthén, du Toit, & Spisic, 1997).

Like the standard errors, the χ^2 is calculated using the full weight matrix but avoids its inversion. In addition, a scaling factor, akin to the one employed in the Satorra-Bentler scaled χ^2, is employed to more thoroughly adjust the χ^2. Specifically, WLSM produces a mean-adjusted chi-square. WLSMV differs from WLSM in that the χ^2 is both mean- and variance-adjusted. The standard errors and parameter estimates from WLSM and WLSMV are equivalent. An additional distinction exists between WLSM and WLSMV in that WLSMV does not calculate model degrees of freedom in the standard way. Instead, degrees of freedom are "estimated" to approximate a χ^2 distribution and are lower in value than standard degrees of freedom (Muthén & Muthén, 2001). Therefore, fit indices (e.g., CFI, NNFI) will differ due to both the different degrees of freedom and different χ^2 values used in the calculations. Table 9.4 provides a brief outline of the differences and similarities between these three estimation techniques available in Mplus.

With the robust WLS estimators comes a new index for use with categorical data, the weighted root mean square residual (WRMR). This index is well suited for categorical data because it incorporates the asymptotic variances into the computation. The WRMR is also appropriate to employ with non-normal continuous data, or if variables have large variances (Muthén & Muthén, 2001). WRMR values under 1.0 have been recom-

Table 9.4. Mplus Estimation Techniques for Ordered Categorical Data

	Description	*Chi-square Estimation*	*Parameter Estimates*	*Standard Errors*	*Applied When?*
WLS	• Weighted least squares parameter estimates • Conventional χ^2 • Conventional standard errors	Full weight matrix used and inverted	Full weight matrix used	Full weight matrix used and inverted	Categorical or continuous endogenous variables
WLSM	• Weighted least squares parameter estimates • Mean-adjusted χ^2 • Scaled standard errors	Full weight matrix used but not inverted	Diagonal weight matrix used	Full weight matrix used but not inverted	At least one categorical endogenous variable
WLSMV	• Weighted least squares parameter estimates • Mean- and variance-adjusted χ^2 • Scaled standard errors	Full weight matrix used but not inverted	Diagonal weight matrix used	Full weight matrix used but not inverted	At least one categorical endogenous variable

mended to represent good fit with continuous or categorical data, with smaller values indicating better fit (Yu & Muthén, 2002).

WLSM and WLSMV: Empirical Results

There is limited research examining the functioning of the robust WLS estimators. Results have shown that these estimators perform much better than the conventional WLS when the tested model is large (15 variables) or sample size is small (N =1,000), yielding less biased χ^2 and standard errors (Muthén, 1993). Tentative results suggest that WLSMV outperforms WLSM, with WLSM showing higher Type I error rates (Muthén, 1999, 2003; Muthén et al., 1997). WLSMV tends to perform well except under conditions of small sample sizes (N = 200) and markedly skewed variables (Muthén et al., 1997).

WLSM and WLSMV: Software Implementation

Neither LISREL 8.54 nor EQS 6.1 have the capabilities to employ WLSM or WLSMV estimation techniques. Therefore, discussion of the implementation of these estimators will be limited to Mplus. In Mplus, when ordered categorical dependent variables are analyzed, the default estimator is WLSMV. The syntax for using the WLSM and WLSMV is very similar to the Mplus WLS syntax reported in Appendix A. The categorical indicators are identified by including the heading "CATEGORICAL ARE:" when defining the "VARIABLES" command. Once the variables are defined to be categorical, CVM will be conducted. To request one of the robust estimation techniques, changes are made to the "ESTIMATOR IS:" heading under the "ANALYSIS" command. Insert "WLSM" to request the mean-adjusted WLS estimation procedure or "WLSMV" to use the mean- and variance-adjusted estimation technique. Appendix C outlines these procedures.

Bootstrapping

General Description

When a model is correctly specified and data are multivariate normal, the expected value of the χ^2 statistic equals the model's degrees of freedom. The degrees of freedom are used to identify the corresponding central χ^2 distribution necessary to evaluate the probability value associated with the obtained χ^2. The S-B adjustment described above also uses the theoretical χ^2 distribution in order to evaluate statistical significance even though data are not multivariate normal. Because data are not multivariate normal, the obtained ML-based χ^2 is adjusted when using the S-B cor-

rection in order to better approximate the expected χ^2 distribution under conditions of normality. Instead of using the theoretical sampling distribution and adjusting the obtained χ^2 for non-normality of the data, bootstrapping techniques can be conceptualized as using the obtained χ^2 and adjusting the sampling distribution used to compute the probability value. More specifically, bootstrapping can be used to construct an empirical distribution of model test statistics that incorporates the non-normality of the data and relieves researchers from relying on the theoretical χ^2 distribution and its underlying assumptions.

In general, bootstrapping is a resampling technique that treats the observed sample data as an estimate of the population (Efron & Tibshirani, 1993). A large number of cases are then drawn with replacement from the observed data (parent sample) in order to create B bootstrap samples of the same size (N). Sampling with replacement implies that a given case may appear more than once in the same bootstrap sample. Using the sample data, the statistic of interest is computed from each of the B bootstrap samples. The estimates computed from the B samples then form an empirical sampling distribution of the statistic of interest.

In SEM, there are two methods of conducting bootstrapping, the naive bootstrap and the Bollen-Stine bootstrap. They differ in the type of empirical distribution formed. We first discuss the naive bootstrap and its application to estimating standard errors and then discuss the Bollen-Stine bootstrap and its application to estimating the probability value associated with the obtained χ^2 value.

Assume we are estimating a structural model using non-normal data. In addition to testing model fit, we also wish to evaluate the statistical significance of the estimated parameters. Given non-normal data, the ML-based standard errors will be underestimated. In contrast, the bootstrap standard errors take into account the distribution of the data, producing more accurate standard errors. Using the naive approach, B samples of size N are drawn (with replacement), and the parameter estimate of interest (e.g., factor loading, error variance) is estimated using each of the B bootstrap samples. This distribution of bootstrap parameter estimates is then used to calculate the standard error for that parameter. Specifically, the standard deviation of the bootstrap estimates represents the bootstrap standard error.

While the procedure above may provide better estimates of standard errors, it is not appropriate for bootstrapping the empirical distribution of the χ^2 statistic in SEM. Specifically, the resulting sampling distribution is incorrect because it reflects not only non-normality and sampling variability, but also model misfit (i.e., the null hypothesis of perfect fit is false). As Bollen and Stine (1992) pointed out, the parent sample must first be transformed to reflect the covariance structure underlying the

hypothesized model (i.e., bootstrap samples are drawn from a parent sample where the null hypothesis is true). Transforming the data so that the parent sample conforms to the specified model is necessary in order to generate χ^2 values that reflect only sampling variation and the impact of non-normality, and not model misfit.

For example, suppose we are estimating a model with 60 degrees of freedom using non-normal data. The sample data are first transformed to have the same covariance structure as the implied covariance matrix (using a matrix transformation given in Bollen & Stine, 1992), then B random samples of size N are taken from the transformed parent sample data matrix, and the χ^2 value for each sample is computed in order to form the empirical sampling distribution of χ^2 values. This empirical sampling distribution can then be used as the reference distribution to identify the probability of the ML-based χ^2. Suppose that the mean χ^2 value across the bootstrap samples equaled 72. This implies that the expected value of the empirical distribution is larger than that of the theoretical distribution (in this case 60). One would therefore expect differing probability values associated with the ML-based χ^2 when employing the empirical versus the theoretical distribution. The probability value associated with the obtained χ^2 using the empirical distribution is simply the proportion of bootstrap χ^2 values that exceed the ML-based χ^2 obtained from the original analysis (i.e., from fitting the proposed model to the untransformed sample data).

Bootstrapping with Continuous Non-Normal Data: Empirical Results

Chi-square. There have been very few studies that have examined the performance of the Bollen-Stine bootstrap for estimating the χ^2 probability value. Fouladi (1998) found that, similar to the S-B scaled χ^2, the Bollen-Stine bootstrap controlled Type I error rates better than NT methods when data were non-normal. In addition, when modeling non-normal data, the Bollen-Stine bootstrap generally provided more accurate probability values than the S-B scaled χ^2 in all conditions except when sample size was large. Fouladi did not necessarily advocate one method over the other but instead suggested that readers realize the liberal or conservative bias of each statistic and use this information to inform decisions pertaining to the plausibility of a model.

Similar to Fouladi (1998), Nevitt and Hancock (2001) found that the model rejection rates based on the Bollen-Stine bootstrap and the S-B χ^2 were more accurate than those from the ML-based χ^2 when estimating correctly specified models under conditions of moderate non-normality. Even under conditions of extreme non-normality and small sample sizes, the Type I error rate associated with the Bollen-Stine bootstrap was con-

trolled, outperforming the S-B χ^2. It is important to note, however, that the bootstrap showed less power to identify misspecified models than the S-B χ^2. Again, this tradeoff between controlling Type I error rate and sensitivity to misspecification complicates the decision of choosing one of the two techniques over the other.

Standard Errors. Nevitt and Hancock (2001) also investigated the performance of the naive bootstrap standard errors. Under conditions of non-normality, the bootstrap standard errors displayed less bias than the ML-based standard errors, and to a lesser extent, the S-B robust standard errors. However, it must be noted that the bootstrap standard error estimates displayed notable variability, signifying a possible concern with the stability of these estimates. A related finding concerning the stability and bias of the naive bootstrap standard errors suggests that small samples ($N \leq 100$) should be avoided due to a dramatic increase in both the variability and bias of the bootstrap standard errors. In addition, increasing the number of bootstrap samples beyond 250 has seemingly no benefits in terms of decreasing the bias in standard errors or model rejection rates.

Bootstrap: Software Implementation

As noted by Fan (2003), there are very few applications of bootstrapping in substantive research, which may be due to the limited automated procedures in SEM software programs. Of the three software programs presented here (EQS 6.1, Mplus 3.01, and LISREL 8.54), EQS (version 6 or higher) is the only program that has an automated bootstrapping option that can produce Bollen-Stine χ^2 probability values, accompanied by naive bootstrap standard errors (see Appendix D for syntax). Mplus 3.01 has an automated bootstrapping option that can produce naive bootstrap standard errors and bootstrap confidence intervals for the parameter estimates (Muthén & Muthén, 2004). AMOS (Arbuckle & Wothke, 1999), another popular SEM software program, also has automated bootstrapping capabilities that can produce the Bollen-Stine χ^2 probability values and the naive bootstrap standard errors (Byrne, 2001).

SUGGESTIONS FOR DEALING WITH NON-NORMALITY AND ORDERED CATEGORICAL DATA

Recommendations

Because so much information is associated with issues surrounding non-normal and ordered categorical data, Table 9.5 summarizes our recommendations when analyzing data of this type. This table is not meant to trivialize the complex issues surrounding this topic; instead, it may be

treated as a supplement to the information presented in this chapter. Also, as explained throughout the chapter, the type of estimation method or technique employed is closely tied to the degree of non-normality and/or the crudeness of the categorization. Therefore, researchers need to recognize the type of data they are modeling, in terms of both metric and distribution, before selecting a technique.

In brief, ML has been shown to be fairly robust if continuous data are only slightly non-normal; therefore, we recommended its use in this situation (e.g., Chou et al., 1991; Green et al, 1997). If data are continuous and non-normally distributed, we recommend the use of either the S-B scaling method or bootstrapping. Given the availability of the S-B scaling method, the ease of its use, and the empirical studies showing promising results, we can easily understand why this method is becoming increasingly popular.

When modeling ordered categorical data, the research seems to indicate that if there are a large number of ordered categories the data could be treated as continuous in nature. If the variables have five categories or

Table 9.5. Recommendations for Dealing with Non-Normal and Ordered Categorical Data

Type of Data	*Suggestions*	*Caveats/Notes*
Continuous Data		
1. Approximately normally distributed	• Use ML estimation	• The assumptions of ML are met and estimates should be unbiased, efficient, and consistent.
2. Moderately non-normal (skew < 2, kurtosis < 7)	• Use ML estimation; fairly robust to these conditions • Use S-B scaling to correct χ^2 and standard errors for even slight non-normality	• Given the availability of S-B scaling methods in the software packages, one could always employ and report findings from both ML estimation and S-B scaling method.
3. Severely non-normal (skew > 2, kurtosis > 7)	• Use S-B scaling • Use bootstrapping	• S-B correction works well, currently much easier to implement than the bootstrap, and tends to be more sensitive to model misspecification than the bootstrap. • Fit indices that are not adjusted for the S-B correction should be adjusted by hand.

(Table continues)

more, the data are approximately normally distributed, and Mplus is not available, we recommend treating the data as continuous in nature and employing ML estimation. If the variables have five categories or more, the data are non-normally distributed, and Mplus is not available, we recommend S-B scaling methods or bootstrapping. On the other hand, if ordered categorical variables have fewer than five categories, we recommend employing CVM to address both the metric and distribution. We specifically recommend employing CVM using Mplus's robust estimator

Table 9.5. (Continued)

Type of Data	*Suggestions*	*Caveats/Notes*
Ordered Categorical Data		
1. Approximately normally distributed	• Use Mplus's WLSMV estimator • Use ML estimation if there are at least five categories • Use S-B scaling methods if there are at least four categories	• WLSMV will adjust the parameter estimates, standard errors, and fit indices for the categorical nature of the data. • Realize that if employing ML estimation that the parameter estimates will be attenuated. • Parameter estimates from S-B scaling equal ML-based estimates implying that they too will be attenuated.
2. Moderately non-normal (skew < 2, kurtosis < 7)	• Use Mplus's WLSMV estimator • Use ML estimation if there are at least five categories • Use S-B scaling methods if there are at least four categories	• WLSMV will adjust the parameter estimates, standard errors, and fit indices for the categorical nature of the data. • Realize that if employing ML estimation that the parameter estimates will be attenuated. • Parameter estimates from S-B scaling equal ML-based estimates, implying that they too will be attenuated.
3. Severely non-normal (skew > 2, kurtosis >7) or very few categories (e.g., 3)	• Use Mplus's WLSMV estimator • If Mplus is not available then employ S-B scaling method	• Fit indices recommended with WLSMV because they show promise with non-normal ordered categorical data: WRMR and RMSEA. • Realize that S-B correction doesn't correct parameters for attenuation.

WLSMV, because unlike ADF/WLS, it avoids inverting a large asymptotic covariance matrix and has exhibited promising results.

Strategies Not Recommended

An obvious omission from the above recommendations table is the ADF/WLS estimator. Given the requirements of the ADF/WLS estimator (e.g., large sample size) and the lack of sensitivity to model misspecification (e.g., Olsson et al., 2000), we cannot recommend the use of this estimator as a method to analyze non-normal or ordered categorical data. As noted throughout the chapter, other techniques outperform this estimator and should be employed.

A second technique that we cannot recommend, though commonly used to construct "more normally distributed" data (e.g., Marsh, Craven, & Debus, 1991), involves parceling items together (e.g., sum or average a subset of items). For example, parceling items with opposite skew has been conducted in order for the resulting parcel to have a better approximation of a normal distribution. Following the same logic, the parceling of ordered categorical items has been conducted to achieve a more continuous normal distribution allowing for the use of NT methods. While it is true that the parcel may have properties that better approximate the assumptions underlying NT estimators, we cannot recommend the uncritical use of this technique as a strategy to deal with non-normal or categorical data because it results in ambiguous findings. As detailed at length in other sources (Bandalos, 2002, 2003; Bandalos & Finney, 2001), parceling can obscure the true relations among the variables leading to biased parameters estimates and fit indices.

DIRECTIONS FOR FUTURE RESEARCH AND CONCLUSIONS

The purpose of this chapter was to review techniques used to accommodate non-normal and categorical data and summarize previous research investigating their utility. Much of the previous research involving non-normal and/or categorical data was concerned with comparing the performance of different estimation techniques (e.g., ML, WLS) under various conditions such as model size, sample size, and the observed variable distribution characteristics. An appropriate question at this point is where do we go from here with respect to researching the effects of modeling categorical and non-normal data? We believe the most pressing questions concern the functioning of the robust estimators (WLSM and WLSMV) available in Mplus (Muthén & Muthén, 2004). Unlike ML and WLS, lim-

ited research has been conducted that evaluates the performance of these estimators. Additional studies exploring the functioning of WLSM and WLSMV under various conditions and in relation to other techniques are needed in order to better understand the utility of these estimators. Also, very recent advances in software allow categorical dependent variables to be analyzed using ML estimation techniques. Specifically, when using Mplus v3.0 to analyze categorical variables, a full-information ML estimator can be employed. This estimator uses information from the full multiway frequency table of all categorical variables, which is why it is referred to as a "full information" technique. This differs from WLS, which is a "limited-information technique" because it uses bivariate information, or two-way frequency tables between pairs of variables. This full-information estimator uses the two-parameter logistic model, common in item response theory, to describe the variation in the probability of the item response as a function of the factor(s) (L. Muthén, personal communication, November 14, 2003). The availability of the full-information ML estimation technique provides opportunities for new research in the area of analyzing categorical data (e.g., feasibility with large models, comparability of WLSMV-based versus ML-based parameter estimates and standard errors).

In closing, given the presence of non-normal and ordered categorical data in applied research, researchers need to not only recognize the properties of their data but to also utilize techniques that accommodate these properties. Simply using the default estimator from a computer package does not guarantee valid results. Understanding the issues surrounding various techniques, such as assumptions, robustness, and implementation in software programs, makes a researcher much more competent to handle issues that they may encounter.

ACKNOWLEDGMENTS

We thank Deborah Bandalos, Craig Enders, Gregory Hancock, David Kaplan, and especially Linda Muthén, for their helpful comments on this chapter.

APPENDIX A: ADF SYNTAX

The following syntax illustrates how to employ the ADF estimator with continuous and categorical data. The model being estimated is a two-factor model with six indicators per factor. Each observed variable serves as an indicator to only one factor, all error covariances are fixed at zero, and

the factor correlation is freely estimated. The metric of the factor is set by constraining the factor variance to a value of 1.00.

Continuous Data

LISREL v8.54

```
! First run PRELIS to obtain covariance matrix and asymptotic covariance
   matrix
DA NI=12
LA
q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12
RA=example.dat
OR ALL
OU MA=CM SM=example.cov AC=example.acc BT XM
```

SIMPLIS command language employed using LISREL v8.54

```
!Second read matrices into SIMPLIS program
Title illustrating ADF estimation with continuous data
Observed Variables q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12
Covariance matrix from file example.cov
Asymptotic matrix from file example.acc
Sample size 1000
Latent variables: fact1 fact2
Relationships
q1 q2 q3 q4 q5 q6 = fact1
q7 q8 q9 q10 q11 q12 = fact2
Options: WLS
Path Diagram
End of Problem
```

EQS v6.1

```
/TITLE
 illustrating ADF estimation with continuous data
/SPECIFICATIONS
 VARIABLES= 12; CASES= 1000; DATAFILE = 'example.ess';
 MATRIX= raw; METHOD = agls;
 EQUATIONS
 V1 = *F1 + E1;
 V2 = *F1 + E2;
 V3 = *F1 + E3;
```

```
V4 = *F1 + E4;
V5 = *F1 + E5;
V6 = *F1 + E6;
V7 = *F2 + E7;
V8 = *F2 + E8;
V9 = *F2 + E9;
V10 = *F2 + E10;
V11 = *F2 + E11;
V12 = *F2 + E12;
/VARIANCES
F1 to F2 = 1;
E1 to E12 = *;
/COVARIANCES
F2, F1 = *;
/END
```

Mplus v3.01

```
TITLE: illustrating ADF estimation with continuous data
Data: FILE IS example.dat;
VARIABLE: NAMES ARE q1 – q12;
ANALYSIS: ESTIMATOR = WLS;
MODEL: f1 by q1* q2* q3* q4* q5* q6*;
 f2 by q7* q8* q9* q10* q11* q12*;
f1 @ 1;
f2 @ 2;
```

Ordered Categorical Data

LISREL v8.54

```
!PRELIS run to obtain correct correlation matrix and asymptotic covari-
   ance matrix
Title ANALYZING ORDERED CATEGORICAL DATA
DA NI=12
LA
q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12
RA=cat.dat
OR ALL
OU MA=KM SM=poly.cm AC=catex.acc BT XM
```

SIMPLIS command language employed using LISREL v8.54

```
Title ANALYZING ORDERED CATEGORICAL DATA
Observed Variables q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12
Correlation matrix from File poly.cm
Asymptotic matrix from File catex.acc
Sample size 1000
Latent variables: fact1 fact2
Relationships
q1 q2 q3 q4 q5 q6 = fact1
q7 q8 q9 q10 q11 q12 = fact2
Options: WLS
Path diagram
End of problem
```

EQS v6.1

```
/TITLE
 illustrating CVM with categorical data
/SPECIFICATIONS
 VARIABLES= 12; CASES= 1000; DATAFILE = 'example.ess';
 MATRIX= RAW; METHOD = AGLS;
 CATEGORY=V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12;
/EQUATIONS
V1 = *F1 + E1;
V2 = *F1 + E2;
V3 = *F1 + E3;
V4 = *F1 + E4;
V5 = *F1 + E5;
V6 = *F1 + E6;
V7 = *F2 + E7;
V8 = *F2 + E8;
V9 = *F2 + E9;
V10 = *F2 + E10;
V11 = *F2 + E11;
V12 = *F2 + E12;
/VARIANCES
 F1 to F2 = 1;
 E1 to E12 = *;
/COVARIANCES
 F2, F1 = *;
/END
```

Mplus v3.01[5]

```
TITLE: Mplus with ordered categorical data
DATA: FILE IS cat.dat;
VARIABLE: NAMES ARE q1-q12;
 CATEGORICAL ARE q1-q12;
ANALYSIS: ESTIMATOR = WLS;
MODEL: f1 BY q1* q2* q3* q4* q5* q6*;
 f2 BY q7* q8* q9* q10* q11* q12*;
 f1 @1;
 f2 @1;
```

APPENDIX B: S-B SCALING SYNTAX

The following syntax illustrates how to employ the S-B scaling methodology with continuous and ordered categorical data. The model being estimated is a two-factor model with six indicators per factor. Each observed variable serves as an indicator to only one factor, all error covariances are fixed at zero, and the factor correlation is freely estimated. The metric of the factor is set by constraining the factor variance to a value of 1.00.

Continuous and Ordered Categorical Data

SIMPLIS command language employed using LISREL v8.54

```
Title illustrating SB chisq and standard errors
Observed variables q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12
Covariance matrix from file example.cov
Asymptotic matrix from file example.acc
Sample size 1000
Latent variables: fact1 fact2
Relationships
q1 q2 q3 q4 q5 q6 = fact1
q7 q8 q9 q10 q11 q12 = fact2
Options: ML
Path diagram
End of problem
```

EQS v6.1

```
/TITLE
 illustrating SB chisq and standard errors
/SPECIFICATIONS
```

```
 VARIABLES= 12; CASES= 1000; DATAFILE = 'example.ess';
 MATRIX= raw; METHOD = ml, robust;
/EQUATIONS
V1 = *F1 + E1;
V2 = *F1 + E2;
V3 = *F1 + E3;
V4 = *F1 + E4;
V5 = *F1 + E5;
V6 = *F1 + E6;
V7 = *F2 + E7;
V8 = *F2 + E8;
V9 = *F2 + E9;
V10 = *F2 + E10;
V11 = *F2 + E11;
V12 = *F2 + E12;
/VARIANCES
 F1 to F2 = 1;
 E1 to E12 = *;
/COVARIANCES
 F2, F1 = *;
/END
```

Mplus v3.01

```
TITLE: illustrating SB chisq and standard errors
Data: FILE IS example.dat;
VARIABLE: NAMES ARE q1 – q12;
ANALYSIS: ESTIMATOR = MLM;
MODEL: f1 by q1* q2* q3* q4* q5* q6*;
       f2 by q7* q8* q9* q10* q11* q12*;
       f1 @ 1;
       f2 @ 1;
```

APPENDIX C: ROBUST WLS (WLSM, WLSMV) SYNTAX

The following syntax illustrates how to employ the robust estimation techniques using Mplus v2.11. The model being estimated is a two-factor model with six indicators per factor and the factor correlation is freely estimated. The metric of the factor is set by constraining the factor variance to a value of 1.00. It is noted that indicator error variance terms are not estimated in Mplus when indicators are identified as categorical. To employ either WLSM or WLSMV, simply type "WLSM" or "WLSMV" on

the command line that specifies the estimation method. In the current example, WLSMV would be employed.

```
TITLE:    MPLUS with ordered categorical data – robust estimation
          procedures
DATA:     FILE IS cat.dat;
VARIABLE: NAMES ARE q1-q12;
          CATEGORICAL ARE q1-q12;
ANALYSIS: ESTIMATOR=WLSMV;
MODEL:    f1 by q1* q2* q3* q4* q5* q6*;
          f2 by q7* q8* q9* q10* q11* q12*;
          f1 @ 1;
          f2 @ 1;
```

APPENDIX D: BOOSTRAPPING SYNTAX

The following syntax illustrates how to employ the bootstrapping technique. The model being estimated is a two-factor model with six indicators per factor. Each observed variable serves as an indicator to only one factor, all error covariances are fixed at zero, and the factor correlation is freely estimated. The metric of the factor is set by constraining a path from a factor to an indicator equal to a value of 1.00 (see Hancock & Nevitt, 1999, for an explanation of why it is necessary).

EQS v6.1
Naive Bootstrap Standard Errors

```
/TITLE
naive bootstrap
/SPECIFICATIONS
VARIABLES= 12; CASES=1000; DATAFILE='example.ess';
MATRIX=RAW; METHOD=ML;
/EQUATIONS
V1 = 1F1 + E1;
V2 = *F1 + E2;
V3 = *F1 + E3;
V4 = *F1 + E4;
V5 = *F1 + E5;
V6 = *F1 + E6;
V7 = 1F2 + E7;
V8 = *F2 + E8;
V9 = *F2 + E9;
```

```
V10 = *F2 + E10;
V11= *F2 + E11;
V12 = *F2 + E12;
/VARIANCES
 F1 to F2 = *;
 E1 to E12 = *;
/COVARIANCES
 F1, F2 = *;
/technical              !itr increases number of iterations to increase the
itr = 500;                bootstrap success rate
/SIMULATION             !The keyword bootstrap indicates the naïve boot-
bootstrap = 1000;         strap
replication = 250;      !1000 refers to the number of cases in the bootstrap
seed = 123456789;         samples
/OUTPUT                 !Number of replications equals 250
parameters;             !Default seed for the random number generator
/END                    !The output will contain the mean parameter esti-
                          mates and the standard
                        !deviations, which are the empirical standard errors
```

Bollen-Stine Bootstrap χ^2 probability value

```
/TITLE
   Bollen-Stine bootstrap chisq probability value
/SPECIFICATIONS
 VARIABLES= 12; CASES=1000; DATAFILE='example.ess';
 MATRIX=RAW; METHOD=ML;
/EQUATIONS
V1 = 1F1 + E1;
V2 = *F1 + E2;
V3 = *F1 + E3;
V4 = *F1 + E4;
V5 = *F1 + E5;
V6 = *F1 + E6;
V7 = 1F2 + E7;
V8 = *F2 + E8;
V9 = *F2 + E9;
V10 = *F2 + E10;
V11= *F2 + E11;
V12 = *F2 + E12;
/VARIANCES
F1 to F2 = *;
```

```
E1 to E12 = *;
/COVARIANCES
F1, F2 = *;
/technical
itr = 500;          !itr increases the number of iterations
/SIMULATION         !The keyword mbb indicates the model-based, or
mbb = 1000;           Bollen-Stine, bootstrap
replication = 250;  ! 1000 refers to the number of cases in the boot-
seed = 123456789;     strap samples
/OUTPUT             !Number of bootstrap samples (B) drawn equals
parameters;           250
/END                !Default seed for the random number generator
                    !Output presents information concerning the
                      empirical distribution of
                    !the model-based chi-square values including the
                      value that represents
                    !the upper 5% of the distribution, which can be
                      used as the critical chi-!square value to assess sig-
                      nificance of the ML-based chi-square
```

Mplus v3.01

```
Naive Bootstrap Standard Errors and Confidence Intervals
TITLE:      MPLUS with naive bootstrap standard errors and CI
DATA:       FILE IS example.dat;
VARIABLE:   NAMES ARE q1-q12;
ANALYSIS:   BOOTSTRAP = 250;          !Number of bootstrap sam-
                                        ples (B) drawn = 250;
MODEL:      f1 BY q1 q2 q3 q4 q5 q6; !  !The size of the B samples=
                                        size of original sample;
            f2 BY q7 q8 q9 q10 q11 q12; !Other sample sizes of B can-
                                        not be specified;
OUTPUT:     CINTERVAL;
```

NOTES

1. Technically, this gives the reweighted least squares fit function, which is asymptotically equivalent to ML's well-known fit function,

   $F = \log|\Sigma(\hat{\theta})| + tr[S\Sigma(\hat{\theta})^{-1}] - \log|S| - p$, where p equals the number of observed variables.

2. The standard linear measurement model specifies that a person's score is a function of the relation (b) between the variable (y^*) and the factor (F) plus error (E): $y^* = bF + E$

3. This formulation is equivalent to the two-parameter normal ogive model (probit model) applied to dichotomous items in item response theory (Muthén & Asparouhov, 2002; Thissen & Orlando, 2001), $P(y \geq k|\mathrm{F}) = \Phi[a\ (\mathrm{F} - b_k)]$, where a is the item discrimination and b is the item difficulty.
4. The formula used to calculate these scaled standard errors is complex but can be found in Arminger and Schoenberg (1989) and Satorra and Bentler (1994).
5. A unique feature of Mplus concerns the indicator error variance terms. Whereas LISREL and EQS allow estimation of these parameters, Mplus does not estimate indicator error variance terms if indicators are identified as categorical. Muthén and Muthén (2001) state that this is related to the use of a correlation matrix as input for analyses. With correlation input, the diagonal elements (values of 1) do not enter into the computations. Values related to each item variance are considered by Mplus to be "residual correlations" rather than as item variance terms.

REFERENCES

Arbuckle, J., & Wothke, W. (1999). *AMOS 4.0 user's guide*. Chicago: Smallwaters Corporation.

Arminger, G., & Schoenberg, R. (1989). Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika, 54*, 409–425.

Austin, J. T., & Calderón, R. F. (1996). Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 105–175.

Babakus, E., Ferguson, C. E., & 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*, 2228.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 78–102.

Bandalos, D. L. (2003, April). *Identifying model misspecification in SEM analyses: Does item parceling help or hinder?* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Erlbaum.

Bentler, P. M. (1990). Comparative fit indexes in structural equation models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. (1995). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M. (2004). *EQS for Window* (Version 6.1) [Computer software]. Encinco, CA: Multivariate Software, Inc.

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

Bentler, P. M., & Wu, E. J. C. (2002). *EQS for Windows user's guide*. Encino, CA: Multivariate Software, Inc.

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

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21*, 205–229.

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

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

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. 37–55). 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.

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

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 327–346.

DiStefano, C. (2003, April). *Considering the number of categories and item saturation levels with structural equation modeling*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

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.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63*, 24–50.

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: A Multidisciplinary Journal, 4*, 87–107.

Fouladi, R. T. (1998, April). *Covariance structure analysis techniques under conditions of multivariate normality and non-normality—Modified and bootstrap based test statistics*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Gierl, M. J., & Mulvenon, S. (1995, April). *Evaluating the application of fit indices to structural equation models in educational research: A review of the literature from 1990 through 1994*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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 confir-

matory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 4*, 108–120.

Hancock, G. R., & Nevitt, J. (1999). Bootstrapping and the identification of exogenous latent variables within structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 394–399.

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

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55.

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

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 344–364.

Jöreskog, K., & Sörbom, D. (1996). *PRELIS 2: User's reference guide*. Chicago: Scientific Software International.

Jöreskog, K., & Sörbom, D. (2004). *LISREL 8.54 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Marsh, H. W., Craven, R. G., & Debus, R. (1991). Self-concepts of young children 5 to 8 years of age: Measurement and multidimensional structure. *Journal of Educational Psychology, 83*, 377–392.

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., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved March 15, 2004, from http://www.statmodel.com/mplus/examples/webnotes/CatMGLong.pdf

Muthén, B. O., du Toit, S., & 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. 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., & 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, L. (1999, August 10). Mplus and non-normal data. Message posted to SEMNET discussion list, archived at http://bama.ua.edu/cgi-bin/wa?A2=ind9908&L=semnet&D=0&T=0&P=9802

Muthén, L. (2003, April 29). Mplus and other questions. Message posted to SEMNET discussion list, archived at http://bama.ua.edu/cgi-bin/wa?A2=ind0304&L=semnet&D=0&I=1&P=52337.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide* (2nd ed). Los Angeles: Authors.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed). Los Angeles: Authors.

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for non-normal 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: A Multidisciplinary Journal, 8*, 353–377.

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

Olsson, U., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and non-normality. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 557–595.

Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretical fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavioral Research, 34*, 31–58.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology, 46*, 273–286.

Rosenberg, M. (1989). *Society and the adolescent self-image* (Rev. ed.). Middletown, CT: Wesleyan University Press.

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.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Erlbaum.

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 93–104.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. 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.