The Dissertation Committee for Tiffany Ann Whittaker
certifies that this is the approved version of the following dissertation:

The Performance of Cross-Validation Indices Used to Select Among

Competing Covariance Structure Models

Committee:

Laura M. Stapleton, Supervisor

Barbara G. Dodd

Susan N. Beretvas

William R. Koch

Steven J. Fitzpatrick

The Performance of Cross-Validation Indices Used to Select Among Competing Covariance Structure Models

by

Tiffany Ann Whittaker, B.A., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2003

Dedicated to my parents

Harold and Sandra Whittaker

Acknowledgments

I would first like to extend my sincerest gratitude to my dissertation advisor Laura Stapleton, not only for her much valued support and guidance throughout the entire dissertation process, but also for her encouragement and optimism during the difficult times. I would also like to thank Laura for the excellent Structural Equation Modeling course she taught, without which I would not have been able to develop a dissertation topic, much less complete this dissertation.

I would also like to acknowledge the additional members of my dissertation committee. Thanks are extended to Barbara Dodd for her helpful insights and her confidence in my abilities during my four years in the doctoral program in Educational Psychology. Thanks are extended to Tasha Beretvas for her enthusiasm when learning and teaching and her willingness to help me regardless of her many other commitments and busy schedule. Thanks are extended to Bill Koch for his invaluable advice during my fours years at the University of Texas and his help during the prospectus process, which prepared me well for the dissertation process. Thanks are extended Steve Fitzpatrick for the excellent SAS course he taught and the opportunity to develop my programming skills at the Measurement and Evaluation center, without which I would not have been able to conduct this study.

I would like to acknowledge Natasha Williams for her friendship and endless support during my graduate career, and for her reassurance during the times of self-doubt. I would like to thank Natasha Williams, Aimee Boyd, and Laurie Davis for kindly allowing me to use their computers to run several conditions for this simulation study. I would also like to thank Carolyn Furlow for her friendship and comic relief during the dissertation process.

Finally, I would like to extend my deepest gratitude to my parents, Harold and Sandra Whittaker, for teaching me by example to strive to achieve and for giving me the courage to attain my goals while welcoming the challenges along the way.

The Performance of Cross-Validation Indices Used to Select Among Competing Covariance Structure Models

Publication No._____

Tiffany Ann Whittaker, Ph.D.

The University of Texas at Austin, 2003

Supervisor: Laura M. Stapleton

When testing structural equation models, researchers attempt to establish a model that will generalize to other samples from the same population. Unfortunately, researchers tend to test and respecify models during this attempt, capitalizing on the characteristics inherent within the sample data in which the model is being developed. Several measures of model fit exist to aid researchers when trying to select a model that fits the sample data well. However, these measures fail to consider the predictive validity of a model, or how well it will generalize to other samples from the same population.

In 1983, Cudeck and Browne proposed using cross-validation as a model selection technique. They recommended cross-validating several plausible models and selecting the model with the most predictive validity. Several cross-validation indices have been proposed in the past twenty years, including the single-sample AIC,

CAIC, and BCI; the multiple-sample *C*; the two-sample CVI; and the "pseudo" single-sample *C** and $\overline{C}$ *. Previous studies have investigated the performance of these various indices, but have been limited with respect to their study design characteristics. The purpose of this study is to extend the literature in this area by examining the performance of the previously mentioned cross-validation indices under additional study design characteristics, such as nonnormality and cross-validation design. Factor loading, sample size, and model misspecification conditions were also manipulated.

The performance of each cross-validation index was measured in terms of how many times out of 1,000 replications they selected the correct confirmatory factor model. The results indicated that the performance of the cross-validation indices tended to improve as factor loading and sample size increased. The double cross-validated indices outperformed their simple cross-validated counterparts in certain conditions. The performance of the cross-validation indices tended to decrease as nonnormality increased. Recommendations are provided as to which cross-validation methods would optimally perform in a given condition. It is hoped that this study provides researchers with useful information concerning the use of cross-validation as a model selection technique and that researchers will begin to focus on the predictive validity of a structural equation model in addition to overall model fit.

Table of Contents

List of Tables

Chapter I

Introduction

Cross-validation originated within the realm of multiple regression analysis as a method to ensure that multiple regression equations developed within one sample would be applicable within other samples from the same population with an unknown criterion variable. Before cross-validation methods were employed in multiple regression, multiple regression equations developed within one sample demonstrating adequate fit would typically demonstrate less than adequate fit in other samples from the same population, much to the dismay of researchers. The estimation procedure used in multiple regression analysis tends to capitalize on chance occurrences in the sample data set in which the multiple regression equation is developed. Consequently, the multiple regression equation developed within the sample would indicate better fit than would be found in other sample data sets collected separately according to typical predictive validity measures (R^2 and the mean square error of prediction). As a result, cross-validation was suggested as a technique to better estimate the fit of a multiple regression equation in other samples from the same population and was advocated as a model selection technique.

The use of cross-validation as a model selection technique has extended beyond the realm of multiple regression analysis into other areas, such as covariance structure analysis. Several statistical techniques are encompassed within the classification of covariance structure analysis, such as path analysis, structural equation modeling, and confirmatory factor analysis. In covariance structure

analysis, the covariances among observed variables are analyzed to determine how well hypothesized models account for these covariances. Similar to model testing in multiple regression analysis, the testing and respecification of structural equation models within a sample tends to capitalize on the idiosyncrasies of the sample in which the structural equation model is being developed. While several measures of model fit in structural equation modeling exist, they do not reflect the predictive accuracy of structural equation models in other samples from the same population. In fact, the measures of model fit within structural equation modeling only reflect hypothesis tests of exact fit of the model to the data.

Cudeck and Browne (1983) rebuke the stringent test of exact fit typically relied upon within the context of structural equation modeling. Instead, they advise that structural equation models are unknown in practice and that hypothesized models only represent approximations of the truth. Cudeck and Browne have recommended using cross-validation as a method to select from among several structural equation models that are theoretically reasonable. Thus, several models are tested and selected in terms of their predictive accuracy.

Several cross-validation indices have been proposed in the literature, including two-sample and single-sample measures. In general, research has found that single-sample cross-validation indices perform comparably to, if not better than, two-sample cross-validation indices under certain conditions when selecting the correct confirmatory factor model (Bandalos, 1993; De Gooijer & Koopman, 1988). Several variables have been found to affect the performance of the cross-validation

indices, such as sample size, factor loading size, and model misspecification (Bandalos, 1993; De Gooijer & Koopman, 1988). However, the studies in this area are limited with respect to the types of structural equation models investigated in that they are not realistic representations of models typically tested in practice. Moreover, the studies have not examined all of the proposed cross-validation indices that have demonstrated competitive performance. To date, research in this area has not examined the impact of nonnormality among the observed variables on the performance of cross-validation indices. Further, most studies have typically employed the cross-validation design necessary to calculate two-sample cross-validation indices when calculating proposed single-sample measures, which contradicts the purpose of single-sample cross-validation indices.

The purpose of this study is to extend the literature in this area by examining the performance of a comprehensive list of proposed cross-validation indices. The performance of the cross-validation indices was documented in terms of selecting the correct confirmatory factor model under similar conditions investigated in previous studies, such as sample size, factor loading size, and model misspecification. The current study, however, extends previous research by including additional conditions, such as nonnormality and cross-validation design conditions.

It is hoped that this study provides further knowledge concerning the use of cross-validation indices under various realistic conditions. In particular, the extended range of model misspecification, which is commonly encountered when selecting models, allows researchers to be aware of these effects and to thus base model

selection on both theoretical and statistical criteria. In addition, the performance of cross-validation indices under conditions of multivariate nonnormality, which is also commonly encountered in applied settings, may also provide researchers with a solution as to which cross-validation index will perform more consistently under nonnormality.

Chapter II

Literature Review

The present study examined the performance of cross-validation indices when selecting among competing covariance structure models. However, the cross-validation procedure originated as a method to determine the predictive accuracy of multiple regression equations. Thus, it is important to understand the history and the use of cross-validation in multiple regression analysis in order to better understand its application in the realm of covariance structure analysis.

This chapter begins with a brief clarification of the cross-validation design and the designs which are mistakenly referred to as cross-validation. An introduction to cross-validation techniques used in multiple regression and subset selection techniques used in multiple regression is then presented. This is followed by an introduction to covariance structure analysis. Typical model modification and selection techniques in covariance structure analysis are reviewed. Finally, cross-validation methods used to select among competing structural equation models are explained.

Clarification of the Cross-Validation Design

Although Mosier (1951) distinguished the differences between the cross-validation design and other similar designs a little more than fifty years ago, confusion about the true cross-validation design still exists. Mosier (1951) defined five different experimental designs, including cross-validation, validity-generalization, validity-extension, simultaneous validation, and replication. These

five methods are similar in terms of using separate samples to estimate regression coefficients, however, they differ in terms of the types of samples used and their purpose.

In the true cross-validation design, regression coefficients are estimated in one sample and their efficiency is tested on a second sample from the same population with the purpose of determining how accurate the regression equation will be in the relevant population. In the validity-generalization design, the purpose is to determine whether an accurate regression equation for one population will generalize to a different population. Thus, regression coefficients are calculated in one sample (e.g., military ROTC cadets) and applied to a second sample from a different population (e.g., civilian ROTC cadets) to test their efficiency. Validity-extension is the same as validity-generalization except that the criterion measure in each sample is different. In simultaneous validation, regression coefficients are determined in two samples from different populations (e.g., male versus female) in order to develop a regression equation that will be efficient in either population, where only valid predictors for both samples are retained in the final model. In the replication design, which is probably the most frequently used term to refer to actual cross-validation designs, regression coefficients are sequentially estimated in a set of samples from the same population. The final regression equation is determined by the consistent performance of predictors across these samples in terms of their standardized regression coefficients. For example, to retain a predictor in the multiple regression model, the standardized regression coefficient associated with the predictor in the

equation must exceed a specified value in a designated percentage of all the replication samples (Mosier, 1951).

The term “cross-validation” in the current paper will refer to the cross-validation design described by Mosier (1951). Variants of the true cross-validation design will also be discussed in the current study. These variant designs are still considered to fall under the rubric of cross-validation in which the purpose is to determine the predictive accuracy of a model in the relevant population.

Multiple Regression Analysis

Cross-validation methods were originally employed within the context of multiple regression analysis. Multiple regression analysis is used to predict a dependent or criterion variable from two or more independent or predictor variables. Consider the sample regression model for n observations with p independent variables:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e},$$

where **Y** is an n x 1 column vector containing data on the dependent variable (criterion);

X is an n x (p + 1) matrix containing data on the predictor (independent) variables;

B is a (p + 1) x 1 column vector containing estimated regression coefficients; and **e** is an n x 1 column vector containing error terms (residuals).

One of the main purposes of multiple regression analysis is to establish a regression equation that can be applied in other samples to predict an unknown

criterion variable with known data on the predictor variables. When conducting multiple regression analysis, regression coefficients (**B**) are typically estimated using the Ordinary Least Squares (OLS) method, which finds values of the regression coefficients that provide the most accurate fit of the data to the model by minimizing prediction error (residuals),

$$\mathbf{e} = \mathbf{Y} - \mathbf{XB} = \mathbf{Y} - \hat{\mathbf{Y}},$$

where $\hat{\mathbf{Y}}$ is an n x 1 column vector containing predicted Y values for the sample. OLS provides optimal, unbiased regression coefficients when the following assumptions are met: (1) lack of measurement error; (2) the residuals have a mean of zero; (3) the residuals have constant variance; (4) the residuals are not intercorrelated; and (5) the residuals are normally distributed (Hamilton, 1992).

After establishing the regression equation and obtaining predicted Y values ($\hat{\mathbf{Y}}$) in a sample for which data on both the predictor and criterion variables are available, the sample squared multiple correlation may be used to determine the proportion of variance in the criterion variable that is accounted for by the predictor variables and is an estimate of the population squared multiple correlation (ρ^2). The squared multiple correlation is simply the squared zero-order correlation between the observed criterion variables (**Y**) and the predicted Y values determined for the sample ($\hat{\mathbf{Y}}$): R^2 (Pedhazur, 1982). In the past, the squared multiple correlation was used as a measure of predictive accuracy, or how well the sample equation would perform in other representative samples from the population, with higher values indicating

higher predictive accuracy. The sample mean square error of prediction (MSEP) was also used as a measure of predictive accuracy, which is the mean squared difference between the criterion variables (**Y**) and the predicted *Y* values from the sample ($\hat{\mathbf{Y}}$):

$$\text{MSEP} = \sum \left(Y - \hat{Y}\right)^2 / (n - p - 1).$$

The MSEP is an estimate of the population MSEP (σ_e^2), with lower values indicating less predictive error, and, thus, better predictive accuracy of the regression model. Unfortunately, when researchers would apply regression equations with the fixed regression coefficients established in one sample to other data, they would typically find that the squared multiple correlation was smaller than that found in the original sample and that the MSEP was underestimated in the original sample (Mosteller & Tukey, 1977).

Finding smaller squared multiple correlations in samples other than the one in which the regression equation was developed is referred to as shrinkage and may be estimated by the difference between the multiple squared correlation obtained in the original sample and those obtained from other samples (Pedhazur, 1982). Shrinkage occurs because the model developed in a sample is dependent upon the idiosyncrasies of that particular sample. Thus, the regression model will reflect the characteristics of that particular sample and not necessarily reflect the characteristics of the population or samples from the population. Consequently, the regression model developed in a sample will always overestimate the fit of the model and will fit better in that particular sample than in other samples. The problem of fit overestimation is due to

OLS regression estimation, which produces optimal regression coefficients that provide the most accurate fit of the model to the data. Hence, it "capitalizes on chance" occurrences within that particular data set (Mosteller & Tukey, 1977, p. 37). Because the sample squared multiple correlation and the sample MSEP are optimistic measures of the sample regression model's validity in the population, Browne (1975a, 1975b) proposed two measures of predictive validity for the sample equation in the population which are analogous to the squared multiple correlation and the MSEP, denoted as w^2 and d^2, respectively.

Cross-Validation Methods

In order to obtain estimates of these measures (w^2 and d^2), cross-validation procedures must be employed. The most commonly employed cross-validation procedure, called simple cross-validation, begins with two subsamples, *a* and *b*, which are randomly selected from the same sample. Sample *a* is used as the calibration sample and sample *b* as the validation sample. The regression model is fitted to the data from the calibration sample, producing estimated regression coefficients. These estimated regression coefficients are employed as fixed values in the regression model, which is then applied to the validation sample data, producing predicted *Y* values for the validation sample ($\hat{Y}_{b|a}$). The squared zero-order correlation is calculated between the criterion variables from the validation sample (Y_b) and the predicted *Y* values from the validation sample ($\hat{Y}_{b|a}$):

$$\hat{w}^2 = r^2_{Y_b \hat{Y}_{b|a}},$$

and is employed as a measure of predictive accuracy of the regression model developed in the calibration sample. This measure represents how accurate the regression model will be in other samples, with higher values representing higher predictive validity of the model. The value of this measure is typically smaller than the squared multiple correlation obtained within the calibration sample (Pedhazur, 1982). The MSEP, or $\hat{d}^2$ as denoted by Browne (1975b), may also be used to determine predictive accuracy in the cross-validation paradigm by calculating the mean squared difference between the criterion variable (Y_b) and the predicted Y values from the validation sample ($\hat{Y}_{b|a}$):

$$\hat{d}^2 = \sum \left(Y_b - \hat{Y}_{b|a}\right)^2 / n_b .$$

Again, lower values of the MSEP ($\hat{d}^2$) indicate better predictive accuracy of the regression model.

Ideally, the validation sample is a sample in which data are collected separately from the data collected in the calibration sample. Cross-validation of a multiple regression model on data which are collected separately from the calibration data would alleviate the problem of capitalizing on chance occurrences that may have occurred during the data collection process. Unfortunately, collecting new data is not always feasible, and may result in lengthy delays of the assessment of the multiple regression equation (Pedhazur, 1982). Thus, the most typical cross-validation methods include randomly splitting an available sample of data in half (Snee, 1977). Splitting the sample to yield both the calibration and validation

samples is a serious drawback of the simple cross-validation procedure, especially for smaller samples. The drawback is the loss of power when estimating the regression coefficients, which are not as precise as they would be if the entire sample of data were used when determining the regression model (Mosier, 1951). The lack of precision in the regression coefficients is due to calculating the standard errors of the regression coefficients on only half the sample data, with precision deteriorating as sample size decreases. Nonetheless, using the entire sample to determine the regression model would not allow for cross-validation (Picard & Cook, 1984).

Another type of cross-validation procedure, termed double cross-validation, has been discussed by Mosier (1951) to overcome the wasteful nature of simple cross-validation in terms of estimating regression coefficients on only half of the data. In double cross-validation, regression coefficients are obtained in both subsamples separately. The regression coefficients are then employed as fixed values and applied to the opposing subsample's data. The predictive accuracy is then calculated to indicate the model's predictive ability within each of the subsamples. Because the goal is to determine the predictive accuracy of the regression model, optimal regression coefficients may be estimated using the entire sample, given that each of the regression equations in the double cross-validation procedure cross-validated well in the opposing subsample. The regression coefficients determined using the entire sample would then be applied to other samples in which data on the criterion variable are not available. There would not be an estimate of predictive accuracy for the entire sample in this case since the estimation was based upon the entire sample, but the

average of the predictive accuracy coefficients obtained from both subsamples may be used as an estimate.

Other ways of utilizing all of the data in the entire sample have also been proposed. If the entire sample is large enough, cross-validation may be employed with more than one validation sample. For instance, the calibration sample can be used to determine the regression coefficients and then cross-validated on two or more validation samples (Geisser, 1975). Another sampling procedure used in cross-validation, similar to the jacknife sampling technique (Efron, 1982), is the "leave-one-out" method in which one observation is put aside and the regression coefficients are determined on the $n - 1$ remaining observations. The coefficients are then applied to the "left out" observation to determine predictive accuracy. This is then repeated for every observation, leaving out a different observation each time until all observations have been left out and tested with the regression coefficients from the $n - 1$ remaining observations (Camstra & Boomsma, 1992). The prediction error sum of squares (PRESS) statistic (Allen, 1971; 1974), which was suggested as an aid in model selection, is an alternate estimator of the MSEP and may be calculated in the "leave-one-out" cross-validation method and is the sum of the squared prediction errors:

$$\text{PRESS} = \sum (Y_i - \hat{Y}_i)^2,$$

where Y_i is the criterion variable for the observation "left out" and $\hat{Y}_i$ is the predicted Y value for that observation obtained from using the regression coefficients determined

on the $n - 1$ remaining observations. The average of the PRESS statistic over every observation is then calculated as an estimate of predictive accuracy, with smaller values indicating greater accuracy.

Due to the problems associated with splitting samples for cross-validation, alternative single sample cross-validation indices, without any type of data splitting, have been advocated as estimates of predictive accuracy. To estimate the squared zero-order correlation between the criterion variable from the validation sample and the predicted Y values from the validation sample (w^2) without using a validation sample, Browne (1975a) derived a single sample estimate of the expected value of w^2 which would be obtained over all possible random calibration samples:

$$\hat{\omega}^2 = \frac{(N-p-3)\hat{\rho}^4 + \hat{\rho}^2}{(N-2p-2)\hat{\rho}^2 + p},$$

where

$$\hat{\rho}^2 = \max(0, r^2 - \frac{p(1-r^2)}{N-p-1}$$

and

$$\hat{\rho}^4 = \max\left(0, (\hat{\rho}^2)^2 - \frac{2p(1-\hat{\rho}^2)^2}{(N-1)(N-p-1)}\right)$$

are less biased estimates of the model's squared multiple correlation in the population. Similarly, Browne (1975b) derived a single sample estimate of the expected value of d^2 (the mean square error of prediction):

$$\hat{\delta}^2 = \frac{(N+1)(N-2)}{(N-1)(N-p-2)} s_e^2,$$

where s_e^2 is the sample mean square error of prediction. The single sample estimates of predictive validity proposed by Browne (1975a, 1975b) assume that the predictor variables are random and are multivariate normal.

Model Selection

The cross-validation methods discussed above assume that the multiple regression model has already been selected based upon substantive theory (Camstra & Boomsma, 1992). That is, cross-validation methods are intended to determine the predictive accuracy of one predetermined multiple regression model. Realistically, predetermined multiple regression models, even those based on substantive theory, may not demonstrate reasonable squared multiple correlations or MSEPs in the calibration data, yielding cross-validation a moot point. Further, researchers may not have determined the multiple regression model in advance. Instead, a researcher may have collected data on several predictor variables in order to determine which subset model predicts the criterion the best in order to implement this multiple regression equation for future prediction purposes (Weisberg, 1985). Hence, subset selection of predictor variables from a larger set of predictor variables will most likely ensue in order to obtain a multiple regression model that fits the data adequately. While the most important approach in model selection is the researcher's knowledge of the variables under study, the use of statistical techniques in determining the fit of various subset models is necessary.

Subset selection is sometimes accomplished using standard search procedures frequently included in statistical software packages (e.g., SPSS, SAS), such as backward, forward, and stepwise selection techniques. Aside from the lack of the researcher's input when using these selection techniques, other problems often arise, such as increased Type I error rates for significance tests associated with the regression coefficients and the omission of intercorrelated variables which may be theoretically important (Hamilton, 1992).

Subset selection may also be achieved by comparing various criterion measures obtained for different subset models. One commonly used criterion measure is the adjusted squared multiple correlation (R^2_{adj}):

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\left(1-R^2_j\right),$$

where R^2_j is the squared multiple correlation for subset model *j* (Neter & Wasserman, 1974). The subset model with the highest R^2_{adj} value is selected as the "best fitting" model. Another criterion measure used to select among competing subset models is Mallows' C_p statistic (Mallows, 1973):

$$C_p = \frac{\text{SSE}_j}{\text{MSEP}} + 2(p+1) - n,$$

where SSE_j is the sum of squared errors, $\Sigma\left(Y - \hat{Y}\right)^2$, for subset model *j* and MSEP is the mean square error of prediction for the full model including all relevant predictor

variables. According to the C_p statistic, the best fitting model will have the smallest positive value of:

$$p + 1 - C_p.$$

If none of the subset models presents a positive value of this difference, the full model is selected as the best fitting model, in which $p + 1 - C_p = 0$ (Weisberg, 1985). Akaike's Information Criterion (AIC; Akaike, 1973) is also used as a criterion measure:

$$\mathrm{AIC} = n \ln\left(\frac{\mathrm{SSE}_j}{n}\right) + 2(p+1),$$

where ln is the natural log. The subset model with the lowest AIC value is considered the best fitting model. The MSEP is also used to select among competing subset models, with the smallest MSEP value indicating the best fit (Weisberg, 1980).

Cross-Validation and Model Selection

All of these criterion measures contain a penalty for the number of predictor variables included in the multiple regression model, indicating that the inclusion of too many predictors in the model may decrease precision in prediction. This penalty illustrates one of the goals of subset selection, which is parsimony (Hamilton, 1992). Unfortunately, exploratory subset selection that is performed on the data in which regression coefficients are estimated introduces the common problem of capitalizing on chance. Interestingly, model selection techniques have typically been discussed separately from cross-validation methods. Model selection techniques are generally discussed with a final recommendation to cross-validate the selected subset model to

determine its predictive accuracy. These two processes should not be considered separately, but instead should be considered simultaneously when validating models (Snee, 1977).

The suggestion to employ cross-validation methods in the model selection process is not surprising considering the final recommendation to cross-validate the model once it is selected. For instance, if the selected model did not cross-validate well, the search for another subset regression model with adequate fit would begin again. The selected model from the search would then be cross-validated. Depending upon the model's predictive accuracy, this cycle may stop or continue again until a valid model was selected. In theory, then, cross-validation methods, if employed after model selection, are involved in the model selection process. An illustration of the application of cross-validation methods in the model selection process can be seen in Snee (1977).

Employing cross-validation methods in the model selection process has not been formally outlined. Following the procedure in Snee's (1977) study and other suggestions (e.g., Camstra & Boomsma, 1992), the basic procedure would entail:

1. The selection of competing multiple regression models with good fit in the calibration sample as determined by a primary criterion (e.g., AIC, Mallows' C_p, R^2_{adj}, MSEP);
2. The application of the estimated coefficients for each competing model obtained from the calibration sample in the validation sample;

3. Selection of the multiple regression model with the best predictive accuracy as determined by a cross-validation index (e.g., $\hat{d}^2$, $\hat{w}^2$).

Variants of the simple cross-validation method (i.e., splitting the data in half) may also be employed, such as the double cross-validation method, the "leave-one-out" method, or the single-sample estimates of predictive accuracy. Similar to the criterion measures used in subset selection, cross-validation indices also include a penalty for the number of predictor variables included in the model and will aid in the selection of a parsimonious model with predictive validity.

Covariance Structure Analysis

The evolution of cross-validation methods employed in multiple regression analysis has foreshadowed the application of cross-validation methods in other analyses, such as covariance structure analysis. In covariance structure analysis, also called structural equation modeling, procedures are performed on the covariances among observed variables. The underlying principle in covariance structure analysis is to minimize the discrepancy between the covariances among observed variables in a sample and the covariances implied by the hypothesized structural model. While the purpose of multiple regression is prediction, the purpose of structural equation modeling is model or theory testing. More specifically, structural equation models are tested to determine how well they account for the covariances among observed variables. Consider the basic equation used in structural equation procedures:

$$\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix for *p* observed variables;

$\boldsymbol{\theta}$ is the vector containing model parameters;

and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix implied by the function of model parameters ($\boldsymbol{\theta}$) (Bollen, 1989).

In applications of structural equation modeling, the population covariance matrix ($\boldsymbol{\Sigma}$) is unknown and is estimated by the sample covariance matrix (**S**). The unknown model parameters ($\boldsymbol{\theta}$) are also estimated ($\hat{\boldsymbol{\theta}}$) through a minimization process of a discrepancy function:

$$F[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})].$$

Substituting the estimates of the unknown model parameters in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ results in the implied covariance matrix, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$. The unknown model parameters are estimated to reduce the discrepancy between the implied covariance matrix and the sample covariance matrix. An indication of the discrepancy between the sample covariance matrix and the implied covariance matrix may be deduced from the residual matrix:

$$(\mathbf{S} - \hat{\boldsymbol{\Sigma}}),$$

with values closer to zero indicating better fit of the structural model to the data (Bollen, 1989).

To estimate a unique set of model parameters, the structural model must be identified. A simple definition of identification involves calculating the difference

between the number of non-redundant observations in the covariance matrix (p^*) with p observed variables,

$$p^* = \frac{p(p+1)}{2},$$

and the number of model parameters (q) that must be estimated. The model parameters to be estimated consist of the number of variances and covariances of observed or latent exogenous variables and the direct effects (paths) on the endogenous variables. Exogenous variables may be thought of as independent variables that influence other variables in the model whereas endogenous variables may be thought of as dependent variables, but may also influence other endogenous variables in the model (Kline, 1998). For a more detailed explanation of identification, see Bollen (1989). A just-identified model contains the same number of non-redundant observations in the covariance matrix as model parameters and is often referred to as the saturated model. An over-identified model contains more non-redundant observations in the covariance matrix than model parameters. Under-identified models (i.e., more model parameters than observations) are mathematically impossible to analyze because there are an infinite set of solutions that will satisfy the structural model, which makes estimation of a unique set of model parameters unattainable (Kline, 1998).

To better illustrate structural equation modeling, consider the confirmatory factor analysis model in Appendix A. According to this model, four indicators, or observed endogenous variables, (x_1 to x_4) measure a single latent exogenous

(independent) factor (ξ_1). The errors of measurement associated with the observed variables (δ_1 to δ_4), which are also exogenous (independent) variables, are not intercorrelated in this particular model, however, structural equation modeling will allow errors to be intercorrelated if deemed necessary. The errors are assumed to have expected values equal to zero [$E(\delta_i) = 0$] and to be uncorrelated with the latent factor (ξ_1). The parameters (λ_1 to λ_4), or direct effects (paths), indicate the causal relationship between the latent exogenous factor (ξ_1) and the observed endogenous variables (x_1 to x_4). More specifically, every one-unit change in the latent factor (ξ_1) will lead to a change in the observed variable (e.g., x_1) of λ_1 (Bollen, 1989).

The model in Appendix A contains 10 non-redundant observations ($p^* = 10$) in the covariance matrix with $p = 4$ observed variables,

$$p^* = \frac{p(p+1)}{2} = \frac{4(4+1)}{2} = \frac{20}{2} = 10 .$$

As indicated before, the model parameters that must be estimated are direct paths to the endogenous variables and the variances and covariances of the exogenous variables. For the model in Appendix A, it appears that 9 parameters ($q = 9$) must be estimated, including the four parameters/paths (λ_1 to λ_4), the four error variances VAR(δ_1 to δ_4), and the variance of the exogenous factor VAR(ξ_1), resulting in an over-identified model (more observations, 10, in the covariance matrix than model parameters, 9). Because the exogenous variable is unmeasured (latent), its scale of measurement is unknown and would prohibit the estimation of all of the model

parameters in the model, even if the model is over-identified. Thus, the factor variance must be fixed (typically set to a value of 1.0) or one of the direct paths must be fixed (typically set to a value of 1.0) to estimate all of the model parameters. For the model in Appendix A, then, setting the latent factor variance to 1.0 [VAR(ξ_1) = 1.0] will result in 8 model parameters (q = 8) that must be estimated (Byrne, 1994).

Estimation Procedures

When estimating the parameters in just-identified models, the model implied covariance matrix will exactly recover the sample covariance matrix, $\mathbf{S} = \hat{\mathbf{\Sigma}}$, indicating perfect fit of the data to the model. In contrast, estimating the parameters in over-identified models will result in a model implied covariance matrix different than, but hopefully similar to, the sample covariance matrix if the structural model is supported by the data. Recall that the unknown model parameters ($\boldsymbol{\theta}$) are estimated ($\hat{\boldsymbol{\theta}}$) through a minimization process of a discrepancy function.

Maximum Likelihood (ML). The most widely employed discrepancy function in structural equation modeling, and is typically the default discrepancy function in structural equation modeling statistical software (e.g., AMOS, EQS, LISREL, and MPLUS), is the maximum likelihood (ML) discrepancy function:

$$F_{\mathrm{ML}} = \ln |\, \mathbf{\Sigma}(\boldsymbol{\theta}) \,| + \mathrm{tr}[\mathbf{S}\mathbf{\Sigma}^{-1}(\boldsymbol{\theta})] - \ln |\, \mathbf{S} \,| - p \,,$$

where ln is the natural log and tr is the trace function. Looking at the ML function, it is essentially minimizing the sum of squared differences between the observed elements in $\mathbf{S}$ and the corresponding elements in $\mathbf{\Sigma}(\boldsymbol{\theta})$ after weighting them with the

sum of the diagonal elements of $[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})]$. ML estimation is a simultaneous estimation process in that all of the elements in the implied covariance matrix are calculated simultaneously. ML estimation is also an iterative estimation process. The iterative process begins with initial structural model parameter estimates, or starting values, which are either generated by the model fitting software package or provided by the user. Depending upon these values, the model fitting program will iterate through sequential cycles that calculate improved estimates. That is, the elements of the implied covariance matrix which are based on the parameter estimates from each iteration will become closer to the elements of the observed covariance matrix (Kline, 1998). If the structural model is just-identified, the ML discrepancy function will equal zero because the sample covariance matrix will equal the implied covariance matrix, $\mathbf{S} = \hat{\boldsymbol{\Sigma}}$ (Bollen, 1989). For over-identified structural models, the sample covariance matrix will not equal the implied covariance matrix, but the estimation iterations will proceed until the difference between the discrepancy function in one iteration to the next falls below a specified value (e.g., < .001) or until the maximum number of iterations has been reached (Kline, 1998). When the iterations necessary to obtain parameter estimates exceed the maximum number of iterations without reaching the specified minimum difference between the discrepancy function from one iteration to the next, the estimates have failed to converge on the parameters. Nonconvergent solutions may provide unstable parameter estimates which should not be considered reliable. Nonconvergence in this situation may be corrected by

increasing the maximum number of iterations allowed in the structural equation modeling software, changing the minimum difference stopping criteria (e.g., < .001) between iterations, or providing starting values that are closer to the initial estimates of the model parameter estimates. If these corrections do not result in a convergent solution, other issues may need to be addressed, such as sample size and model complexity (Bollen, 1989). The given explanation of ML estimation is simple compared to its mathematical complexities. For more details concerning ML estimation, the reader is referred to Bollen (1989) and Eliason (1993).

Given that the distribution of the observed variables satisfies the multivariate normality assumption, the popularity of the ML discrepancy function is evident when considering the following strengths of the estimators' $(\hat{\boldsymbol{\theta}})$ properties. Under small sample size conditions, ML estimators may be biased, however they are asymptotically unbiased. Thus, as sample size increases, the expected values of the ML estimators $(\hat{\boldsymbol{\theta}})$ represent the true values in the population ($\boldsymbol{\theta}$). The ML estimator is also consistent, meaning that as sample size approaches infinity, the probability that the estimator $(\hat{\boldsymbol{\theta}})$ is close to the true value ($\boldsymbol{\theta}$) approaches 1:

$$\lim_{N \to \infty} P[|\hat{\theta}_N - \theta| < \delta] = 1, \text{ for any small constant } \delta > 0.$$

Another essential property of ML estimators is asymptotic efficiency. That is, the ML estimator has the lowest asymptotic variance among a class of consistent estimators. Further, the ML estimator is scale invariant in that the values of the ML

discrepancy function will be the same for any change in the scale (e.g., correlation matrix vs. covariance matrix) (Bollen, 1989).

The fundamental hypothesis in covariance structure analysis is that the population covariance matrix of p observed variables, $\boldsymbol{\Sigma}$, is equal to the reproduced implied covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, based on the hypothesized structural model, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. For over-identified structural models, ML estimation provides a chi-square (χ^2) test statistic to test this hypothesis of overall model fit. Once the ML fitting function is minimized, the value of F_{ML} is multiplied by $(n - 1)$, yielding an approximately chi-square distributed statistic (Bollen, 1989):

$$\chi^2 = F_{ML}(n - 1),$$

with $p^* - q$ (i.e., number of observations - number of parameters) degrees of freedom (df). The structural equation model is rejected if:

$$\chi^2 > c_\alpha,$$

where c_α is the critical value of the chi-square test and α is the significance level of the chi-square test (e.g., $\alpha = .05$).

ML estimation is based on the assumption of multivariate normality among the observed variables and is often referred to as normal theory ML. Another normal theory estimator is generalized least squares (GLS) (for a review of GLS, see Bollen, 1989). When the assumption of multivariate normality is violated, the parameters estimated by F_{ML}, and F_{GLS}, are typically robust, yielding consistent estimates. However, skewed and kurtotic distributions may sometimes render an incorrect

asymptotic covariance matrix (Bollen, 1989). Further, higher levels of skewness (greater than 3.0) and/or kurtosis (greater than 8.0) generally invalidates asymptotic efficiency of the estimated parameters and creates inaccurate χ^2 test statistics (Kline, 1998). Consequently, observed variables with nonnormal distributions may affect statistical significance tests of overall model fit (χ^2) as well as the consistency and efficiency of the estimated parameters.

Asymptotically Distribution Free (ADF). Other discrepancy functions that produce asymptotically efficient estimators have been proposed that do not require multivariate normality among the observed variables. One of these discrepancy functions is the asymptotically distribution free (ADF) function (Browne, 1984):

$$F_{\text{ADF}} = [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})]' \mathbf{W}^{-1} [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})],$$

where **s** is a *p** x 1 vector of the non-duplicated elements in the sample covariance matrix (**S**), $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is a *p** x 1 vector of the non-duplicated elements in the model implied covariance matrix ($\hat{\boldsymbol{\Sigma}}$), and **W** is a *p** x *p** covariance matrix of **s**. In ADF, the weight matrix (**W**) is the fourth-order multivariate product moment matrix:

$$\mathbf{W}_{\text{ADF}(ij,kl)} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl},$$

where σ_{ijkl} is the fourth-order multivariate moment around the mean:

$$\sigma_{ijkl} = E(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l),$$

and σ_{ij} and σ_{kl} are the covariances between X_i and X_j and between X_k and X_l, respectively. The sample estimator of the fourth-order moment, σ_{ijkl}, is:

$$s_{ijkl} = \sum_{t=1}^{n} (X_{ti} - \bar{X}_i)(X_{tj} - \bar{X}_j)(X_{tk} - \bar{X}_k)(X_{tl} - \bar{X}_l) / n,$$

and the estimators of σ_{ij} and σ_{kl} are:

$$s_{ij} = \sum_{t=1}^{n} (X_{ti} - \bar{X}_i)(X_{tj} - \bar{X}_j) / (n-1)$$

and

$$s_{kl} = \sum_{t=1}^{n} (X_{tk} - \bar{X}_k)(X_{tl} - \bar{X}_l) / (n-1),$$

respectively (Bollen, 1989). The ADF discrepancy function minimizes the weighted sum of squared differences between **s** and $\boldsymbol{\sigma}(\boldsymbol{\theta})$. Similar to ML estimation, for over-identified models, ADF estimation will provide a chi-square (χ^2) test statistic of overall model fit with $p^* - q$ degrees of freedom (df):

$$\chi^2 = F_{\text{ADF}}(n-1),$$

where F_{ADF} is the value of the ADF discrepancy function at the end of the minimization process.

ADF was proposed as an efficient estimator for any arbitrary distribution of observed variables (Browne, 1984). Muthén & Kaplan (1985) demonstrated that ADF estimation is more efficient and that ADF chi-square test values (χ^2) are more consistent than ML under the presence of nonnormality among categorical variables. Unfortunately, the performance of ADF in other studies has been questionable under certain conditions. For instance, ADF chi-square values (χ^2) have been shown to

reject the correct factor model too frequently, even under normal distributions at small sample sizes (Hu, Bentler, & Kano, 1992), and increasingly overestimate the expected χ^2 value as nonnormality and model misspecification increase (Curran, West, & Finch, 1996). ADF chi-square values (χ^2) have also been shown to reject the correct factor model too often as model complexity increases under small sample size conditions (Muthén & Kaplan, 1992). Although ADF has demonstrated better efficiency under nonnormal distributions than ML (Chou, Bentler, & Satorra, 1991; Muthén & Kaplan, 1985), ADF efficiency is adversely affected under conditions of increasing nonnormality, small sample sizes, and large model size (Muthén & Kaplan, 1992). Thus, ADF estimation requires very large sample sizes (approximately 2,500 to 5,000) for accurate chi-square tests (χ^2) and parameter estimates (Hu et al., 1992; Loehlin, 1998). In addition, ADF estimation is more computationally intensive than other estimation procedures due to taking the inverse of the weighting matrix (**W**), which increases in size as the number of observed variables increases (Loehlin, 1998).

Elliptical. When observed variables have a multivariate elliptical distribution (i.e., a symmetric distribution, skew = 0, with longer (positive kurtosis) or shorter (negative kurtosis) tails than those in the normal distribution), an elliptical discrepancy function may be used:

$$F_{\mathrm{E}} = \frac{1}{2}(K+1)^{-1}\,\mathrm{tr}[(\mathbf{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta}))\mathbf{V}^{-1}]^2 - C_1[\mathrm{tr}(\mathbf{S}-\boldsymbol{\Sigma}(\boldsymbol{\theta}))\mathbf{V}^{-1}]^2,$$

where K is the kurtosis parameter which may be estimated with:

$$\hat{K}_1 = \frac{b_2 - \left(\frac{p(p+1)}{2}\right)\left(\frac{p(p+1)}{2}+2\right)}{\left(\frac{p(p+1)}{2}\right)\left(\frac{p(p+1)}{2}+2\right)},$$

where b_2 is a measure of multivariate kurtosis proposed by Mardia (1970):

$$b_2 = \frac{1}{N}\sum_{i=1}^{N}\{(\mathbf{z}_i - \bar{\mathbf{z}})' \, \mathbf{S}^{-1}(\mathbf{z}_i - \bar{\mathbf{z}})\}^2,$$

where $\mathbf{z}_i$ is a column vector of all the variables' values for the ith observation and $\bar{\mathbf{z}}$ is the column vector of corresponding sample means. C_1 is a constant given by:

$$C_1 = K \Big/ \left[4(K+1)^2 + 2\left(\frac{p(p+1)}{2}\right)K(K+1)\right],$$

and $\mathbf{V}$ is a weight matrix (typically $\mathbf{S}$ or $\hat{\mathbf{\Sigma}}$) (Bollen, 1989). When the weight matrix ($\mathbf{V}$) used is the sample covariance matrix ($\mathbf{S}$), the estimation method is Elliptical Generalized Least Squares (EGLS), which is similar to GLS estimation. When the weight matrix ($\mathbf{V}$) used is the implied covariance matrix ($\hat{\mathbf{\Sigma}}$), the estimation method is Elliptical Reweighted Least Squares (ERLS; Bentler, 1995), which is similar to ML estimation. Other estimators of the kurtosis parameter, K, are provided by Bentler (1995). The elliptical estimator is asymptotically efficient and provides a chi-square test statistic of overall model fit:

$$\chi^2 = F_{\mathrm{E}}(n-1),$$

with p^* - q *df*.

The research that has investigated the performance of Elliptical estimation has offered unsettled results. Hu, Bentler, and Kano (1992) found that under nonnormal

distributions, Elliptical chi-square tests (χ^2) are more accurate than those based on normal theory estimators (e.g., ML, GLS). Henly (1993) also found that Elliptical estimation reduced the bias in the standard errors of the estimates as compared to normal theory estimators (e.g., MWL) under conditions of high levels of kurtosis. Nonetheless, Hu et al. (1992) also found that the Elliptical chi-square tests (χ^2) tended to reject the correct model too often under the elliptical distribution condition, even at the largest sample size (n = 5,000). This is similar to the results of Satorra and Bentler (1994) who found that the mean Elliptical chi-square value (χ^2) over 1,000 replications was extremely smaller (M = .84, SD = .73) than the theoretical mean for the correct factor model's chi-square value with 3 df (M = 3.00, SD = 2.45). Moreover, the bias in the standard errors of the estimates under Elliptical estimation has been shown to increase under leptokurtic distributions (Hoogland & Boomsma, 1998).

Correction Procedures

Satorra-Bentler Scaled Statistics. ML and Elliptical chi-square tests (χ^2) are based upon the multivariate normal and elliptical distribution assumption, respectively. As discussed above, the chi-square tests (χ^2) obtained when using these estimation methods may not achieve the expected chi-square (χ^2) distribution when their distributional assumptions are violated (Hu et al., 1992). Bias in the standard errors of the parameter estimates has also been demonstrated under increased conditions of nonnormality (Hoogland & Boomsma, 1998). ADF estimation may be

used alternatively, however, ADF chi-square tests (χ^2) and standard errors for the model parameters have demonstrated inaccuracies under small sample size and increased model complexity conditions (Muthén & Kaplan, 1992). Satorra and Bentler (1994) have offered a scaling solution to provide a chi-square test (χ^2) statistic that more closely approximates the chi-square (χ^2) distribution and robust standard errors when assumptions of normality are violated. The Satorra-Bentler scaled χ^2 test statistic is given by:

$$\chi^2_{\mathrm{SB}} = c^{-1}\chi^2,$$

where χ^2 is the chi-square value obtained after a normal theory discrepancy function is minimized [e.g., $\chi^2 = F_{\mathrm{ML}}(n-1)$ or $\chi^2 = F_{\mathrm{GLS}}(n-1)$], and c is the scaling constant:

$$c = (df)^{-1}\,\mathrm{tr}(\mathbf{U\Gamma}),$$

where $\mathbf{U}$ is the residual weight matrix:

$$\mathbf{U} = \mathbf{W} - \mathbf{W\Delta}(\mathbf{\Delta'W\Delta})^{-1}\mathbf{\Delta'W}),$$

where $\mathbf{W}$ is the weight matrix used in the normal theory estimation procedure after the discrepancy function is minimized, $\mathbf{\Delta}$ is the p^* x q partial derivative matrix with respect to the parameters:

$$\mathbf{\Delta} = \partial\boldsymbol{\sigma}(\boldsymbol{\theta}) / \partial\boldsymbol{\theta},$$

$\mathbf{\Gamma}$ is the p^* x p^* covariance matrix containing fourth-order moments ($s_{ijkl} - s_{ij}s_{kl}$) as in the ADF weight matrix, and df are the $p^* - q$ degrees of freedom for the model. For normal theory estimators, such as ML and GLS, $\mathbf{W}$ is given by:

$$\mathbf{W} = 2\mathbf{K}'_p(\hat{\mathbf{\Sigma}} \otimes \hat{\mathbf{\Sigma}})\mathbf{K}_p,$$

where $\mathbf{K}_p$ is the p^2 x p^* transition matrix (see Browne, 1974; Nel, 1980) containing elements:

$$\mathbf{K}_p(ij,kl) = \begin{cases} 1/2, \text{ if } i \neq j \text{ but } (i=k,\ j=l) \text{ or } (i=l,\ j=k); \\ 1, \text{if } i=j=k=l; \\ 0, \text{else; where } i<p,\ j<p,\ k<l<p, \end{cases}$$

and $\otimes$ represents the Kronecker product. Robust standard errors are given by:

$$\text{acov}(\hat{\theta}) = n^{-1}(\mathbf{\Delta'W\Delta})^{-1}(\mathbf{\Delta'W\Gamma W\Delta})(\mathbf{\Delta'W\Delta})^{-1}.$$

The Satorra-Bentler scaled χ^2 test statistic (SB χ^2) has been shown to perform more accurately than ML or ADF chi-square tests (Chou et al., 1991) and Elliptical chi-square tests (Hu et al., 1992) under increasing nonnormality conditions. The SB χ^2 has also performed more accurately than ML or ADF chi-square tests under model misspecification conditions (Curran et al., 1996). The Satorra-Bentler scaled standard errors for the model parameters have also been shown to be more robust than ADF and ML under increasing nonnormality conditions, demonstrating less bias (Chou et al., 1991). Similar to ADF, however, a drawback of the SB χ^2 is that it is based upon the fourth-order moments which requires more computation time than other estimation methods.

Overall Model Fit

Each estimation procedure discussed above is able to provide a chi-square test statistic which is typically used as a test of overall model fit in structural equation

modeling. The unfortunate problem with using the χ^2 test statistic, however, is its sensitivity to sample size, which is evident from the formula:

$$\chi^2 = F(n - 1).$$

Thus, as sample size increases, the χ^2 test statistic is more likely to identify small differences between the observed and implied covariance matrices as being significant. Consequently, the null hypothesis, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, is more likely to be rejected, indicating poor fit of the hypothesized structural model. The χ^2 test statistic has often been criticized for its sample size dependency (Bentler & Bonett, 1980; Hu & Bentler, 1999; Kline, 1998), which has led to the proposal of numerous alternative fit indices that evaluate model fit and supplement the χ^2 test statistic. These fit indices have been classified as either incremental or absolute fit indices (Hu & Bentler, 1999).

Incremental Fit Indices. Incremental fit indices measure the proportionate improvement in a model's fit to the data by comparing a specific structural equation model *k* to a baseline structural equation model. The typical baseline comparison model is the null model in which all of the variables in the model are independent of each other or uncorrelated (Byrne, 1994). Commonly used incremental fit indices include the Normed Fit Index (NFI; Bentler & Bonnett, 1980), Bollen's (1986) Incremental Fit Index (IFI), the Non-normed Fit Index (NNFI) which is also called the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), the Comparative Fit Index

(CFI; Bentler, 1990), and the Relative Noncentrality Index (RNI; McDonald & Marsh, 1990).

Absolute Fit Indices. Absolute fit indices measure how well a structural equation model reproduces the data. Commonly used absolute fit indices include the Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI; Jöreskog & Sörbom, 1984), Akaike's (1987) Information Criterion (AIC), McDonald's (1989) Fit Index (MFI), the Standardized Root Mean-Square Residual (SRMR; Bentler, 1995), and the Root Mean-Square Error of Approximation (RMSEA; Steiger & Lind, 1980).

Reporting Fit Indices. Incremental and absolute fit indices indicate different aspects of model fit. Thus, it is recommended to report values of several fit indices. According to Kline (1998), a minimum collection of fit indices to report would consist of the χ^2 test statistic with corresponding degrees of freedom and level of significance, a fit index indicating the overall proportion of variance explained (e.g., CFI, GFI, or NFI), a fit index that adjusts for model parsimony (e.g., NNFI or IFI), and a fit index that considers the standardized residuals (e.g., SRMR). Hu and Bentler (1999) recommend reporting the SRMR with a supplemental fit index that meet certain cutoff values depending upon the combination of indices reported. When evaluating the fit of a model, the cutoff value for the SRMR should be about .09 with a supplemental fit index cutoff value around .95 for the NNFI, IFI, RNI, MFI, or CFI, or a cutoff value around .05 for the RMSEA. It is important to note that

the fit indices only indicate the overall fit of a structural equation model and not necessarily every facet of the model. In addition, fit indices do not reflect the predictive validity of the model.

Model Modification and Selection

When structural equation models do not demonstrate adequate fit through the use of the χ^2 test statistic and/or fit indices, models are typically modified, or respecified, and subsequently re-tested (MacCallum, Roznowski, & Necowitz, 1992). While the modification of structural equation models should be based upon substantive theory, modification indices are sometimes relied upon to either add or drop parameters in the model. For instance, the Lagrange Multiplier modification index estimates the decrease in the χ^2 test statistic that would occur if a parameter were to be freely estimated. More specifically, it indicates which parameters could be added to increase the fit of the model by significantly decreasing the χ^2 test statistic of overall fit. In contrast, the Wald stastistic estimates the increase in the χ^2 test statistic that would occur if a parameter was fixed to zero. Thus, the Wald statistic estimates whether a parameter may be dropped from the model without significantly increasing the χ^2 test statistic (Kline, 1998).

The two models being compared in this procedure are said to be nested or hierarchically related models. In other words, one of these models is a subset of the model to which it is compared. For instance, if the Lagrange Multiplier index indicated that the χ^2 test statistic would significantly decrease by adding a parameter

to Model 1, Model 1 would be nested in the model with the new parameter incorporated (Model 2). Moreover, if the Wald statistic indicated that the χ^2 test statistic would not significantly increase by eliminating a parameter from Model 1, the model formed by eliminating the parameter (Model 2) would be nested in Model 1 (Kline, 1998).

These modification indices determine if a significant difference between two models' χ^2 test statistics exists when adding or eliminating a model parameter and is reflected by the chi-square difference test ($\chi^2_{\text{difference}}$):

$$\chi^2_{\text{difference}} = \chi^2_{\text{restricted}} - \chi^2_{\text{unrestricted}},$$

where $\chi^2_{\text{restricted}}$ is the chi-square value (χ^2) for the nested, less parameterized (restricted) model and $\chi^2_{\text{unrestricted}}$ is the chi-square value (χ^2) for the hierarchically more parameterized, less restricted (unrestricted) model, with corresponding degrees of freedom for the $\chi^2_{\text{difference}}$ test:

$$df_{\text{difference}} = df_{\text{restricted}} - df_{\text{unrestricted}}.$$

When the $\chi^2_{\text{difference}}$ test indicates a significant difference between two hierarchically related models ($\chi^2_{\text{difference}} > c_\alpha$), the nested model with less parameters has been oversimplified. That is, the less parameterized (nested) model has significantly decreased the overall fit of the model when compared to the model with more parameters. In this situation, then, the more parameterized model would be selected over the less parameterized model. On the other hand, when the $\chi^2_{\text{difference}}$ test is not

significant ($\chi^2_{\text{difference}} < c_\alpha$), the two models are comparable in terms of overall model fit. In this situation, the less parameterized would most likely be selected over the more parameterized model in support of parsimony (Kline, 1998). It is important to note here that the $\chi^2_{\text{difference}}$ test is possible because it uses likelihood ratio χ^2 statistics as opposed to Pearson χ^2 statistics.

When structural equation models are not hierarchically related, the $\chi^2_{\text{difference}}$ test is an inappropriate method to assess significant model fit differences because neither of the two models can serve as a baseline comparison model. Still, there are instances in which non-hierarchically related models are compared in terms of model fit, such as when testing different theoretical models posited to support the data. Values of the χ^2 test statistic for each non-hierarchical model are sometimes compared, however, it is important to note that models with more parameters tend to have smaller χ^2 values, indicating better fit than less parameterized models (Kline, 1998). To compensate for the number of parameters in the model, and to encourage parsimony, Akaike's (1987) Information Criterion (AIC) is often used to compare the fit of non-hierarchically related models. The AIC adjusts the regular χ^2 test statistic by including a penalty for the number of parameters estimated in the model:

$$\text{AIC} = \chi^2 - 2df ,$$

in which more parameterized models have fewer degrees of freedom (*df*) than less parameterized models. Consequently, the AIC reduces the χ^2 value more for less

complex models than for more complex models. Structural equation models with the lowest AIC value, which may be a negative value, are selected as a model demonstrating better fit than its comparison model.

Similar to subset selection in multiple regression, modifying and re-testing structural equation models on the basis of model fit indices capitalizes on the chance occurrences within the sample in which the models are being tested. Again, cross-validation of structural equation models, especially once one has been selected through modification procedures, is highly recommended to ensure its predictive validity (MacCallum et al., 1992). Although there is no procedural rule of thumb when conducting structural equation models cross-validation, multiple group analysis has been suggested as a pragmatic cross-validation method. In multiple group analysis, the structural equation model developed in the calibration sample is applied to the data in the validation sample to simultaneously test for invariance (model equivalency) across the two samples by constraining different sets of model parameters to be equal in both groups. The model parameter constraints employed in the analysis, which may vary in terms of their restrictive nature, are dependent upon the structural equation model as well as the particular model hypotheses. Model parameters that may be constrained to be equal across the two samples include: a) direct paths to endogenous variables, b) factor variances/covariances, and c) error variances/covariances. When testing for multiple group invariance, one or all of these constraints may be imposed on the parameters (Bollen, 1989; Byrne, 1994). It has been argued, however, that imposing all three constraints may be too restrictive and

that the invariance of error variances/covariances across samples is of least interest (Bentler, 1995) and may be freely estimated in both groups instead of constrained to be equal (Byrne, 1994).

After the constraints are imposed and simultaneously tested across both samples, fit indices are typically examined to determine how well the model developed in the calibration sample cross-validated in the validation sample. The Lagrange Multiplier modification index is also examined. If the Lagrange Multiplier index indicates that some constraints should be released, indicating that these constrained parameters should be freely estimated separately for each sample, the model is re-tested without these constraints. If the fit indices do not indicate poor fit, there is support that the model cross-validates or is equivalent across the two samples (Byrne, 1994).

Unfortunately, cross-validation of structural equation models is a rare practice (Breckler, 1990). For instance, MacCallum, Roznowski, and Necowitz (1992) examined 100 studies published in psychological journals that conducted covariance structural modeling. Of these 100 studies, 37 of them indicated that modification of the initial model occurred in order to improve the fit of the model to the data. Of these 37 studies, only 4 indicated that they cross-validated the final model selected through a modification process. MacCallum et al. (1992) also demonstrated the instability of model modification and the effect of model modification on cross-validation outcomes. In two sampling studies, several pairs of samples were created by randomly selecting cases from an existing data set to form various sample size

conditions. Calibration samples were used to fit an initial model to the data, modify the model using modification indices, and refit the modified model to the data. This sequence (initial fit, modification, and refitting) was conducted three to four times. For the initial and modified models, two cross-validation indices (see Cudeck & Browne, 1983; Browne & Cudeck, 1989) were calculated based on the validation samples. (Note: These cross-validation indices will be discussed in the following subsection). The results indicated that the particular model modifications in each sequence were inconsistent across replication samples, especially when sample size was small (n = 100). Accordingly, the cross-validation indices demonstrated inconsistency across replication samples, especially under small to moderate sample size conditions. Hence, modified models did not demonstrate consistent cross-validation outcomes.

Recall that the hypothesis tested by the χ^2 test statistic, as well as fit indices and modification indices, is that the structural equation model fits the data exactly. Cudeck and Browne (1983) advocate a departure from the hypothesis of exact fit because models are never truly known and are only approximations of the truth. Instead, structural equation models should be considered as one of several reasonable models that may adequately describe theory. More specifically, theoretically reasonable models should be selected a priori and compared systematically in order to determine which structural equation model will optimally generalize to samples from the same population.

Cross-Validation and Model Selection

Two-Sample Cross-Validation Index (CVI). Cudeck and Browne (1983) have recommended the use of cross-validation when analyzing theoretically competing or alternative, a priori models for covariance structures. They described a procedure in which two subsamples, *a* and *b*, are randomly selected from the same sample and have covariance matrices $\mathbf{S}_a$ and $\mathbf{S}_b$, respectively. Sample *a* is used as the calibration sample and sample *b* as the validation sample. The model, *k*, is fitted to the data from the calibration sample, producing the implied covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta}_{k|a})$. The cross-validation index given by Cudeck and Browne is:

$$\text{CVI} = \text{F}[\mathbf{S}_b; \mathbf{\Sigma}(\boldsymbol{\theta}_{k|a})],$$

which is the discrepancy between the covariance matrix for the validation sample, $\mathbf{S}_b$, and the implied covariance matrix from the calibration sample, $\mathbf{\Sigma}(\boldsymbol{\theta}_{k|a})$. The minimization process is not actually conducted when calculating the CVI. Instead, the CVI is a measure indicating the discrepancy between $\mathbf{S}_b$ and $\mathbf{\Sigma}(\boldsymbol{\theta}_{k|a})$. The CVI is calculated for each of the competing models, and the model with the smallest CVI is selected as the one with the most predictive validity and most likely to cross-validate.

Double cross-validation may also be conducted when selecting among competing models by calibrating the model in two subsamples from the same population and then measuring the discrepancy between the sample covariance matrix for each subsample and the opposing subsample's implied covariance matrix, yielding two CVI values:

$$\mathrm{CVI}_{\mathrm{a}} = \mathrm{F}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})]$$

and

$$\mathrm{CVI}_{\mathrm{b}} = \mathrm{F}[\mathbf{S}_a; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})].$$

The model obtaining the smallest CVI in each subsample demonstrates the most predictive accuracy. If no model obtains the smallest CVI in both subsamples, the models with small CVI values in either subsample may then be evaluated in terms of theoretical viability.

Cudeck and Browne (1983) suggest that the set of competing structural equation models should include a lower bound comparison model, such as the null model or models with only a few parameters (e.g., a one-factor model), and a higher bound comparison model, such as the saturated model with the maximum number of possible parameters. These lower and upper reference models will assist in the comparison process of other theoretically reasonable a priori models and in the judgment of which competing model demonstrates the most predictive validity. The relative performance of competing models in terms of cross-validation is necessary since there is no cutoff value for the CVI.

Multiple-Sample Cross-Validation Index (C). De Gooijer and Koopman (1988) proposed a cross-validation method in which the two-sample CVI may be employed using the "leave-one-out" method with more than two subsamples (i.e., L > 2). Similar to the jacknife sampling technique, De Gooijer and Koopman (1988) suggested fitting each model in a calibration sample comprised of all the subsamples

(c_i) except the one that is "left-out," which is subsequently employed as the validation sample (v_i). This process is repeated for each of the subsamples, leaving out a different subsample (v_i) each time until each of the subsamples has been employed as a validation sample (v_i, i = 1…L). This procedure will yield the same number of cross-validation indices as the number of subsamples employed. This multiple-sample cross-validation index is given by:

$$C = \sum_{i=1}^{\mathrm{L}} \mathrm{F}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|c_i})],$$

which is the sum of the discrepancies between the covariance matrix for each of the validation samples, $\mathbf{S}_{v_i}$, and the implied covariance matrix from each of the calibration samples, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|c_i})$. The model with the smallest summed cross-validation index is selected as the one with the most predictive validity. When the number of subsamples is equal to 2 (L = 2), the value of C is simply the sum of the double cross-validation indices using the two-sample CVI.

The use of cross-validation is important in terms of selecting a reliable model that will fit the data available in other samples. An advantage of utilizing the cross-validation, model selection procedure is that the competing a priori structural equation models do not have to be hierarchically related. Thus, the cross-validation procedure may be used to select among competing non-nested models. Similar to cross-validation methods in multiple regression, however, a drawback of the two-sample CVI, and especially the multiple-sample C, is that the sample must be

randomly split to yield both the calibration and validation samples. This is not always feasible, especially when working with already small samples which may result in nonconvergent estimation solutions. In addition, estimates are not as precise as they would be if based on the entire sample. As a result, single sample cross-validation indices have been proposed.

Single-Sample Akaike Information Criterion (AIC). One single sample cross-validation index was proposed by Cudeck and Browne (1983). They recommended using the rescaled Akaike (1973) Information Criterion (AIC), which they expressed in terms of the Maximum Likelihood discrepancy function:

$$\text{AIC} = F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k /(n-1),$$

where $F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is the ML discrepancy function of model *k*, q_k is the number of freely-estimated parameters in the model, and *n* is the sample size. The rationale for using the AIC as a single-sample cross-validation measure is that it estimates the difference between the fit of the model in the sample data and the fit of the model in the population using the log likelihood ratio test (see Akaike, 1987 for a more detailed explanation). The AIC also includes a penalty for model complexity in which more complex models generally yield higher AIC values than less complex models, indicating the importance placed upon parsimony when selecting structural equation models. Among competing alternative models, the model obtaining the smallest AIC value would indicate more predictive accuracy.

Cudeck and Browne (1983) compared the performance of the rescaled AIC and the two-sample CVI in terms of model selection among 17 competing factor analysis models within a real data set. The sample size was varied by randomly selecting observations from 2,676 cases with data and forming two equal subsample sizes (75, 100, 200, 400, 800, and 1338). Using Maximum Likelihood (ML) estimation, the two-sample CVI and the AIC were used to double cross-validate the models in each of the two subsamples for all models, resulting in 2 values of each index for each model. The results indicated that the AIC tended to be more liberal than the CVI with respect to reaching a minimum value for models that are highly parameterized. Thus, the AIC tended to select more highly parameterized models than the two-sample CVI, especially when sample size was small. However, the CVI and AIC values became more comparable as sample size increased. Both the CVI and the AIC resulted in larger values for the saturated model under smaller sample size conditions, but resulted in smaller values for the saturated model under larger sample size conditions. Cudeck and Browne (1983) warned that this does not necessarily indicate that the saturated model is the preferred model. On the contrary, highly parameterized models are difficult to interpret and to explain. Cudeck and Browne (1983) recommend a compromise between goodness-of-fit and parsimony, given that the selected parsimonious model indicates predictive validity with a fairly small cross-validation index value.

Single-Sample Consistent Akaike Information Criterion (CAIC). Due to the sample size dependency of the CVI and the AIC, Bozdogan (1987) extended the

Akaike Information Criterion to incorporate a penalty for adding model parameters that is dependent upon sample size. This single sample index, called the consistent AIC (CAIC), is given as:

$$\text{CAIC}(k) = -2\ln L(\boldsymbol{\theta}_k) + q_k[\ln(n)+1],$$

where ln is the natural log, $L(\boldsymbol{\theta}_k)$ is the likelihood function, q_k is the number of parameters for model *k*, and *n* is the sample size. In a Monte Carlo study, Bozdogan (1987) compared the performance of the AIC and the CAIC in terms of selecting the correct degree of a polynomial model under different sample size (50, 100, and 200) and residual variance conditions (.25, .50, 1.00, and 5.00). It was found that, overall, the AIC tended to select an overfitted polynomial model more often than the CAIC. The AIC also tended to select an overfitted polynomial model as sample size and residual variance increased more often than the CAIC. The CAIC performed fairly consistently across all conditions. The AIC and CAIC both demonstrated the worst performance under the smallest sample size (n = 50) and highest residual variance conditions ($\sigma_e^2 = 5.00$), indicating the incorrect degree of the polynomial 27 out of 100 and 19 out of 100 times, respectively. The rescaling procedure applied to the AIC in Cudeck and Browne (1983) may also be employed to rescale the CAIC in terms of the ML discrepancy function:

$$\text{CAIC} = F_{\text{ML}}[\mathbf{S};\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + q_k[\ln(n)+1]/(n-1),$$

where $F_{\text{ML}}[\mathbf{S};\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is the ML discrepancy function of model *k*, q_k is the number of freely-estimated parameters in the model, and *n* is the sample size. The model that

obtains the smallest rescaled CAIC value among competing alternative models would indicate more predictive validity.

Browne and Cudeck (1989) further studied the performance of the single sample rescaled AIC and the two-sample CVI, employing them again in the real data set used in Cudeck and Browne (1983). From 2,676 observations of data, 20 sample pairs for each sample size (100, 400, and 800) were randomly selected with replacement. Using ML estimation, four different factor analysis models were double cross-validated with the CVI and the AIC for each of the 20 sample pairs in each sample size, resulting in 2 values of each index for each model. The means of the two indices were calculated over the forty replications (20 replications in each of the subsamples). The CVI and AIC agreed in terms of reaching a minimum value for the same factor analysis models across conditions and their values became more comparable as sample size increased. However, the AIC values tended to be smaller for the competing models as compared to the CVI, especially for the saturated factor analysis model under the smallest sample size condition (100).

Single-Sample Browne-Cudeck Index (BCI). In the same study, Browne and Cudeck (1989) proposed another single sample cross-validation index, the Browne-Cudeck Index (BCI), which incorporates a correction to the AIC to adjust for its underestimation of the CVI. The BCI, also called the expected cross-validation index (ECVI; see Browne & Cudeck, 1992), is given by:

$$\text{BCI} = F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k/(n-p-2),$$

where $F_{\mathrm{ML}}[\mathbf{S};\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is the ML discrepancy function of model *k*, q_k is the number of parameters, *p* is the number of observed variables, and *n* is the sample size. Browne and Cudeck (1989) showed proofs demonstrating that the single-sample BCI approximates the expected two-sample CVI value, using only the calibration sample data, under multivariate normality when using the ML discrepancy function. When compared to the CVI and AIC values, the correction included in the BCI improved the values of the AIC in the real data. However, the values of the BCI still remained slightly lower than the CVI values, even for the saturated factor analysis model for which the values should be equal. Browne and Cudeck (1989) attributed this to using real data which were not normally distributed. They conducted a Monte Carlo study comparing the performance of the BCI and the two-sample CVI in order to verify that the BCI was correct under multivariate normality conditions for the saturated model. For 500 sample pairs of equal size (*n* = 100), it was determined that the mean double cross-validated values for the CVI and the BCI were equivalent.

"Pseudo" Single-Sample C and* $\overline{C}$ **.* De Gooijer and Koopman (1988) proposed another single-sample cross-validation index (*C**) that approximates their summed multiple-sample cross-validation index (*C*) and is given by:

$$C^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i};\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k \mathrm{L}^2/(n-1),$$

where $F_{\mathrm{ML}}[\mathbf{S}_{v_i};\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is the ML discrepancy function of model *k*, q_k is the number of freely-estimated parameters, *n* is the number of observations in the entire sample,

and L is the number of subsamples employed to validate each competing model. The "pseudo" single-sample cross-validation index proposed by De Gooijer and Koopman (1988) is similar to other single-sample cross-validation indices in that the implied covariance matrix $[\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is produced from fitting the model, *k*, to the entire sample of data. It differs, however, from other single-sample cross-validation indices in that the entire sample is then split into several subsamples (L) which serve as validation samples (v_i, $i = 1 \ldots \mathrm{L}$). Thus, the discrepancy between the covariance matrix for each subsample, $\mathbf{S}_{v_i}$, and the covariance matrix implied by the model for the entire sample $[\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)]$ is calculated, yielding the same number of discrepancy values as the number of validation subsamples employed. These discrepancy values are then summed and corrected, $2q_k \mathrm{L}^2/(n-1)$, to provide a single cross-validation index, in which models with the smallest *C** are selected as models with the most predictive accuracy. It is important to note that when L = 1, this cross-validation index is equivalent to the rescaled AIC. The correction applied to the rescaled AIC in Browne and Cudeck (1989), yielding the BCI, may also be applied to De Gooijer and Koopman's (1988) "pseudo" single-sample cross-validation index:

$$\overline{C}^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k \mathrm{L}^2/(n-p-2).$$

De Gooijer and Koopman (1988) conducted a Monte Carlo study investigating the performance of the two-sample CVI, the multiple-sample *C*, and single-sample cross-validation indices, including the AIC, BCI, *C**, and $\overline{C}^*$, with respect to

selecting the correct factor model. Factor models were investigated under varying factor number (1 and 2), error variance (1 and 4), factor loading, model misspecification (3 levels), and sample size (n = 50, 100, 200, and 400) conditions. For each sample size, data sets were generated from multivariate normal distributions to yield 40,000/n replications (i.e., 800, 400, 200, and 100 data sets, respectively). Each of these data sets was randomly split into calibration and validation samples, c and v, respectively. Using ML estimation, each of the factor models was double cross-validated using each cross-validation index under investigation. Using double cross-validation resulted in: 1) two CVI values; 2) one C value; 3) two AIC values; 4) an average AIC value; 5) two BCI values; 6) an average BCI value; 7) one C* value; and 8) one $\overline{C}$ * value for each model. The findings indicated that the two-sample CVI and the multiple-sample C tended to incorrectly select more parameterized models, even at larger sample sizes. The double cross-validated AIC and BCI indices performed better than the two-sample cross-validation indices and performed fairly comparably in terms of selecting the correct model, especially as sample size increased. In addition, the averaged AIC and BCI indices demonstrated improved performance over their two separate calibration and validation index counterparts. The single-sample cross-validation indices proposed by De Gooijer and Koopman (1988), C* and $\overline{C}$ *, tended to select the correct factor model more frequently than the other cross-validation indices investigated. All of the cross-validation indices, however, performed the worst when error variance was high, meaning that factor

loadings were low, and the factor model was more complex, especially under small sample sizes.

The "pseudo" single-sample cross-validation indices proposed by De Gooijer and Koopman (1988) outperformed the other cross-validation indices in their study using only two subsamples (L = 2), even though they recommended using more than two subsamples (L > 2). De Gooijer and Koopman (1988) note that splitting the sample into more than two subsamples may not be realistic considering the small data sets with which researchers typically work, which will most likely limit the use of their cross-validation indices to only two subsamples. Nevertheless, their "pseudo" single-sample indices performed well under this realistic limitation of sample splitting.

Bandalos (1993) suggested that the decision of which cross-validation index to use should depend upon the researcher's intention. For instance, if one cannot use the two-sample CVI, but would like an estimate of the CVI, the AIC or the BCI could be employed under large sample and/or high factor loading conditions. This was concluded in the Monte Carlo study conducted by Bandalos (1993) which compared the performance of the CVI, the AIC, the CAIC, and the BCI under different sample size conditions (200 and 600), factor loading conditions (.4, .6, and .8) with corresponding error variances (.8, .6, .4), as well as model misspecification conditions (three increasing levels). Using the CVI as the standard, the results indicated that the mean value of the CAIC over all conditions differed greatly from the CVI, indicating that the mean value of the CAIC was significantly larger than the mean CVI value.

Separate analyses indicated that all four of the cross-validation indices produced lower values when the sample size was large than when sample size was small. Factor loading effects demonstrated that as factor loading size and sample size increased, all four indices were comparable in terms of selecting the correct model. However, when factor loadings and sample size were low, the CAIC produced minimum values for an incorrect model as compared to the other three cross-validation indices. Model misspecification effects were also demonstrated, in that the more misspecified factor models had higher cross-validation values.

Statement of the Problem

To summarize the findings discussed concerning the cross-validation model selection procedure, sample size has been shown to affect cross-validation index values. Studies have consistently found that as sample size increases, values for the cross-validation indices decrease (e.g., Cudeck & Browne, 1983; Browne & Cudeck, 1989; Bandalos, 1993). Because the underlying statistical theory in structural equation modeling is asymptotic, large sample sizes are necessary to provide stable parameter estimates (Bentler, 1995). The chi-square test statistic (χ^2), as well as commonly used supplemental fit indices (e.g., IFI, NFI), have been shown to be affected as sample size increases, sometimes indicating inaccurate conclusions concerning model fit (Marsh, Balla, & McDonald, 1988). Thus, it is necessary to examine how the performance of cross-validation indices is affected by sample size with respect to selecting the model with the most predictive accuracy.

Cross-validation index values have also been shown to depend upon the interaction between sample and model size. Cudeck and Browne (1983) demonstrated that lower cross-validation index values are obtained when sample size is small and models are not highly parameterized, whereas models more highly parameterized are more likely to cross-validate when sample size is large. Because more parameterized models explain more of the covariance among observed variables than less parameterized models, they tend to demonstrate better fit to the data. This does not indicate that more parameterized models are more theoretically reasonable than less parameterized models. On the contrary, the underlying objective of model selection is a theoretical balance between parsimony and goodness-of-fit. Thus, structural equation models should be theoretically viable, parsimonious, and demonstrate predictive validity. Unfortunately, Cudeck and Browne (1983; and Browne & Cudeck, 1989) examined the performance of cross-validation indices by sampling from an extant data set in which the true factor model is not known. Consequently, the performance of the cross-validation indices in terms of selecting the true model cannot be ascertained from these studies.

Bozdogan (1987) conducted a Monte Carlo study to investigate the performance of the AIC and the CAIC in terms of selecting the correct model. The Monte Carlo study indicated that the AIC tended to select more parameterized models than the CAIC. Nonetheless, the study did not examine the performance of the two-sample CVI, the multiple-sample *C*, or other single-sample indices (i.e., the BCI, *C**,

and $\overline{C}$*). Further, the models under consideration were polynomials, which are fundamentally different from linear structural equation models in terms of the relationships among the variables (linear versus curvilinear).

Bandalos (1993) conducted a Monte Carlo study to examine the performance of the two-sample CVI and the single-sample AIC, CAIC, and BCI in terms of selecting the correct confirmatory factor analysis model from among other misspecified models. The most parameterized model was designated as the true model, which was consistently selected as correct by the cross-validation indices as sample size increased. Because the correct model was the most parameterized, it is not known whether the performance of the cross-validation indices was a function of their accuracy in model selection or a function of the increased fit of more parameterized models to the data. Bandalos also suggested that a more extensive range of model misspecification should be further investigated.

In contrast to Bandalos (1993), De Gooijer and Koopman (1988) allowed the correct factor models to be low to moderately parameterized, which allows the examination of whether cross-validation indices will be able to accurately select less parameterized models that are indeed the true model. In their Monte Carlo study, the two-sample CVI and the multiple-sample *C* cross-validation indices tended to incorrectly select more parameterized models too frequently at larger sample sizes and were outperformed by the other single-sample cross-validation indices (i.e., the AIC, BCI, *C**, and $\overline{C}$*). Nonetheless, De Gooijer and Koopman examined either

one or two factor models in which the factors were measured by five observed variables. In practice, structural equation models are typically simpler in terms of the number of indicators per latent factor (e.g., 3 or 4) and more complex in terms of the number of factors (3 or 4; Anderson & Gerbing, 1984). Thus, it is important to examine the ability of cross-validation indices to select among more realistic structural equation models.

Factor loading size has also been shown to affect the performance of cross-validation indices when selecting among competing models. Bandalos (1993) also investigated the effects of known factor loading size on the performance of cross-validation indices. It was demonstrated that factor loading size does affect the performance of cross-validation indices and also interacts with sample size. It was determined that the performance of the cross-validation indices becomes more comparable as factor loading and sample sizes increase, selecting the correct factor model. This is consistent with other studies that have found that lower standard errors associated with parameter estimates are obtained when factor loadings are high (Boomsma, 1982; 1985) and that indicator reliabilities are higher when factor loadings are high (Gerbing & Anderson, 1985). The influence of factor loading size on the performance of cross-validation indices is important in structural equation modeling in that models with higher factor loadings indicate more reliability in the factor structure, which, in turn, may affect predictive validity. A clarification is due here because it was originally thought that Bandalos (1993) generated data with given

standardized factor loadings. In contrast, Bandalos (1993) actually generated data in her study with given unstandardized factor loadings.

Based on these findings, researchers could use the cross-validation index that is most appropriate for their research purposes, or several of the cross-validation indices, when selecting the model that will most likely cross-validate. Researchers could report the cross-validation indices employed, the values of these indices, and which model was selected as having the most predictive validity. However, the previous Monte Carlo studies have been conducted assuming conditions of multivariate normality. Indeed, Browne and Cudeck (1989) found that the single sample BCI values were lower than expected as compared to the two-sample CVI when using empirical data, which they attributed to the presence of nonnormality among the variables. Browne and Cudeck did not report how extreme the nonnormality was, but this demonstrates how rare multivariate normality is when performing analyses within authentic data sets (Micceri, 1989).

To date, no studies have investigated the performance of cross-validation indices under conditions of nonnormality. The studies investigating the performance of cross-validation indices in terms of selecting structural equation models have all employed the ML discrepancy function (i.e., Cudeck & Browne, 1983; Browne & Cudeck, 1989; Bandalos, 1993), which assumes multivariate normality. While ML estimation is fairly robust to deviations from multivariate normality, the corresponding χ^2 statistic and the estimator's efficiency are questionable as deviations

from multivariate normality increase. Consequently, other asymptotically efficient estimation methods have been suggested to compensate for nonnormality among the observed variables without the normality assumption, such as ADF and Elliptical. The performance of ADF and Elliptical estimation procedures, however, has been questionable. On the other hand, the Satorra-Bentler scaled statistic has been shown to perform well under increasing deviations from multivariate normality, yielding efficient parameter estimates and accurate chi-square test statistics. Because the cross-validation indices are based on the ML discrepancy function, the use of the Satorra-Bentler scaled statistic should be used in conjunction with ML estimation under conditions of nonnormality.

Interestingly, several of the studies (e.g., Browne & Cudeck, 1989; Cudeck & Browne, 1983; De Gooijer & Koopman, 1988) performed double cross-validation, even when using single-sample cross-validation indices which were proposed in order to avoid the problem of sample splitting. Moreover, the studies that examined single-sample indices in a simple cross-validation design are limited with respect to designating the most parameterized model as correct (viz., Bandalos, 1993) and the single-sample cross-validation indices examined in the study. Accordingly, it is not clear whether single-sample indices perform adequately when employed within the context of a simple cross-validation design for which they were intended. De Gooijer and Koopman (1988) proposed both multiple-sample and single-sample cross-validation indices (C, C^*, and $\overline{C}^*$, respectively) with which they recommended

employing more than two subsamples. Still, only two subsamples were investigated in their study. Indeed, De Gooijer and Koopman (1988) noted that splitting the sample into more than two subsamples may not be feasible considering the small data sets with which researchers typically work. Hence, it is not known whether employing more than two subsamples would benefit or hinder the performance of their proposed cross-validation indices, especially under small sample sizes.

Purpose of the Study

The purpose of this Monte Carlo study is to extend the literature concerning the cross-validation model selection method first proposed by Cudeck and Browne (1983). Specifically, this study compares the performance of two-sample cross-validation indices, including the CVI and *C*, to that of the single-sample cross-validation indices, including the AIC, CAIC, BCI, *C**, and $\overline{C}$ *, when selecting among competing models. The current study examines their selection from among realistic confirmatory factor analysis models. The performance of these seven indices are investigated under similar conditions included in previous studies, such as sample size, factor loading size, and model misspecification. However, the current study extends the literature by including additional varied conditions not yet included in previous studies, such as nonnormality and cross-validation design.

Chapter III

Method

A Monte Carlo simulation study was conducted to compare the performance of the two-sample cross-validation indices to the single-sample indices in terms of selecting the correctly specified confirmatory factor analysis model. Several conditions were varied, including sample size, factor loading size, model misspecification, nonnormality, and cross-validation design. The performance of these cross-validation indices was evaluated under each condition by examining how often each index identified the correct model as the one with the most predictive validity, or how generalizable the model will be in other samples from the same population.

Study Design

Cross-Validation Indices

The following seven cross-validation indices were calculated and compared under each condition:

1) the single-sample AIC (Cudeck & Browne, 1983);
2) the single-sample CAIC (Bozdogan, 1987);
3) the single-sample BCI (Browne & Cudeck, 1989);
4) the multiple-sample *C* (De Gooijer & Koopman, 1988);
5) the two-sample CVI (Cudeck & Browne, 1983);
6) the "pseudo" single-sample *C** (De Gooijer & Koopman, 1988); and

7) the "pseudo" single-sample $\overline{C}$ * (De Gooijer & Koopman, 1988).

Sample Size (n)

Large sample sizes are necessary when testing structural equation models. Sample sizes of at least 100 are recommended in order to be fairly confident in the statistical results of model testing (Kline, 1998; Loehlin, 1998). However, sample sizes of 200 are recommended in order to avoid frequent convergence failure and improper solution problems (Loehlin, 1998). In the current study, the small sample size condition was represented by a sample of 200 generated observations. Medium and large sample size conditions were represented by samples of 500 and 1,000 generated observations, respectively. The choice of these sample sizes is based on Breckler's (1990) review of 72 structural equation modeling studies published in personality and social psychology journals in which the range of sample sizes used in these studies was between 40 and 8,650. Twenty-five percent of these studies used sample sizes greater than 500 and approximately 50% of the studies used sample sizes between 100 and 500. Thus, data were generated for sample sizes of 200, 500, and 1,000 to represent a range of realistic sample sizes.

Factor Loading Size

Unstandardized factor loading sizes of .40, .60, and .80 were used in Bandalos (1993) with corresponding error variances of .80, .60, and .40, respectively. Bandalos (1993) found that the cross-validation indices did not vary much across the different factor models when factor loadings were low (.40), but began to vary as factor

loading size increased. In addition, factor loadings of .40 may be considered by some as being too low in terms of factor reliability, making cross-validation a moot point. In the current study, then, standardized factor loading sizes of .50 and .70 were examined.

Model Misspecification

The confirmatory factor models used in the current study consist of five different models that ascend in terms of model complexity (see Appendices B through F). Four of the five models were taken from Bandalos (1993; Models 1 through 4). The correct factor model in the current study, however, is less parameterized than the correct factor model used in Bandalos (1993; Model 1). In the current study, the correct factor model consists of three correlated factors, each with four indicators. Two of the variables each load on two of the three factors, with one primary loading and one secondary loading (see Appendix D; Model 3). The additional four models are misspecified by incorrectly adding 2 or 4 secondary loadings to the correct model, omitting two secondary loadings from the correct model, or setting the correlation between two factors equal to 1.0. Primary factor loadings depend upon the factor loading size condition as described above (either .50 or .70). The secondary loadings are equal to .30 and the factors are correlated at .40. Residual variances were dependent upon the factor loading condition described above. When factor loading size was .50, residual variances were .75 for the variables with only a primary loading on a factor and .54 for the variables with a secondary loading on a factor. When factor loading size was .70, residual variances were .51 for the variables with only a

primary loading on a factor and .252 for the variables with a secondary loading on a factor. Residual variances were modeled but are not shown in the model figures for simplicity. Error paths and factor variances were set to equal 1.0.

Distribution of Observed Variables

Three population distributions were generated in which the univariate skew and kurtosis of the observed variables were: 1) normally distributed (0, 0); 2) moderately nonnormal (2, 7); and 3) extremely nonnormal (3, 25). The selection of these distributions is partially based on Kline's (1998, p. 82) description of extreme levels of skew and kurtosis.

Cross-Validation Design

All of the cross-validation indices were calculated using both the simple and double cross-validation designs in order to compare which design provides more accurate indices, resulting in a total of 14 cross-validation indices. Because the definitions and calculations of the various cross-validation indices differ (viz., two-sample, multiple-sample, single-sample, "pseudo" single-sample), the application of the simple and double cross-validation designs to these various indices may not be translated identically in their traditional sense.

Simple Cross-Validation. The single-sample cross-validation indices (AIC, CAIC, and BCI) were simply calculated using the entire sample:

$$\text{AIC} = F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k/(n-1),$$

$$\text{CAIC} = F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + q_k[\ln(n)+1]/(n-1),$$

and

$$\text{BCI} = F_{\text{ML}}[\mathbf{S}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k/(n-p-2),$$

respectively. For the multiple-sample cross-validation index proposed by De Gooijer and Koopman (1988), *C*, the simple cross-validation design consists of randomly splitting the data into two subsamples (L = 2). The model was fitted to the data in each subsample (c_1 and c_2) and then cross-validated in the opposing subsample (v_1 and v_2), yielding two cross-validation indices which were then summed:

$$C = \sum_{i=1}^{\text{L}} \text{F}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|c_i})].$$

It is important to note that this is the sum of the double cross-validation indices obtained when using the two-sample CVI. For the two-sample CVI, the sample was randomly split into one calibration and one validation sample. The model was fitted to the data in the calibration sample, and the discrepancy between the validation sample covariance matrix and the implied covariance matrix reproduced from the calibration sample was calculated:

$$\text{CVI} = \text{F}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})].$$

For the "pseudo" single-sample indices proposed by De Gooijer and Koopman (1988), *C** and $\overline{C}$ *, the simple cross-validation design consists of fitting the model to the entire sample and then randomly splitting the sample into two subsamples (L = 2). The sum of the discrepancies between each subsample's covariance matrix

($\mathbf{S}_{v_1}$ and $\mathbf{S}_{v_2}$) and the implied covariance matrix from the entire sample, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)$, were then summed:

$$C^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k \mathrm{L}^2/(n-1)$$

and

$$\overline{C}^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k \mathrm{L}^2/(n-p-2).$$

Double Cross-Validation. For the single-sample cross-validation indices (AIC, CAIC, and BCI), the sample was randomly split into two subsamples (*a* and *b*). The model was then fitted to the data in both subsamples. The discrepancy between the sample covariance matrix for each subsample ($\mathbf{S}_a$ and $\mathbf{S}_b$) and the opposing subsample's implied covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})$, respectively, was then calculated:

$$\mathrm{AIC_a} = F_{\mathrm{ML}}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})] + 2q_k/(n-1)$$

and

$$\mathrm{AIC_b} = F_{\mathrm{ML}}[\mathbf{S}_a; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})] + 2q_k/(n-1);$$

$$\mathrm{CAIC_a} = F_{\mathrm{ML}}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})] + q_k[\ln(n)+1]/(n-1)$$

and

$$\mathrm{CAIC_b} = F_{\mathrm{ML}}[\mathbf{S}_a; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})] + q_k[\ln(n)+1]/(n-1);$$

$$\mathrm{BCI_a} = F_{\mathrm{ML}}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})] + 2q_k /(n - p - 2)$$

and

$$\mathrm{BCI_b} = F_{\mathrm{ML}}[\mathbf{S}_a; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})] + 2q_k /(n - p - 2);$$

respectively. This yields two cross-validation indices for each method, which were then averaged to represent the double cross-validation index for the AIC, CAIC, and the BCI, respectively. The double cross-validation design for the multiple-sample *C* consists of randomly splitting the entire sample into four equal subsamples (L = 4). The model was fitted to the data in a calibration sample comprised of all the subsamples (c_i) except the one that is "left-out," which is subsequently employed as the validation sample (v_i). This process was repeated for each of the subsamples, leaving out a different subsample (v_i) each time until each of the 4 subsamples has been employed as a validation sample (v_i, i = 1…4). The discrepancy between the covariance matrix for each of the validation samples, $\mathbf{S}_{v_i}$, and the implied covariance matrix from each of the calibration samples, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|c_i})$, was calculated, yielding four cross-validation indices which were then summed:

$$C = \sum_{i=1}^{\mathrm{L}} \mathrm{F}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|c_i})].$$

For the two-sample CVI, the sample was randomly split into two subsamples (*a* and *b*). The model was then fitted to the data in both subsamples. The discrepancy between the sample covariance matrix for each subsample ($\mathbf{S}_a$ and $\mathbf{S}_b$) and the

opposing subsample's implied covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})$, respectively, was then calculated:

$$\mathrm{CVI_a} = \mathrm{F}[\mathbf{S}_b; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|a})]$$

and

$$\mathrm{CVI_b} = \mathrm{F}[\mathbf{S}_a; \boldsymbol{\Sigma}(\boldsymbol{\theta}_{k|b})].$$

This yields two cross-validation indices, $\mathrm{CVI_a}$ and $\mathrm{CVI_b}$, which were then averaged. It is important to note that the two cross-validation indices, $\mathrm{CVI_a}$ and $\mathrm{CVI_b}$, are the same as the two indices obtained from the multiple-sample *C* in the simple cross-validation design. For the "pseudo" single-sample *C** and $\overline{C}$*, the model was fitted to the entire sample, which was subsequently randomly split into four equal subsamples (L = 4). The discrepancy between each subsample's covariance matrix ($\mathbf{S}_{v_1}$ through $\mathbf{S}_{v_4}$) and the implied covariance matrix from the entire sample, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)$, was then summed:

$$C^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k\mathrm{L}^2/(n-1)$$

and

$$\overline{C}^* = \sum_{i=1}^{\mathrm{L}} F_{\mathrm{ML}}[\mathbf{S}_{v_i}; \boldsymbol{\Sigma}(\boldsymbol{\theta}_k)] + 2q_k\mathrm{L}^2/(n-p-2),$$

respectively.

Study Design Overview

The conditions examined in the current study were fully crossed [7 (cross-validation indices) x 5 (model misspecification) x 2 (cross-validation design) x 3 (sample size) x 3 (distribution type) x 2 (factor loading size)], with repeated measures on the cross-validation measure. Thus, each of the seven cross-validation indices was calculated in each of the 180 conditions (see Table 1).

Table 1

Dimensions of the Study Design

Cross-Validation Indices (7 Levels)

1. Single-Sample AIC
2. Single-Sample CAIC
3. Single-Sample BCI
4. Multiple-Sample *C*
5. Two-Sample CVI
6. "Pseudo" Single-Sample *C**
7. "Pseudo" Single-Sample $\overline{C}$ *

Sample Size (3 Levels)

1. $N = 200$
2. $N = 500$
3. $N = 1{,}000$

Factor Loading Size (2 Levels)

1. $\lambda = .50$
2. $\lambda = .70$

Model Misspecification (5 Levels)

1. Model 1 (see Appendix B)
2. Model 2 (see Appendix C)
3. Model 3 (Correct Model; see Appendix D)
4. Model 4 (see Appendix E)
5. Model 5 (see Appendix F)

Distribution of Observed Variables (3 Levels)

1. Normal: (0, 0)
2. Moderately Nonnormal: (2, 7)
3. Extremely Nonnormal: (3, 25)

Cross-Validation Design (2 Levels)

1. Simple Cross-Validation
2. Double Cross-Validation

Data Generation

The EQS Structural Program Software (Bentler, 1995) was used to generate raw data according to specific sample size, nonnormality, and model structure conditions (see Table 2). Using the covariance matrix implied by the model parameters (see Table 3), EQS generated univariate normal and nonnormal observed variables with specified skew and kurtosis based on the Vale and Maurelli (1983) method which incorporates formulae developed by Fleishman (1978).

Table 2

Data Generation Characteristics

	Distribution Type	Factor Loading Size
Sample Size		
N = 200	(0, 0)	λ = .50 and .70
	(2, 7)	λ = .50 and .70
	(3, 25)	λ = .50 and .70
N = 500	(0, 0)	λ = .50 and .70
	(2, 7)	λ = .50 and .70
	(3, 25)	λ = .50 and .70
N = 1,000	(0, 0)	λ = .50 and .70
	(2, 7)	λ = .50 and .70
	(3, 25)	λ = .50 and .70

Table 3

Generating Population Covariance Matrix as a Function of Factor Loading Size

Generating Covariance Matrix: Factor Loading (λ) = .50

$$\begin{bmatrix}
1 & .310 & .250 & .250 & .100 & .100 & .100 & .100 & .100 & .100 & .100 & .250 \\
.310 & 1 & .310 & .310 & .160 & .160 & .160 & .160 & .250 & .250 & .250 & .436 \\
.250 & .310 & 1 & .250 & .100 & .100 & .100 & .100 & .100 & .100 & .100 & .250 \\
.250 & .310 & .250 & 1 & .100 & .100 & .100 & .100 & .100 & .100 & .100 & .250 \\
.100 & .160 & .100 & .100 & 1 & .250 & .250 & .250 & .100 & .100 & .100 & .160 \\
.100 & .160 & .100 & .100 & .250 & 1 & .250 & .250 & .100 & .100 & .100 & .160 \\
.100 & .160 & .100 & .100 & .250 & .250 & 1 & .250 & .100 & .100 & .100 & .160 \\
.100 & .160 & .100 & .100 & .250 & .250 & .250 & 1 & .100 & .100 & .100 & .160 \\
.100 & .250 & .100 & .100 & .100 & .100 & .100 & .100 & 1 & .250 & .250 & .310 \\
.100 & .250 & .100 & .100 & .100 & .100 & .100 & .100 & .250 & 1 & .250 & .310 \\
.100 & .250 & .100 & .100 & .100 & .100 & .100 & .100 & .250 & .250 & 1 & .310 \\
.250 & .436 & .250 & .250 & .160 & .160 & .160 & .160 & .310 & .310 & .310 & 1
\end{bmatrix}$$

Generating Covariance Matrix: Factor Loading (λ) = .70

$$\begin{bmatrix}
1 & .574 & .490 & .490 & .196 & .196 & .196 & .196 & .196 & .196 & .196 & .406 \\
.574 & 1 & .574 & .574 & .280 & .280 & .280 & .280 & .406 & .406 & .406 & .652 \\
.490 & .574 & 1 & .490 & .196 & .196 & .196 & .196 & .196 & .196 & .196 & .406 \\
.490 & .574 & .490 & 1 & .196 & .196 & .196 & .196 & .196 & .196 & .196 & .406 \\
.196 & .280 & .196 & .196 & 1 & .490 & .490 & .490 & .196 & .196 & .196 & .280 \\
.196 & .280 & .196 & .196 & .490 & 1 & .490 & .490 & .196 & .196 & .196 & .280 \\
.196 & .280 & .196 & .196 & .490 & .490 & 1 & .490 & .196 & .196 & .196 & .280 \\
.196 & .280 & .196 & .196 & .490 & .490 & .490 & 1 & .196 & .196 & .196 & .280 \\
.196 & .406 & .196 & .196 & .196 & .196 & .196 & .196 & 1 & .490 & .490 & .574 \\
.196 & .406 & .196 & .196 & .196 & .196 & .196 & .196 & .490 & 1 & .490 & .574 \\
.196 & .406 & .196 & .196 & .196 & .196 & .196 & .196 & .490 & .490 & 1 & .574 \\
.406 & .652 & .406 & .406 & .280 & .280 & .280 & .280 & .574 & .574 & .574 & 1
\end{bmatrix}$$

Procedure

For each of the 18 combinations of sample size, distribution type, and factor loading size, 1,000 sets of raw data were generated. The seed numbers employed in the random number generator of EQS that was used for each data generation condition was documented in order to subsequently reproduce the same sets of data if necessary. Each of the samples were randomly split according to the cross-validation design used, except when the single-sample cross-validation indices (AIC, CAIC, and BCI) were employed in the simple cross-validation design. After randomly splitting the data into two or four subsamples, the cross-validation indices were calculated using the same split data sets. Each of the proposed models was fitted to the appropriate data set using ML estimation. Because nonnormality was introduced among the observed variables, the Satorra-Bentler scaled statistic was used to obtain robust χ^2 statistics and standard errors. Each of the models was then cross-validated according to the cross-validation design employed. Cross-validation indices were calculated based on the sample and implied covariance matrices.

Maximum likelihood parameter estimates, maximum likelihood standard error estimates, and Satorra-Bentler standard error estimates obtained from fitting the correct confirmatory factor model to the full sample data set were saved to external files. The maximum likelihood and Satorra-Bentler chi-square test statistics obtained for the correct confirmatory factor model in the full sample data set were also saved to external files. The rationale for keeping the parameter estimates, standard error

estimates, and chi-square test statistics obtained from fitting the models to the full sample data set is that researchers typically examine the parameters and test statistics estimated in the entire data set as opposed to split-sample data sets.

If memory allocation problems were encountered in a replication, which indicated that the computer was unable to allocate enough memory to use the EQS software program to fit all of the models to the appropriate data set, the condition in which this occurred was documented. If improper solutions were encountered, indicating that the correlations estimated may be greater than 1.0 or that the variances estimated may be negative or equal to zero (Byrne, 1994), the condition and model in which this occurred was documented. In addition, if the estimates failed to converge on the model parameters, the condition and model in which this occurred was documented. When any of the memory allocation, improper solution, or convergence problems occurred in any replication, new data were generated accordingly and the model parameters were estimated again. This process was repeated until a total of 1,000 solutions had converged without improper solutions or memory allocation problems for each condition. The number of iterations allowed for estimates to converge was set to be 100.

Data Analysis

Performance of Cross-Validation Indices

For each condition, the number of times out of 1,000 that each cross-validation index selected the correct model was documented. These hit rates were based on 1,000 converged solutions. Because the actual values of each cross-

validation index are not all on the same scale of measurement, traditional mean difference tests (i.e., Analysis of Variance) were not used to examine the performance of the cross-validation indices.

Parameter and Standard Error Bias

Recall that maximum likelihood estimation may render asymptotically incorrect covariance matrices and inefficient parameter estimates under conditions of nonnormality among the observed variables (Bollen, 1989). Because nonnormality among the observed variables was introducted in the current study, it is important to examine parameter and standard error estimate bias. Thus, additional descriptive analyses were conducted to examine relative bias in the parameter and standard error estimates for the correct, generating factor model (see Appendix D). Bias in the maximum likelihood standard error estimates and in the Satorra-Bentler scaled standard error estimates, which are corrected for nonnormality, were calculated for comparison reasons.

The formula for parameter relative bias is:

$$B(\hat{\theta}_i) = \frac{\bar{\hat{\theta}}_i - \theta_i}{\theta_i}, \; i = 1, 2, \ldots, \; t \, ,$$

where θ_i is the value of the *i*th population parameter ($\theta_i \neq 0$), $\bar{\hat{\theta}}_i$ is the mean of the *i*th parameter estimates across the 1,000 replications, and *t* is the number of parameters estimated. The formula for standard error relative bias is:

$$B(\hat{s}e_{\hat{\theta}_i}) = \frac{\bar{\hat{s}}e_{\hat{\theta}_i} - \hat{s}e_{\theta_i}}{\hat{s}e_{\theta_i}}, i = 1, 2, \ldots, t,$$

where $\hat{s}e_{\hat{\theta}_i}$ is the estimated value of the population standard error for $\hat{\theta}_i$, and $\bar{\hat{s}}e_{\hat{\theta}_i}$ is the mean of the standard error estimates for $\hat{\theta}_i$ across the number of replications. The estimated value of the population standard error ($\hat{s}e_{\hat{\theta}_i}$) for $\hat{\theta}_i$ was estimated by the standard deviation of $\hat{\theta}_i$ over the 1,000 replications.

Chi-Square Test Statistics

Recall that maximum likelihood chi-square test statistics may become inaccurate under conditions of nonnormality among the observed variables (Kline, 1998). Because nonnormality was introduced among the observed variables in the present study, it is important to examine the effects of nonnormality on the commonly used maximum likelihood chi-square test statistic as well as the Satorra-Bentler chi-square test statistic, which is corrected for nonnormality. Thus, maximum likelihood and Satorra-Bentler chi-square rejection rates for the correct confirmatory model were calculated in each condition at the nominal alpha level of 0.05.

Memory Allocation, Improper Solution, and Convergence Problems

Memory allocation problems occurred at times when fitting the models to the appropriate data set using the EQS software program. In addition, improper solution and convergence problems were sometimes encountered when fitting the factor models to the appropriate data set. The number of times a memory allocation,

improper solution, or convergence problem was encountered was documented within each condition.

Chapter IV

Results

This chapter first presents the results concerning the performance of the cross-validation indices when selecting the correct factor model. The results of the parameter and standard error estimate bias are presented next, followed by the performance of the chi-square test statistics. Memory allocation, improper solution, and convergence problems are subsequently presented.

Performance of Cross-Validation Indices

The number of times (out of 1,000 replications) each cross-validation index selected each factor model was documented in each condition. To date, nonnormality among the observed variables and cross-validation design are conditions not yet examined in the relevant literature. Due to this, the results are displayed in Tables 4 through 9 as a function of nonnormality and cross-validation design for each factor loading and sample size condition.

Single-Sample AIC

Simple Cross-Validation Design. The AIC correctly selected model 3 more frequently than the misspecified models in each condition. When the AIC did not select the correct factor model in a replication, however, it tended to select either the most parameterized model 1 or the more parameterized model 2. The percentage of times the AIC selected the correct factor model tended to increase slightly as factor loading and sample size increased, but tended to decrease as nonnormality increased.

Double Cross-Validation Design. The AIC correctly selected model 3 more frequently than the misspecified models in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it incorrectly selected the less parameterized model 4 more frequently (see Table 4). When the AIC did not select the correct factor model in a replication, it tended to select the less parameterized model 4. The percentage of times the AIC selected the correct model tended to increase as factor loading and sample size increased, but tended to decrease as nonnormality increased.

Single-Sample CAIC

Simple Cross-Validation Design. The CAIC correctly selected model 3 more often than the misspecified models in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the correct model and the less parameterized model 4 almost equally (see Table 4). When the CAIC did not select the correct model in a replication, it tended to select the less parameterized model 4. The percentage of times the CAIC selected the correct model tended to increase as factor loading and sample size increased, but tended to decrease slightly as nonnormality increased.

Double Cross-Validation Design. The CAIC incorrectly selected the less parameterized models 4 and 5 more frequently in the low factor loading (.5) and small sample size (N = 200) condition (see Table 4). In the low factor loading (.5) and medium sample size (N = 500) condition, the CAIC incorrectly selected the less parameterized model 4 more frequently (see Table 5). Under the extremely

nonnormal distribution condition in the high factor loading (.7) and small sample size (N = 200) condition, the CAIC selected the correct model and the less parameterized model 4 almost equally (see Table 7). The CAIC did, however, correctly select model 3 more often than the misspecified models in the remaining conditions. When the CAIC did not select the correct factor model in a replication, it tended to select the less parameterized model 4. The percentage of times the CAIC selected the correct model tended to increase as factor loading and sample size increased, but tended to decrease as nonnormality increased.

Single-Sample BCI

Simple Cross-Validation Design. The BCI correctly selected model 3 more frequently than the misspecified models in each condition. When the BCI did not select the correct factor model in a replication, however, it tended to select either the most parameterized model 1 or the more parameterized model 2. The percentage of times the BCI selected the correct model tended to increase slightly as factor loading and sample size increased, but tended to decrease as nonnormality increased.

Double Cross-Validation Design. The BCI correctly selected model 3 more often than the misspecified models in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it incorrectly selected the less parameterized model 4 more frequently (see Table 4). When the BCI did not select the correct factor model in a replication, it tended to select the less parameterized model 4. The percentage of times the BCI selected the correct model

tended to increase as factor loading and sample size increased, but tended to decrease as nonnormality increased.

Multiple-Sample C

Simple Cross-Validation Design. The *C* correctly selected model 3 more frequently than the misspecified models in each condition, with the exception of the low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the correct model and the less parameterized model 4 almost equivalently (see Table 4). When the *C* did not select the correct factor model in a replication, it tended to generally select either the most parameterized model 1 or the more parameterized model 2. The percentage of times the *C* selected the correct model tended to increase as sample size increased in the low factor loading size (.5) condition (see Tables 4 through 6), but remained fairly comparable across sample size conditions in the high factor loading size (.7) condition (see Tables 7 through 9). In addition, the percentage of times the *C* selected the correct model tended to decrease slightly as nonnormality increased.

Double Cross-Validation Design. The *C* correctly selected model 3 more often than the misspecified models in each condition. When the *C* did not select the correct model in a replication, however, it generally tended to select either the most parameterized model 1 or the more parameterized model 2 and, in the small sample size conditions, the less parameterized model 4. The percentage of times the *C* correctly selected model 3 increased slightly as factor loading and sample size increased, but tended to decrease slightly as nonnormality increased.

Two-Sample CVI

Simple Cross-Validation Design. The CVI correctly selected model 3 more frequently than the misspecified models in each condition, with the exception of the low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the correct model and the less parameterized model 4 almost equivalently (see Table 4). When the CVI did not select the correct model in a replication, it generally tended to select either the most parameterized model 1, the more parameterized model 2, or the less parameterized model 4. The percentage of times the CVI selected the correct model tended to increase slightly as factor loading and sample size increased, but tended to decrease slightly as nonnormality increased.

Double Cross-Validation Design. The CVI correctly selected model 3 more often than the misspecified models in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the correct model and the less parameterized model 4 almost equally (see Table 4). When the CVI did not select the correct model in a replication, it tended to generally select either the most parameterized model 1 or the more parameterized model 2. The percentage of times the CVI selected the correct model tended to increase slightly as sample size increased in the low factor loading size (.5) condition (see Tables 4 through 6), but remained fairly comparable across sample size conditions in the high factor loading size (.7) condition (see Tables 7 through 9). In addition, the percentage of times the CVI selected the correct factor model tended to slightly decrease as nonnormality increased.

*"Pseudo" Single-Sample C**

Simple Cross-Validation Design. The *C** selected the correct model more frequently than the misspecified models in each condition. When the *C** did not select the correct model in a replication, it tended to generally select the most parameteized model 1 or the more parameterized model 2, with the exception of the low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the less parameterized model 4 (see Table 4). The percentage of times the *C** selected the correct factor model tended to increase as factor loading and sample size increased, but tended to slightly decrease as nonnormality increased.

Double Cross-Validation Design. The *C** correctly selected model 3 in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it incorrectly selected the less parameterized model 4 more often (see Table 4). When the *C** did not select the correct model in a replication, it generally tended to select the less parameterized model 4. The percentage of times the *C** selected the correct model generally increased as factor loading and sample size increased, but tended to slightly decrease as nonnormality increased.

"Pseudo" Single-Sample $\overline{C}$ *

Simple Cross-Validation Design. The $\overline{C}$ * correctly selected model 3 more frequently than the misspecified models in each condition. When the $\overline{C}$ * did not select the correct factor model in a replication, however, it tended to select the most parameterized model 1 or the more parameterized model 2, with the exception of the

low factor loading (.5) and small sample size (N = 200) condition in which it tended to select the less parameterized model 4 (see Table4). The percentage of times the $\overline{C}$ * selected the correct factor model tended to increase as factor loading and sample size increased, but generally decreased as nonnormality increased.

Double Cross-Validation Design. The $\overline{C}$ * correctly selected model 3 in all of the conditions except in the low factor loading (.5) and small sample size (N = 200) condition in which it incorrectly selected the less parameterized model 4 more frequently (see Table 4). When the $\overline{C}$ * did not select the correct model in a replication, it tended to select the less parameterized model 4. The percentage of times the $\overline{C}$ * selected the correct model generally increased as factor loading and sample size increased, but tended to slightly decrease as nonnormality increased.

Table 4

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 200 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	5.2	0.0	4.0	5.6	8.4	0.4	0.3	0.1	0.0	0.0	5.7	5.6	0.0	0.0
Model 2	13.1	0.1	11.9	12.3	15.5	1.6	1.0	0.8	0.0	0.8	11.6	12.3	0.1	0.0
Model 3	75.5	44.2	76.3	50.1	43.2	68.4	66.2	34.5	6.1	30.3	64.8	50.1	27.7	24.6
Model 4	6.2	55.2	7.8	30.4	28.9	29.6	32.5	59.1	64.9	62.5	17.7	30.4	70.6	72.9
Model 5	0.0	0.5	0.0	1.6	4.0	0.0	0.0	5.5	29.0	6.4	0.2	1.6	1.6	2.5
Moderately Nonnormal Distribution (2, 7)														
Model 1	10.6	0.0	9.0	5.1	10.0	1.1	0.5	0.4	0.0	0.1	6.2	5.1	0.0	0.0
Model 2	17.9	0.4	16.8	11.8	14.2	3.6	3.0	1.2	0.0	0.6	11.2	11.8	0.1	0.1
Model 3	64.8	49.5	66.7	38.5	34.0	69.0	67.2	29.1	7.3	26.5	54.2	38.5	32.8	30.4
Model 4	6.7	49.4	7.5	37.6	31.7	26.3	29.2	53.4	54.9	55.1	27.8	37.6	65.0	66.5
Model 5	0.0	0.7	0.0	7.0	10.1	0.0	0.1	15.9	37.8	17.7	0.6	7.0	2.1	3.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	12.7	0.2	11.0	5.1	8.5	1.9	1.3	0.3	0.0	0.3	4.6	5.1	0.2	0.2
Model 2	17.6	1.0	16.3	9.1	10.3	5.6	5.1	1.6	0.0	1.3	10.0	9.1	0.2	0.1
Model 3	63.3	52.7	65.1	27.8	26.5	68.7	66.5	21.1	7.3	19.8	39.4	27.8	39.2	35.9
Model 4	6.4	45.2	7.6	37.8	34.2	23.7	27.0	47.6	44.8	48.0	37.5	37.8	58.3	60.9
Model 5	0.0	0.9	0.0	20.2	20.5	0.1	0.1	29.4	47.9	30.6	8.5	20.2	2.1	2.9

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Table 5

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 500 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	4.8	0.0	4.2	8.3	10.9	0.1	0.1	0.0	0.0	0.0	6.7	8.3	0.0	0.0
Model 2	13.6	0.1	13.1	15.6	19.0	2.5	2.3	0.5	0.0	0.5	16.0	15.6	0.1	0.1
Model 3	81.2	86.1	82.3	69.7	59.9	94.6	94.7	80.2	29.5	79.4	76.2	69.7	82.2	80.6
Model 4	0.4	13.8	0.4	6.4	10.2	2.8	2.9	19.3	70.4	20.1	1.1	6.4	17.7	19.3
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Moderately Nonnormal Distribution (2, 7)														
Model 1	8.7	0.0	7.9	6.1	9.8	0.8	0.5	0.0	0.0	0.0	5.8	6.1	0.0	0.0
Model 2	15.0	0.2	14.5	14.8	17.4	2.7	2.6	1.0	0.0	0.7	12.8	14.8	0.0	0.0
Model 3	75.6	84.5	76.9	63.7	53.0	92.4	92.5	68.9	32.0	68.7	75.4	63.7	81.0	79.7
Model 4	0.7	15.3	0.7	15.4	18.9	4.1	4.4	30.1	66.6	30.6	6.0	15.4	19.0	20.3
Model 5	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	18.1	0.2	16.7	6.4	9.2	2.4	2.2	0.2	0.0	0.2	8.1	6.4	0.1	0.1
Model 2	18.3	1.2	17.8	13.2	15.5	6.3	6.2	1.9	0.0	1.8	14.2	13.2	0.8	0.6
Model 3	61.9	81.0	63.7	50.9	44.5	85.5	85.3	53.0	24.5	52.5	62.7	50.9	77.8	76.1
Model 4	1.7	17.6	1.8	27.9	26.3	5.8	6.3	41.8	65.8	42.4	14.7	27.9	21.3	23.2
Model 5	0.0	0.0	0.0	1.6	4.5	0.0	0.0	3.1	9.7	3.1	0.3	1.6	0.0	0.0

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Table 6

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 1000 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	5.6	0.0	5.4	8.1	11.4	0.3	0.3	0.0	0.0	0.0	6.2	8.1	0.0	0.0
Model 2	12.4	0.0	12.3	17.5	19.9	2.1	2.1	0.7	0.0	0.6	15.4	17.5	0.0	0.0
Model 3	82.0	99.7	82.3	74.2	66.1	97.6	97.6	98.5	82.4	98.6	78.4	74.2	99.7	99.5
Model 4	0.0	0.3	0.0	0.2	2.6	0.0	0.0	0.8	17.6	0.8	0.0	0.2	0.3	0.5
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Moderately Nonnormal Distribution (2, 7)														
Model 1	11.0	0.0	10.9	9.8	12.6	0.7	0.7	0.2	0.0	0.1	7.1	9.8	0.0	0.0
Model 2	14.7	0.0	14.5	16.6	20.0	4.1	4.0	1.4	0.0	1.2	14.8	16.6	0.1	0.0
Model 3	74.3	99.3	74.6	72.3	62.7	95.2	95.3	93.6	71.9	93.8	77.9	72.3	99.1	99.2
Model 4	0.0	0.7	0.0	1.3	4.7	0.0	0.0	4.8	28.1	4.9	0.2	1.3	0.8	0.8
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	15.7	0.1	15.3	7.9	10.6	1.7	1.6	0.3	0.0	0.3	7.7	7.9	0.1	0.1
Model 2	18.4	0.6	18.3	14.9	15.6	6.5	6.6	1.9	0.1	1.8	14.5	14.9	0.4	0.4
Model 3	65.9	97.4	66.4	69.0	59.8	91.6	91.6	82.8	60.2	82.9	76.0	69.0	97.7	97.7
Model 4	0.0	1.9	0.0	8.2	13.8	0.2	0.2	15.0	39.7	15.0	1.8	8.2	1.8	1.8
Model 5	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Table 7

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 200 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	5.9	0.0	4.3	8.4	11.7	0.1	0.0	0.0	0.0	0.0	7.9	8.4	0.0	0.0
Model 2	15.8	0.4	14.2	16.3	18.4	2.2	1.6	0.8	0.0	0.4	13.5	16.3	0.1	0.1
Model 3	78.3	98.3	81.5	73.8	63.9	97.3	98.0	92.9	71.2	91.6	78.4	73.8	96.3	95.8
Model 4	0.0	1.3	0.0	1.5	6.0	0.4	0.4	6.3	28.8	8.0	0.2	1.5	3.6	4.1
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Moderately Nonnormal Distribution (2, 7)														
Model 1	14.6	0.1	12.4	7.9	11.3	1.4	1.1	0.1	0.0	0.0	7.9	7.9	0.0	0.0
Model 2	19.5	1.3	18.9	16.8	19.6	6.7	5.5	2.5	0.0	2.1	17.1	16.8	0.5	0.4
Model 3	65.8	95.2	68.7	63.1	51.2	91.3	92.6	74.4	53.8	73.5	70.9	63.1	92.3	90.8
Model 4	0.0	3.4	0.0	12.1	17.4	0.6	0.8	22.8	46.0	24.2	4.1	12.1	7.2	8.8
Model 5	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.2	0.2	0.2	0.0	0.1	0.0	0.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	22.8	1.2	20.5	6.9	11.3	4.7	3.7	0.7	0.0	0.6	8.1	6.9	0.5	0.3
Model 2	21.5	3.6	20.0	11.5	14.4	9.4	8.8	3.9	0.4	3.1	14.4	11.5	1.6	1.1
Model 3	55.3	90.5	59.0	51.1	45.8	84.2	85.4	54.7	42.8	54.6	59.0	51.1	89.5	88.3
Model 4	0.4	4.7	0.5	25.8	21.7	1.7	2.1	35.1	48.9	36.0	17.2	25.8	8.4	10.3
Model 5	0.0	0.0	0.0	4.7	6.8	0.0	0.0	5.6	7.9	5.7	1.3	4.7	0.0	0.0

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Table 8

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 500 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	5.3	0.0	4.7	10.1	13.2	0.3	0.3	0.0	0.0	0.0	7.3	10.1	0.0	0.0
Model 2	12.0	0.1	11.3	18.6	20.2	1.7	1.7	0.5	0.0	0.5	15.4	18.6	0.0	0.0
Model 3	82.7	99.9	84.0	71.3	66.4	98.0	98.0	99.5	100.0	99.5	77.3	71.3	100.0	100.0
Model 4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Moderately Nonnormal Distribution (2, 7)														
Model 1	13.1	0.0	12.4	8.1	11.4	1.1	1.0	0.1	0.0	0.1	6.3	8.1	0.0	0.0
Model 2	16.3	0.3	16.0	17.2	20.5	5.4	5.3	2.2	0.0	1.9	15.4	17.2	0.2	0.2
Model 3	70.6	99.7	71.6	74.5	66.0	93.5	93.7	97.4	98.0	97.7	78.3	74.5	99.8	99.8
Model 4	0.0	0.0	0.0	0.2	2.1	0.0	0.0	0.3	2.0	0.3	0.0	0.2	0.0	0.0
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	26.8	0.2	25.5	10.0	13.7	6.0	5.7	1.0	0.0	0.8	9.2	10.0	0.0	0.0
Model 2	21.0	3.9	20.8	17.2	18.6	10.9	10.6	4.9	0.6	4.4	15.6	17.2	2.7	2.5
Model 3	52.2	95.8	53.7	65.2	56.3	83.0	83.6	83.6	81.3	84.1	72.2	65.2	97.2	97.4
Model 4	0.0	0.1	0.0	7.6	10.9	0.1	0.1	10.5	18.1	10.7	3.0	7.6	0.1	0.1
Model 5	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Table 9

Percentage of Times (Out of 1,000 Replications) Each Cross-Validation Index Selected Each Confirmatory Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 1000 as a Function of Cross-Validation Design and Distribution Type

	Simple Cross-Validation Design							Double Cross-Validation Design						
	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *	AIC	CAIC	BCI	*C*	CVI	*C**	$\overline{C}$ *
Normal Distribution (0, 0)														
Model 1	6.1	0.0	5.5	9.5	12.8	0.4	0.4	0.1	0.0	0.0	6.9	9.5	0.0	0.0
Model 2	10.9	0.0	10.9	18.1	19.3	2.0	2.0	0.6	0.0	0.5	13.7	18.1	0.0	0.0
Model 3	83.0	100.0	83.6	72.4	67.9	97.6	97.6	99.3	100.0	99.5	79.4	72.4	100.0	100.0
Model 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Moderately Nonnormal Distribution (2, 7)														
Model 1	14.3	0.0	13.8	10.9	13.7	1.2	1.1	0.2	0.0	0.1	8.4	10.9	0.0	0.0
Model 2	17.5	0.3	17.8	18.6	20.1	4.2	3.9	1.5	0.0	1.5	15.4	18.6	0.4	0.3
Model 3	68.2	99.7	68.4	70.5	66.1	94.6	95.0	98.3	100.0	98.4	76.2	70.5	99.6	99.7
Model 4	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Model 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extremely Nonnormal Distribution (3, 25)														
Model 1	25.0	0.2	24.7	9.9	13.4	4.0	3.9	0.4	0.0	0.4	6.4	9.9	0.1	0.1
Model 2	23.4	1.7	23.0	18.4	21.9	11.3	11.0	4.6	0.3	4.4	15.6	18.4	1.7	1.7
Model 3	51.6	98.1	52.3	71.3	62.0	84.7	85.1	94.5	98.1	94.7	78.0	71.3	98.2	98.2
Model 4	0.0	0.0	0.0	0.4	2.5	0.0	0.0	0.5	1.6	0.5	0.0	0.4	0.0	0.0
Model 5	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note. Model 1 = Most Parameterized, Incorrect Factor Model. Model 3 = Correct Factor Model. Model 5 = Least Parameterized Incorrect Factor Model.

Parameter and Standard Error Bias

Relative Bias of Parameter Estimates

The relative bias of each of the path loadings (λ), error variances [VAR(δ)], and factor covariances (ϕ) were calculated and are presented in Tables 10 through 15. A cutoff value of 0.05 was used in the current study as the criterion for parameter estimate bias acceptability (see Hoogland & Boomsma, 1998). Thus, relative bias for a parameter estimate was considered acceptable if the bias value was less than 0.05.

Relative bias greater than the 0.05 cutoff appeared for certain parameter estimates in the low factor loading (.5) and small sample size (N = 200) condition as nonnormality increased (see Table 10). In the high factor loading (.7) and small sample size (N = 200) condition, bias greater than 0.05 appeared for certain parameter estimates in the extremely nonnormal (3, 25) distribution condition (see Table 13). There were no distinct patterns in terms of which parameters demonstrated bias. The relative bias demonstrated by these parameter estimates was negative, indicating that the parameter estimates were underestimated in these conditions during the estimation process. The relative bias of parameter estimates was not greater than 0.05 in any of the remaining generating conditions.

Table 10

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 200 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	0.005	0.034	0.019
λ_{21}	0.015	0.023	0.005
λ_{23}	-0.021	-0.010	-0.047
λ_{31}	0.017	0.012	0.021
λ_{41}	0.012	0.013	-0.005
λ_{52}	0.016	0.023	-0.022
λ_{62}	0.014	0.009	0.006
λ_{72}	0.027	0.007	0.013
λ_{82}	0.009	0.011	0.017
λ_{93}	0.010	0.024	0.008
λ_{103}	0.011	0.028	0.023
λ_{113}	0.014	0.019	0.022
λ_{121}	0.016	-0.015	-0.066
λ_{123}	0.007	0.017	-0.009
VAR(δ_1)	-0.017	-0.030	-0.046
VAR(δ_2)	-0.022	-0.046	-0.069
VAR(δ_3)	-0.020	-0.021	-0.032
VAR(δ_4)	-0.017	-0.027	-0.055
VAR(δ_5)	-0.013	-0.015	-0.056
VAR(δ_6)	-0.014	-0.033	-0.059
VAR(δ_7)	-0.025	-0.020	-0.033
VAR(δ_8)	-0.009	-0.023	-0.048
VAR(δ_9)	-0.016	-0.029	-0.055
VAR(δ_{10})	-0.018	-0.024	-0.033
VAR(δ_{11})	-0.020	-0.026	-0.040
VAR(δ_{12})	-0.028	-0.028	-0.083
ϕ_{12}	-0.023	-0.056	-0.100
ϕ_{13}	-0.056	-0.071	-0.091
ϕ_{23}	-0.029	-0.050	-0.074

Table 11

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 500 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	-0.003	0.004	-0.007
λ_{21}	0.005	0.005	0.013
λ_{23}	-0.014	-0.012	-0.035
λ_{31}	0.000	-0.000	-0.008
λ_{41}	0.000	-0.000	-0.000
λ_{52}	0.002	0.003	-0.008
λ_{62}	0.000	-0.000	-0.005
λ_{72}	-0.004	-0.005	-0.005
λ_{82}	-0.003	0.007	0.000
λ_{93}	0.002	-0.004	-0.000
λ_{103}	0.001	0.002	-0.005
λ_{113}	0.002	-0.005	0.002
λ_{121}	-0.006	-0.007	-0.027
λ_{123}	0.004	-0.002	-0.001
VAR(δ_1)	-0.001	-0.008	0.003
VAR(δ_2)	-0.017	-0.015	-0.035
VAR(δ_3)	-0.005	-0.006	-0.012
VAR(δ_4)	-0.005	-0.010	-0.000
VAR(δ_5)	-0.004	-0.009	0.004
VAR(δ_6)	-0.005	-0.008	-0.013
VAR(δ_7)	-0.002	-0.009	-0.020
VAR(δ_8)	-0.007	-0.005	-0.004
VAR(δ_9)	-0.005	-0.002	-0.012
VAR(δ_{10})	-0.005	-0.002	-0.016
VAR(δ_{11})	-0.005	-0.008	-0.019
VAR(δ_{12})	-0.010	-0.021	-0.032
ϕ_{12}	0.000	-0.011	-0.010
ϕ_{13}	0.008	-0.011	0.003
ϕ_{23}	-0.007	-0.016	-0.011

Table 12

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .5 at Sample Sizes of 1000 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	0.003	0.002	-0.010
λ_{21}	0.010	0.001	-0.012
λ_{23}	-0.005	-0.006	0.005
λ_{31}	0.001	0.000	0.006
λ_{41}	0.001	-0.002	-0.005
λ_{52}	-0.001	-0.001	-0.003
λ_{62}	-0.002	0.005	0.007
λ_{72}	0.001	0.000	0.001
λ_{82}	0.004	0.005	-0.002
λ_{93}	-0.001	-0.002	0.000
λ_{103}	0.003	0.001	-0.005
λ_{113}	-0.000	-0.003	-0.001
λ_{121}	0.004	-0.007	-0.017
λ_{123}	0.002	0.004	0.010
VAR(δ_1)	0.000	-0.005	-0.003
VAR(δ_2)	-0.003	-0.010	-0.023
VAR(δ_3)	-0.000	-0.006	-0.004
VAR(δ_4)	-0.002	0.002	-0.007
VAR(δ_5)	-0.001	-0.010	-0.003
VAR(δ_6)	-0.004	-0.001	-0.001
VAR(δ_7)	-0.005	-0.003	-0.003
VAR(δ_8)	-0.004	0.000	-0.000
VAR(δ_9)	0.001	-0.001	-0.003
VAR(δ_{10})	-0.000	-0.003	-0.010
VAR(δ_{11})	0.001	-0.005	-0.000
VAR(δ_{12})	-0.008	-0.006	-0.007
ϕ_{12}	-0.006	0.000	-0.007
ϕ_{13}	-0.003	0.011	-0.002
ϕ_{23}	0.003	0.001	-0.003

Table 13

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 200 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	-0.006	-0.007	-0.017
λ_{21}	-0.005	-0.011	-0.043
λ_{23}	-0.006	-0.017	-0.039
λ_{31}	-0.004	-0.006	-0.015
λ_{41}	-0.005	-0.014	-0.022
λ_{52}	-0.004	0.001	-0.022
λ_{62}	0.001	-0.004	-0.009
λ_{72}	-0.001	-0.013	-0.022
λ_{82}	0.003	-0.006	-0.023
λ_{93}	0.002	-0.010	-0.020
λ_{103}	0.001	-0.001	-0.009
λ_{113}	-0.001	-0.016	-0.016
λ_{121}	-0.006	-0.027	-0.051
λ_{123}	-0.001	-0.010	-0.024
VAR(δ_1)	-0.009	-0.036	-0.066
VAR(δ_2)	-0.012	-0.030	-0.022
VAR(δ_3)	-0.011	-0.028	-0.069
VAR(δ_4)	-0.005	-0.034	-0.061
VAR(δ_5)	-0.007	-0.021	-0.070
VAR(δ_6)	-0.013	-0.012	-0.030
VAR(δ_7)	-0.007	-0.022	-0.048
VAR(δ_8)	-0.014	-0.006	-0.037
VAR(δ_9)	-0.009	-0.020	-0.063
VAR(δ_{10})	-0.006	-0.025	-0.066
VAR(δ_{11})	-0.007	-0.022	-0.050
VAR(δ_{12})	-0.019	-0.026	-0.026
ϕ_{12}	-0.001	0.003	0.000
ϕ_{13}	0.007	-0.005	0.002
ϕ_{23}	0.007	0.002	-0.002

Table 14

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 500 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	-0.002	-0.008	-0.005
λ_{21}	-0.001	-0.004	-0.008
λ_{23}	-0.004	-0.013	-0.017
λ_{31}	-0.001	-0.008	0.001
λ_{41}	-0.001	-0.006	-0.006
λ_{52}	-0.003	-0.001	-0.008
λ_{62}	-0.001	-0.001	-0.008
λ_{72}	0.000	-0.007	0.003
λ_{82}	-0.001	0.000	-0.004
λ_{93}	-0.004	-0.006	-0.005
λ_{103}	-0.001	0.002	-0.005
λ_{113}	-0.003	-0.002	-0.009
λ_{121}	-0.000	-0.006	-0.027
λ_{123}	-0.002	-0.003	-0.006
VAR(δ_1)	-0.004	-0.011	-0.016
VAR(δ_2)	-0.003	-0.032	0.011
VAR(δ_3)	-0.001	-0.008	-0.000
VAR(δ_4)	-0.002	-0.005	-0.008
VAR(δ_5)	-0.001	-0.008	-0.029
VAR(δ_6)	-0.007	-0.003	-0.012
VAR(δ_7)	-0.002	-0.007	0.008
VAR(δ_8)	-0.006	-0.005	-0.013
VAR(δ_9)	0.000	-0.006	-0.013
VAR(δ_{10})	-0.004	-0.009	-0.008
VAR(δ_{11})	-0.003	-0.007	-0.015
VAR(δ_{12})	-0.005	-0.015	-0.033
ϕ_{12}	-0.004	-0.002	0.001
ϕ_{13}	0.002	-0.006	0.008
ϕ_{23}	-0.002	-0.007	0.004

Table 15

Relative Bias of Estimated Parameters for the Correct Factor Model With Factor Loading Sizes of .7 at Sample Sizes of 1000 as a Function of Distribution Type

	Normal (0, 0)	Nonnormal (2, 7)	Nonnormal (3, 25)
Parameter			
λ_{11}	0.000	-0.003	0.000
λ_{21}	-0.000	0.001	-0.008
λ_{23}	-0.001	-0.000	-0.004
λ_{31}	-0.000	-0.000	-0.006
λ_{41}	-0.000	0.001	-0.003
λ_{52}	-0.002	0.002	-0.002
λ_{62}	0.000	-0.001	-0.005
λ_{72}	0.003	-0.002	-0.001
λ_{82}	0.000	-0.001	-0.007
λ_{93}	-0.002	0.003	-0.007
λ_{103}	-0.000	-0.001	-0.007
λ_{113}	0.001	0.001	-0.009
λ_{121}	0.002	0.006	-0.024
λ_{123}	-0.001	-0.001	-0.001
VAR(δ_1)	-0.006	-0.007	0.005
VAR(δ_2)	-0.003	-0.009	-0.009
VAR(δ_3)	0.000	-0.003	-0.010
VAR(δ_4)	-0.001	-0.006	-0.001
VAR(δ_5)	-0.002	-0.004	-0.004
VAR(δ_6)	-0.001	-0.007	-0.001
VAR(δ_7)	-0.004	-0.006	0.006
VAR(δ_8)	-0.001	-0.007	-0.003
VAR(δ_9)	0.000	-0.006	-0.001
VAR(δ_{10})	-0.002	-0.004	-0.005
VAR(δ_{11})	-0.002	0.003	-0.018
VAR(δ_{12})	-0.005	-0.007	-0.009
ϕ_{12}	0.002	-0.002	-0.001
ϕ_{13}	-0.003	-0.003	0.004
ϕ_{23}	-0.001	-0.011	0.001

Relative Bias of Maximum Likelihood (ML) and Satorra-Bentler (SB) Standard Error Estimates

The relative bias of the maximum likelihood (ML) and Satorra-Bentler (SB) standard errors for each of the path loadings (λ), error variances [V(δ)], and factor covariances (ϕ) were calculated and are shown in Tables 16 through 21. A cutoff value of 0.10 was used in the current study as the criterion for standard error estimate bias acceptability (see Hoogland & Boomsma, 1998). Thus, relative bias for a parameter estimate was considered acceptable if the bias value was less than 0.10.

The relative bias of the ML standard error estimates were generally greater than 0.10 in each of the two nonnormal distribution conditions, regardless of factor loading or sample size. The relative bias of the SB standard error estimates increased as nonnormality increased, but decreased as sample size increased. While the SB standard error estimates demonstrated less bias than the ML standard error estimates in the nonnormal distribution conditions, the SB standard error estimates still demonstrated bias greater than 0.10 in the extremely nonnormal distribution condition, regardless of factor loading or sample size. The SB standard error estimates also demonstrated bias when the observed variables were moderately nonnormal in the small sample size condition, regardless of factor loading size. In addition, three of the SB standard error estimates for the error variances demonstrated bias in the low factor loading and medium sample size condition when the observed variables were moderately nonnormal (see Table 17). The bias demonstrated by these

standard error estimates was negative, indicating that these standard error estimates were underestimated in these conditions during the estimation process. Further, there was no distinct pattern in terms of which parameters' standard error estimates demonstrated bias.

Table 16

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .5 at Sample Sizes of 200 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	-0.023	-0.034	-0.251	-0.075	-0.443	-0.179
λ_{21}	-0.043	-0.049	-0.261	-0.107	-0.456	-0.228
λ_{23}	-0.037	-0.048	-0.209	-0.099	-0.397	-0.237
λ_{31}	-0.052	-0.063	-0.259	-0.091	-0.421	-0.134
λ_{41}	-0.040	-0.052	-0.249	-0.088	-0.405	-0.123
λ_{52}	-0.037	-0.049	-0.236	-0.084	-0.407	-0.159
λ_{62}	-0.070	-0.081	-0.224	-0.071	-0.417	-0.164
λ_{72}	-0.053	-0.061	-0.233	-0.085	-0.404	-0.141
λ_{82}	-0.010	-0.022	-0.221	-0.060	-0.405	-0.141
λ_{93}	-0.009	-0.026	-0.232	-0.056	-0.449	-0.179
λ_{103}	-0.038	-0.048	-0.264	-0.102	-0.424	-0.133
λ_{113}	-0.026	-0.037	-0.274	-0.100	-0.444	-0.175
λ_{121}	-0.005	-0.018	-0.247	-0.129	-0.408	-0.248
λ_{123}	-0.051	-0.062	-0.270	-0.121	-0.453	-0.215
$V(\delta_1)$	-0.006	-0.021	-0.454	-0.120	-0.686	-0.238
$V(\delta_2)$	-0.006	-0.022	-0.384	-0.090	-0.639	-0.255
$V(\delta_3)$	-0.040	-0.054	-0.494	-0.159	-0.676	-0.216
$V(\delta_4)$	-0.055	-0.070	-0.464	-0.118	-0.668	-0.216
$V(\delta_5)$	-0.063	-0.077	-0.459	-0.114	-0.657	-0.208
$V(\delta_6)$	-0.084	-0.099	-0.459	-0.128	-0.641	-0.176
$V(\delta_7)$	-0.023	-0.037	-0.435	-0.088	-0.667	-0.221
$V(\delta_8)$	-0.015	-0.031	-0.457	-0.132	-0.678	-0.255
$V(\delta_9)$	-0.027	-0.040	-0.470	-0.125	-0.674	-0.222
$V(\delta_{10})$	0.003	-0.009	-0.473	-0.138	-0.695	-0.260
$V(\delta_{11})$	0.005	-0.012	-0.460	-0.111	-0.690	-0.258
$V(\delta_{12})$	-0.074	-0.089	-0.445	-0.163	-0.639	-0.254
ϕ_{12}	-0.049	-0.063	-0.117	-0.078	-0.193	-0.156
ϕ_{13}	-0.011	-0.022	-0.053	-0.017	-0.049	-0.032
ϕ_{23}	-0.055	-0.067	-0.149	-0.105	-0.180	-0.150

Table 17

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .5 at Sample Sizes of 500 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	-0.036	-0.039	-0.216	-0.011	-0.437	-0.123
λ_{21}	-0.049	-0.055	-0.271	-0.096	-0.450	-0.148
λ_{23}	-0.034	-0.043	-0.213	-0.065	-0.399	-0.161
λ_{31}	0.007	0.001	-0.250	-0.059	-0.478	-0.177
λ_{41}	-0.022	-0.028	-0.240	-0.053	-0.467	-0.158
λ_{52}	-0.014	-0.021	-0.193	0.003	-0.377	-0.067
λ_{62}	-0.009	-0.012	-0.242	-0.068	-0.405	-0.111
λ_{72}	-0.024	-0.030	-0.197	-0.015	-0.391	-0.082
λ_{82}	-0.015	-0.018	-0.255	-0.081	-0.394	-0.084
λ_{93}	-0.013	-0.016	-0.236	-0.040	-0.448	-0.134
λ_{103}	-0.029	-0.037	-0.241	-0.036	-0.429	-0.091
λ_{113}	0.023	0.020	-0.256	-0.066	-0.454	-0.128
λ_{121}	-0.035	-0.040	-0.222	-0.078	-0.391	-0.143
λ_{123}	-0.018	-0.023	-0.239	-0.053	-0.435	-0.130
$V(\delta_1)$	-0.017	-0.020	-0.478	-0.090	-0.702	-0.172
$V(\delta_2)$	-0.002	-0.010	-0.427	-0.082	-0.637	-0.141
$V(\delta_3)$	0.002	-0.004	-0.463	-0.063	-0.690	-0.166
$V(\delta_4)$	-0.006	-0.014	-0.440	-0.023	-0.681	-0.122
$V(\delta_5)$	-0.043	-0.049	-0.450	-0.068	-0.673	-0.131
$V(\delta_6)$	-0.037	-0.042	-0.471	-0.100	-0.676	-0.138
$V(\delta_7)$	-0.045	-0.052	-0.472	-0.102	-0.674	-0.156
$V(\delta_8)$	0.018	0.013	-0.466	-0.088	-0.663	-0.114
$V(\delta_9)$	-0.023	-0.030	-0.467	-0.062	-0.704	-0.194
$V(\delta_{10})$	-0.007	-0.016	-0.491	-0.103	-0.687	-0.159
$V(\delta_{11})$	-0.054	-0.060	-0.471	-0.082	-0.686	-0.153
$V(\delta_{12})$	-0.033	-0.040	-0.427	-0.087	-0.658	-0.167
ϕ_{12}	-0.064	-0.069	-0.113	-0.046	-0.195	-0.087
ϕ_{13}	-0.043	-0.050	-0.091	-0.031	-0.179	-0.083
ϕ_{23}	-0.027	-0.032	-0.120	-0.058	-0.221	-0.115

Table 18

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .5 at Sample Sizes of 1000 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	-0.027	-0.027	-0.196	0.018	-0.417	-0.066
λ_{21}	-0.023	-0.025	-0.234	-0.036	-0.424	-0.077
λ_{23}	-0.025	-0.028	-0.195	-0.033	-0.374	-0.086
λ_{31}	-0.007	-0.008	-0.205	0.010	-0.433	-0.064
λ_{41}	-0.008	-0.011	-0.217	-0.004	-0.464	-0.131
λ_{52}	0.022	0.018	-0.198	-0.006	-0.413	-0.088
λ_{62}	-0.006	-0.008	-0.229	-0.039	-0.417	-0.084
λ_{72}	-0.010	-0.013	-0.215	-0.024	-0.382	-0.040
λ_{82}	-0.033	-0.037	-0.193	0.002	-0.406	-0.075
λ_{93}	0.001	-0.004	-0.233	-0.028	-0.430	-0.079
λ_{103}	0.011	0.009	-0.246	-0.040	-0.406	-0.042
λ_{113}	0.011	0.009	-0.200	0.014	-0.411	-0.033
λ_{121}	0.006	0.004	-0.171	-0.006	-0.360	-0.068
λ_{123}	0.016	0.015	-0.224	-0.026	-0.449	-0.111
$V(\delta_1)$	0.005	0.005	-0.464	-0.045	-0.697	-0.116
$V(\delta_2)$	-0.036	-0.037	-0.422	-0.049	-0.663	-0.125
$V(\delta_3)$	0.033	0.031	-0.441	-0.004	-0.682	-0.075
$V(\delta_4)$	0.039	0.038	-0.492	-0.082	-0.692	-0.118
$V(\delta_5)$	0.013	0.010	-0.456	-0.058	-0.681	-0.091
$V(\delta_6)$	0.016	0.012	-0.428	0.002	-0.679	-0.111
$V(\delta_7)$	-0.005	-0.006	-0.448	-0.026	-0.670	-0.082
$V(\delta_8)$	0.003	-0.001	-0.438	-0.023	-0.675	-0.102
$V(\delta_9)$	0.001	-0.001	-0.447	-0.007	-0.690	-0.108
$V(\delta_{10})$	0.004	0.001	-0.452	-0.014	-0.681	-0.085
$V(\delta_{11})$	-0.004	-0.008	-0.462	-0.043	-0.683	-0.070
$V(\delta_{12})$	-0.022	-0.025	-0.414	-0.019	-0.677	-0.160
ϕ_{12}	-0.041	-0.043	-0.094	-0.019	-0.173	-0.039
ϕ_{13}	0.012	0.009	-0.126	-0.055	-0.214	-0.100
ϕ_{23}	0.007	0.005	-0.087	-0.012	-0.129	0.015

Table 19

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .7 at Sample Sizes of 200 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	0.024	0.014	-0.368	-0.088	-0.585	-0.182
λ_{21}	-0.015	-0.030	-0.380	-0.086	-0.584	-0.172
λ_{23}	-0.007	-0.020	-0.297	-0.112	-0.475	-0.183
λ_{31}	-0.008	-0.021	-0.380	-0.099	-0.588	-0.189
λ_{41}	-0.003	-0.015	-0.385	-0.115	-0.587	-0.177
λ_{52}	-0.021	-0.034	-0.372	-0.096	-0.577	-0.174
λ_{62}	-0.019	-0.035	-0.358	-0.081	-0.583	-0.180
λ_{72}	0.048	0.035	-0.354	-0.086	-0.565	-0.167
λ_{82}	0.006	-0.009	-0.363	-0.087	-0.573	-0.183
λ_{93}	-0.004	-0.015	-0.375	-0.098	-0.582	-0.174
λ_{103}	-0.023	-0.036	-0.392	-0.115	-0.587	-0.175
λ_{113}	0.031	0.017	-0.374	-0.093	-0.581	-0.171
λ_{121}	-0.008	-0.022	-0.285	-0.096	-0.451	-0.153
λ_{123}	0.008	-0.006	-0.375	-0.085	-0.592	-0.187
$V(\delta_1)$	-0.022	-0.035	-0.456	-0.111	-0.687	-0.228
$V(\delta_2)$	-0.006	-0.021	-0.380	-0.076	-0.637	-0.233
$V(\delta_3)$	-0.006	-0.020	-0.465	-0.126	-0.696	-0.267
$V(\delta_4)$	-0.067	-0.079	-0.464	-0.131	-0.670	-0.203
$V(\delta_5)$	0.006	-0.013	-0.454	-0.128	-0.688	-0.271
$V(\delta_6)$	-0.058	-0.076	-0.445	-0.110	-0.695	-0.280
$V(\delta_7)$	-0.006	-0.021	-0.418	-0.071	-0.684	-0.255
$V(\delta_8)$	-0.025	-0.040	-0.450	-0.102	-0.680	-0.245
$V(\delta_9)$	0.016	0.002	-0.483	-0.137	-0.692	-0.261
$V(\delta_{10})$	0.008	-0.009	-0.444	-0.084	-0.670	-0.222
$V(\delta_{11})$	0.006	-0.007	-0.442	-0.075	-0.695	-0.269
$V(\delta_{12})$	0.035	0.014	-0.393	-0.088	-0.671	-0.292
ϕ_{12}	-0.034	-0.043	-0.200	-0.116	-0.238	-0.125
ϕ_{13}	0.028	0.013	-0.124	-0.049	-0.243	-0.131
ϕ_{23}	-0.025	-0.040	-0.151	-0.069	-0.268	-0.151

Table 20

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .7 at Sample Sizes of 500 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	0.029	0.023	-0.379	-0.058	-0.615	-0.150
λ_{21}	-0.013	-0.018	-0.389	-0.052	-0.610	-0.118
λ_{23}	0.007	-0.001	-0.287	-0.064	-0.498	-0.127
λ_{31}	-0.001	-0.006	-0.364	-0.042	-0.596	-0.096
λ_{41}	-0.007	-0.012	-0.362	-0.035	-0.609	-0.150
λ_{52}	0.023	0.016	-0.368	-0.057	-0.593	-0.141
λ_{62}	0.008	0.004	-0.311	0.026	-0.594	-0.142
λ_{72}	-0.012	-0.017	-0.357	-0.050	-0.600	-0.149
λ_{82}	-0.049	-0.056	-0.366	-0.055	-0.562	-0.074
λ_{93}	0.027	0.022	-0.384	-0.069	-0.593	-0.112
λ_{103}	0.013	0.007	-0.392	-0.073	-0.613	-0.149
λ_{113}	-0.009	-0.014	-0.380	-0.051	-0.595	-0.110
λ_{121}	0.017	0.012	-0.280	-0.048	-0.484	-0.108
λ_{123}	0.001	-0.008	-0.404	-0.076	-0.616	-0.132
$V(\delta_1)$	-0.001	-0.008	-0.464	-0.059	-0.684	-0.146
$V(\delta_2)$	-0.027	-0.033	-0.402	-0.059	-0.707	-0.253
$V(\delta_3)$	-0.010	-0.016	-0.478	-0.078	-0.710	-0.174
$V(\delta_4)$	0.007	0.002	-0.487	-0.082	-0.724	-0.241
$V(\delta_5)$	0.003	-0.003	-0.441	-0.051	-0.669	-0.144
$V(\delta_6)$	-0.007	-0.015	-0.441	-0.036	-0.687	-0.175
$V(\delta_7)$	-0.016	-0.017	-0.436	-0.037	-0.706	-0.212
$V(\delta_8)$	0.010	0.003	-0.460	-0.070	-0.679	-0.148
$V(\delta_9)$	0.016	0.010	-0.460	-0.051	-0.690	-0.159
$V(\delta_{10})$	-0.040	-0.048	-0.479	-0.089	-0.706	-0.195
$V(\delta_{11})$	0.020	0.015	-0.455	-0.046	-0.703	-0.191
$V(\delta_{12})$	-0.036	-0.043	-0.393	-0.031	-0.655	-0.179
ϕ_{12}	-0.033	-0.040	-0.147	-0.034	-0.276	-0.092
ϕ_{13}	-0.024	-0.029	-0.155	-0.057	-0.268	-0.091
ϕ_{23}	-0.008	-0.016	-0.160	-0.052	-0.279	-0.087

Table 21

Relative Bias of Maximum Likelihood and Satorra-Bentler Standard Errors for the Correct Model With Factor Loading Sizes of .7 at Sample Sizes of 1000 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Parameter						
λ_{11}	0.061	0.059	-0.410	-0.078	-0.600	-0.068
λ_{21}	0.020	0.019	-0.384	-0.015	-0.613	-0.084
λ_{23}	0.021	0.018	-0.283	-0.024	-0.498	-0.074
λ_{31}	-0.015	-0.017	-0.368	-0.009	-0.583	-0.052
λ_{41}	-0.006	-0.009	-0.380	-0.029	-0.575	-0.030
λ_{52}	-0.018	-0.022	-0.339	0.011	-0.572	-0.062
λ_{62}	-0.005	-0.010	-0.349	-0.005	-0.575	-0.062
λ_{72}	0.018	0.015	-0.386	-0.060	-0.575	-0.056
λ_{82}	0.003	0.001	-0.359	-0.024	-0.569	-0.058
λ_{93}	0.007	0.005	-0.370	-0.019	-0.581	-0.032
λ_{103}	-0.022	-0.024	-0.362	-0.003	-0.600	-0.075
λ_{113}	-0.000	-0.001	-0.359	0.003	-0.590	-0.039
λ_{121}	-0.003	-0.006	-0.294	-0.040	-0.489	-0.074
λ_{123}	0.004	0.001	-0.402	-0.044	-0.607	-0.045
$V(\delta_1)$	-0.024	-0.027	-0.444	-0.007	-0.714	-0.145
$V(\delta_2)$	0.014	0.011	-0.403	-0.022	-0.699	-0.211
$V(\delta_3)$	0.076	0.073	-0.468	-0.035	-0.703	-0.129
$V(\delta_4)$	0.002	-0.001	-0.428	0.032	-0.699	-0.112
$V(\delta_5)$	0.024	0.022	-0.448	-0.036	-0.667	-0.072
$V(\delta_6)$	-0.011	-0.013	-0.470	-0.063	-0.684	-0.091
$V(\delta_7)$	-0.012	-0.016	-0.452	-0.039	-0.689	-0.107
$V(\delta_8)$	-0.043	-0.045	-0.412	0.021	-0.684	-0.099
$V(\delta_9)$	-0.019	-0.023	-0.472	-0.041	-0.707	-0.127
$V(\delta_{10})$	0.015	0.015	-0.478	-0.055	-0.701	-0.124
$V(\delta_{11})$	-0.013	-0.015	-0.477	-0.047	-0.701	-0.130
$V(\delta_{12})$	-0.015	-0.018	-0.410	-0.033	-0.699	-0.209
ϕ_{12}	0.026	0.021	-0.114	0.020	-0.289	-0.066
ϕ_{13}	0.033	0.030	-0.140	-0.021	-0.279	-0.065
ϕ_{23}	-0.013	-0.017	-0.169	-0.049	-0.305	-0.084

Chi-Square Test Statistics

Rejection rates of the maximum likelihood (ML) and Satorra-Bentler (SB) chi-square tests (χ^2) for the correct confirmatory factor model were calculated in each condition and are presented in Table 22. Chi-square rejection rates were considered acceptable if they rejected the correct factor model in only 5% of the 1,000 replications, representing a nominal alpha level of 0.05. The ML χ^2 rejection rates were generally appropriate in the normal distribution condition, with the exception of the high factor loading (.7) and small sample size (N = 200) condition in which the ML χ^2 rejected the correct model too frequently. As nonnormality increased, ML χ^2 rejection rates became higher, rejecting the correct factor model too often. The SB χ^2 performed better than the ML χ^2 in the nonnormal distribution conditions, but still tended to reject the correct model too frequently as nonnormality increased, especially in the high factor loading (.7) and small sample size (N = 200) condition. Interestingly, the SB χ^2 tended to reject the correct model more frequently than the ML χ^2 in the normal distribution condition, particularly in the small sample size condition.

Table 22

Rejection Rates (Out of 1,000 Replications) of Maximum Likelihood and Satorra-Bentler Chi-Square Tests for the Correct Factor Model With Factor Loading Sizes of .5 and .7 at Sample Sizes of 200, 500, and 1000 as a Function of Distribution Type

	Normal (0, 0)		Nonnormal (2, 7)		Nonnormal (3, 25)	
	ML	SB	ML	SB	ML	SB
Factor Loading Size = .5						
Sample Size = 200	4.6	5.7	22.9	5.4	38.3	9.2
Sample Size = 500	5.9	6.2	28.5	5.4	56.8	7.2
Sample Size = 1000	5.2	5.3	30.6	6.6	64.3	7.2
Factor Loading Size = .7						
Sample Size = 200	8.1	9.2	57.5	9.2	86.3	13.3
Sample Size = 500	5.7	6.4	57.5	6.5	94.5	7.8
Sample Size = 1000	4.9	4.9	63.0	6.3	95.2	6.4

Memory Allocation, Improper Solution, and Convergence Problems

The results discussed above are based on a total of 1,000 replications in which memory allocation problems and improper solutions did not exist, and the number of iterations needed for solutions to converge did not exceed 100 iterations. Recall that memory allocation problems indicated that the computer was unable to allocate enough memory to use the EQS software program to fit all of the models to the appropriate data set. Improper solutions indicated that the correlations estimated may be greater than 1.0 or that the variances estimated may be negative or equal to zero (Byrne, 1994). Convergence problems indicated that the estimates failed to converge on the model parameters before reaching 100 iterations. The number of times any of these problems were encountered was documented in each condition, as well as the total number of replications needed to achieve 1,000 complete replications in each condition. This information is presented in Table 23.

To summarize, the percent of usable replications increased as factor loading and sample size increased. Unusable replications were due to two general problems, the lack of sufficient memory allocation to run the EQS software program and noncovergence on the model parameters. In general, the percent of memory allocation problems remained fairly comparable across sample size conditions in the low factor loading size condition, but tended to decrease slightly as sample size increased in the high factor loading size condition. Improper solution and convergence problems were encountered more frequently in the small sample size condition, regardless of factor loading size, and tended to increase as nonnormality

increased. In the low factor loading size condition, the majority of the improper solutions and convergence problems occurred when fitting the more parameterized models to the data in the half-split samples. In the high factor loading size condition, however, the majority of the improper solution and convergence problems were encountered when fitting the less parameterized models to the data in the half-split samples.

Table 23

Memory Allocation, Improper Solution, and Convergence Problems Encountered as a Function of Generating Condition

	Replications Needed	% Usable Replications	% Memory Allocation Problems	% Improper and Convergence Problems
Factor Loading = .5	Small Sample Size (N = 200)			
Normal	1635	61.2%	7.7%	31.1%
Moderately Nonnormal	2271	44.0%	6.9%	49.1%
Extremely Nonnormal	3272	30.6%	7.8%	61.6%
Factor Loading = .5	Medium Sample Size (N = 500)			
Normal	1134	88.0%	11.0%	1.0%
Moderately Nonnormal	1116	89.6%	8.5%	1.9%
Extremely Nonnormal	1195	83.7%	9.4%	6.9%
Factor Loading = .5	Large Sample Size (N = 1000)			
Normal	1100	90.9%	9.0%	0.0%
Moderately Nonnormal	1095	91.3%	8.7%	0.0%
Extremely Nonnormal	1091	91.7%	7.9%	0.4%
Factor Loading = .7	Small Sample Size (N = 200)			
Normal	1114	89.7%	7.3%	3.0%
Moderately Nonnormal	1299	77.0%	8.7%	14.3%
Extremely Nonnormal	1816	55.1%	8.3%	36.6%
Factor Loading = .7	Medium Sample Size (N = 500)			
Normal	1071	93.4%	6.6%	0.0%
Moderately Nonnormal	1084	92.2%	7.3%	0.5%
Extremely Nonnormal	1154	86.7%	8.1%	5.2%
Factor Loading = .7	Large Sample Size (N = 1000)			
Normal	1053	95.0%	5.0%	0.0%
Moderately Nonnormal	1067	93.7%	6.3%	0.0%
Extremely Nonnormal	1074	93.1%	6.4%	0.5%

Chapter V

Discussion

This chapter discusses the results in the order in which they were presented in the previous chapter. Thus, the performance of the cross-validation indices is discussed first, followed by the parameter and standard error estimate bias results. The chi-square rejection rates are then discussed, which are succeeded by a summary of the memory allocation, improper solution, and convergence problems that were encountered. Implications of these findings and recommendations for future practice are then provided. This chapter concludes with a discussion of the limitations in the present study and suggestions for future research, as well as a general conclusion to the study.

Performance of Cross-Validation Indices

The primary question addressed in the present study was which of the several proposed cross-validation indices performs the best in terms of selecting the structural equation model with the most predictive validity, or the model that will generalize to another sample from the same population. Unfortunately, there is no single best cross-validation index that would be appropriate in all of the structural equation modeling situations examined in this study. Instead, it appears that the choice of which cross-validation index to use depends on several factors, such as factor loading size, sample size, model misspecification, cross-validation design, and nonnormality. Recommendations concerning which cross-validation index performs optimally under

these various conditions are presented in the implications and recommendations section.

The accuracy of all the cross-validation indices in terms of selecting the correct factor model generally improved as factor loading and sample size increased. This finding is similar to Bandalos (1993) in which the performance of the AIC, CAIC, BCI, and CVI became more comparable and tended to improve as factor loading and sample size increased. Nonetheless, there were performance differences between the cross-validation indices examined in the current study with respect to selecting the correct factor model. In general, the single-sample CAIC and the "pseudo" single-sample sample C^* and $\overline{C}^*$ outperformed the remaining cross-validation indices. In addition, the single-sample AIC and BCI outperformed the multiple-sample C and the two-sample CVI. These findings corroborate the results in De Gooijer and Koopman (1988) in which the C^* and $\overline{C}^*$ generally outperformed the AIC, BCI, C, and CVI in terms of selecting the correct factor model.

Model misspecification also affected the performance of the cross-validation indices with respect to selecting the correct factor model and interacted with the cross-validation design dimension for certain cross-validation indices. For instance, when the simple cross-validated AIC and BCI did not select the correct factor model in a replication, they tended to incorrectly indicate the more parameterized models as ones with predictive validity. In comparison, when the simple cross-validated CAIC did not select the correct factor model in a replication, it tended to incorrectly indicate

the less parameterized model as one with predictive validity. When the double cross-validated AIC, CAIC, and BCI did not select the correct factor model in a replication, they tended to indicate the less parameterized model as one with predictive validity. Comparable to the findings in De Gooijer and Koopman (1988), when the *C* and CVI did not select the correct factor model in a replication, they tended to indicate the more parameterized models as ones with predictive validity, regardless of cross-validation design. When the simple cross-validated C^* and $\overline{C}^*$ did not select the correct model in a replication, they tended to select the more parameterized models. In comparison, when the double cross-validated C^* and $\overline{C}^*$ did not select the correct factor model in a replication, they tended to indicate the less parameterized model as one with predictive validity.

The effect of cross-validation design on the performance of cross-validation indices had not been explicitly examined previously in the relevant literature, and several of the relevant studies only examined the single-sample indices in the double cross-validation design. Thus, the performance of the indices under both the simple and double cross-validation designs was of great interest in the present study. Based on the results in the current study, the double cross-validated indices generally performed better than their simple cross-validated counterparts with respect to selecting the correct model. For instance, in the low factor loading and large sample size condition, as well as across sample size conditions in the high factor loading size condition, the double cross-validated AIC and BCI performed better than their simple

cross-validated counterparts, selecting the correct model approximately 10 to 16 percentage points more frequently. The double cross-validated CAIC did not outperform its simple cross-validated counterpart, but did begin to perform comparably to the simple cross-validated CAIC when sample size reached 500 in the high factor loading size condition. The double cross-validated *C* and CVI performed better than their simple cross-validated counterparts in all of the conditions, selecting the correct model about 1 to 15 percentage points more often. In the low factor loading and large sample size condition, as well as in the high factor loading and medium and large sample size conditions, the double cross-validated *C** and $\overline{C}$ * performed better than their simple cross-validated counterparts, selecting the correct model approximately 2 to 14 percentage points more often.

The impact of nonnormality among the observed variables on the performance of cross-validation indices had not been examined in previous studies and was also of great interest in the present study. In the current study, nonnormality was introduced to represent normal, moderately nonnormal, or extremely nonnormal distributions of the observed variables. It was found that nonnormality does appear to generally affect the performance of the cross-validation indices. More specifically, the percentage of times the cross-validation indices selected the correct model tended to decrease as nonnormality increased.

The parameters estimated by maximum likelihood tend to be fairly robust under conditions of nonnormality, yielding consistent estimates. However, it is

possible for skewed and kurtotic observed variables to sometimes render an incorrect asymptotic covariance matrix (Bollen, 1989). More specifically, the implied covariance matrix is based on the estimated parameters, which may become biased as nonnormality increases. In the present study, as will be discussed below, the estimated parameters did tend to become slightly biased as nonnormality increased. Thus, the implied covariance matrix tended to become incorrect. All of the cross-validation indices are calculated by determining the discrepancy between the implied covariance matrix and the sample covariance matrix. Accordingly, because the implied covariance matrices became slightly more distorted and variable as nonnormality increased, the cross-validation findings also tended to become more distorted and variable.

While all of the cross-validation indices appeared to be impacted by the presence of nonnormality among the observed variables, some were more adversely affected than others. For instance, nonnormality had a greater adverse effect on the simple cross-validated AIC and BCI than on the remaining simple cross-validated indices. In addition, nonnormality had a greater adverse effect on the double cross-validated AIC, CAIC, and BCI than on the remaining double cross-validated indices.

Parameter and Standard Error Bias

Nonnormality among the observed variables, as mentioned above, appeared to have an impact on the parameter and standard error estimates. Consistent with other studies (e.g., Henly, 1993) negative parameter estimate bias in the current study tended to slightly increase as nonnormality increased, indicating that maximum

likelihood tends to slightly underestimate the parameter estimates more as nonnormality increases. The maximum likelihood and Satorra-Bentler corrected standard error estimates demonstrated negative bias as nonnormality increased, indicating an underestimation of the standard error estimates. Similar to other studies (e.g., Chou et al., 1991), however, the Satorra-Bentler corrected standard error estimates demonstrated less bias than the maximum likelihood standard error estimates under conditions of nonnormality.

Chi-Square Test Statistics

Nonnormality also affected the robustness of the chi-square test statistics in the present study. Consistent with previous research (e.g., Chou et al., 1991), the maximum likelihood chi-square test statistic tended to reject the correct factor model more often than the nominal alpha rate (0.05) as nonnormality increased. Although the Satorra-Bentler corrected chi-square test statistic also tended to reject the correct factor model more frequently than was appropriate as nonnormality increased, it performed better than the maximum likelihood chi-square.

Memory Allocation, Improper Solution, and Convergence Problems

As factor loading and sample size increased, the number of usable replications increased. Memory allocation problems were fairly comparable across sample size conditions in the low factor loading condition, but decreased slightly as sample size increased in the high factor loading condition. Consistent with previous findings (e.g., Anderson & Gerbing, 1984), the proportion of improper solutions and

convergence problems decreased as sample size increased. Nonetheless, these problems were encountered more frequently as nonnormality increased.

Implications and Recommendations

Based on the results, certain cross-validation indices performed more optimally than others in the various conditions examined in the present study with respect to selecting the correct factor model. Recommendations concerning which cross-validation indices perform more accurately in these various conditions are provided below. As previously mentioned, the cross-validation indices generally improved as factor loading and sample size increased. Due to this, recommendations are provided as a function of factor loading and sample size. It is important to note here that the true factor loading sizes will not be known in practice. Instead, the factor loadings of a structural equation model will be estimated. Nonetheless, these estimated factor loadings could be examined when deciding which cross-validation index to use.

When choosing the index to use when factor loadings are low and sample size is small, the simple cross-validated AIC and BCI, as well as the simple cross-validated C^* and $\overline{C}^*$ would be more appropriate than the remaining cross-validation indices. Under low factor loading and medium sample size conditions, the simple cross-validated CAIC and the simple cross-validated C^* and $\overline{C}^*$ would be more accurate than the remaining cross-validation indices. When factor loadings are low and sample size is large, or factor loadings are high and sample size is small, the

simple cross-validated CAIC, *C**, and $\overline{C}$ *, as well as the double cross-validated AIC, BCI, *C**, and $\overline{C}$ * would perform more optimally than the other indices in terms of selecting the correct model. In high factor loading and medium to large sample size conditions, the simple cross-validated CAIC, *C**, and $\overline{C}$ *, as well as the double cross-validated AIC, CAIC, BCI, *C**, and $\overline{C}$ * would perform more accurately.

It appears that De Gooijer and Koopman's (1988) "pseudo" single-sample *C** and $\overline{C}$ * generally perform more accurately than the remaining cross-validation indices examined. Under small sample size conditions, however, researchers should employ the simple cross-validated AIC and BCI in addition to the *C** and $\overline{C}$ * in order to be more assured that a particular model will generalize to other samples from the same population. Unfortunately, there is currently no structural equation modeling software program that provides the values of the *C** and $\overline{C}$ *. To my knowledge, only the values of the AIC and CAIC are typically provided in the output of the standard modeling software programs (e.g., EQS, LISREL). This may unfortunately limit the use of De Gooijer and Koopman's two indices when cross-validating structural equation models in practice.

Researchers should be aware that fitting structural equation models to data with nonnormal distributions among the observed variables may render distorted findings related to the model, such as cross-validation performance, parameter and standard error estimates, and chi-square rejection rates. To correct for nonnormality among the observed variables, one may want to transform the variables. At a

minimum, researchers should incorporate the Satorra-Bentler corrected statistics for more accurate standard errors and chi-square test statistics, but note that they are still possibly biased. In addition, researchers may experience improper solution or convergence problems when testing models in small sample sizes or under conditions of nonnormality among the observed variables.

Limitations and Suggestions For Future Research Directions

There are limitations inherent in any research study and the current study is no exception. This study was intended to examine the performance of the cross-validation indices using structural equation models with higher factor loadings than were examined (viz., .60 and .80 versus .50 and .70, respectively). Unfortunately, when originally testing the correct factor model with primary loadings of .8, factor intercorrelations of .4, and secondary loadings of .3, a non-positive definite matrix was encountered due to the high covariance between variables 2 and 12, which both loaded on two factors (see Appendix D). Thus, the current study is limited with respect to the magnitude of the factor loadings examined. More specifically, some may consider factor loadings of .5 as too low for construct reliability, rendering cross-validation questionable. Because researchers typically strive to establish factor models with higher loadings, future research should examine higher primary factor loadings using models for which this is possible.

This study was proposed to examine extreme conditions of nonnormality among the observed variables. Data in the present study were generated in the extremely nonnormal condition as having univariate skew of 3 and univariate kurtosis

of 25, which may be considered by some as not extreme enough, at least with regards to skew. The standard methods of generating nonnormal data according to correlation matrices are limited in terms of the levels of skew and kurtosis that may be achieved (e.g., Vale & Maurelli, 1983). Thus, future studies should incorporate other methods to generate data that may reach higher levels of skew and kurtosis.

The sample sizes investigated in the present study were selected to represent the minimum sample size in which improper solutions and convergence problems tend to disappear (i.e., $N = 200$), and to represent medium ($N = 500$) and large ($N = 1000$) sample sizes. Sample sizes of at least 100 are recommended in order to be fairly confident in the statistical results of model testing (Kline, 1998; Loehlin, 1998). Cross-validating models in sample sizes of 100 would have most likely introduced more improper solutions and convergence problems than were encountered in the present study. Nevertheless, applied researchers do test models in sample sizes of 100, and sometimes in sample sizes much smaller. Thus, future research could investigate the problems encountered when calculating cross-validation indices in smaller sample size conditions.

The benefit of using cross-validation as a model selection technique is that structural equation models need not be hierarchically related to one another. The present study examined the performance of cross-validation indices when selecting among hierarchically related structural equation models, which may be viewed by some as unnecessary since the chi-square difference test may be used to select a model in this situation. Thus, future studies should investigate the

performance of cross-validation indices when selecting among models that are not hierarchically related or nested. Nonetheless, cross-validation as a model selection technique is not exclusive to selecting non-nested structural equation models and may be used when selecting among nested models. Further, the chi-square difference test has not been investigated in terms of being able to select a model with predictive validity, or one that will generalize to other samples from the same population. Accordingly, future studies should examine the correspondence between cross-validation as a model selection technique and the chi-square difference test in terms of selecting population models when they are hierarchically related.

In the present study, the performance of the cross-validation indices when selecting among measurement models was investigated. It would be interesting to examine the performance of cross-validation indices when selecting among structural models to determine if differences exist in this situation. Thus, future research studies may want to address the issue of selecting structural models that generalize to other samples from the same population.

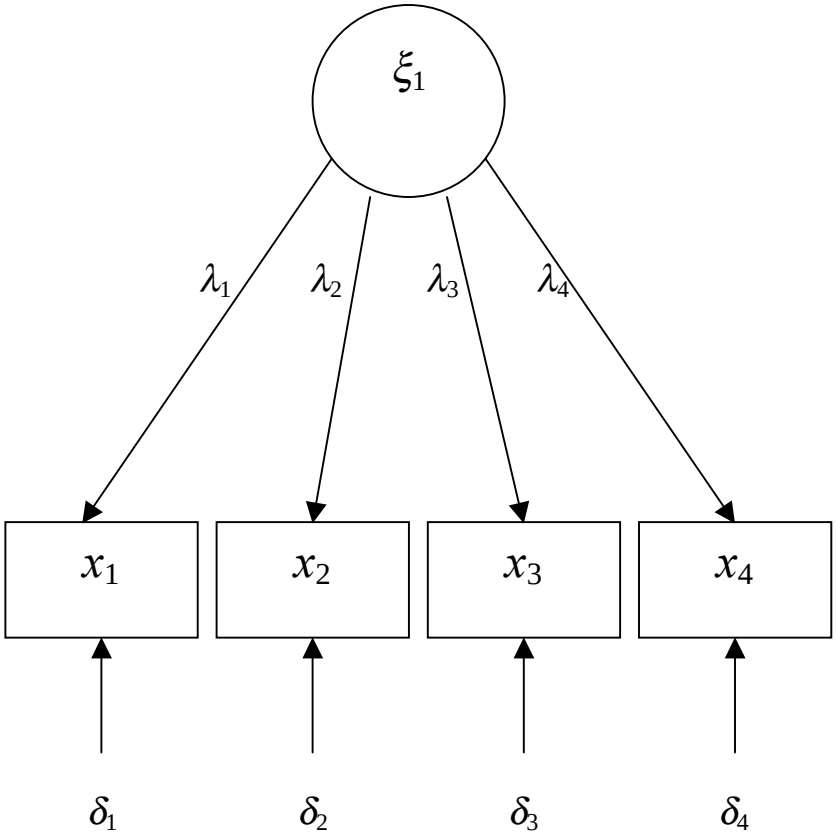
The cross-validation indices examined in the current study consisted of indices proposed in the literature in the past twenty years. Multiple group analysis has also been proposed as a cross-validation method in structural equation modeling. Hence, future research could examine the performance of multiple group analysis as a cross-validation technique, as well as at which point the constraints imposed in multiple group analysis become too restrictive.

General Conclusion

The use of cross-validation is important in terms of selecting a reliable model that is expected to fit the data available in other samples. This study indicates that there is no cross-validation index that performs perfectly well in every scenario. However, this study suggests that there is hope in terms of being fairly confident that a selection of cross-validation indices will perform optimally in certain conditions. It is hoped that this study provides further knowledge concerning the use of cross-validation indices under various realistic conditions.

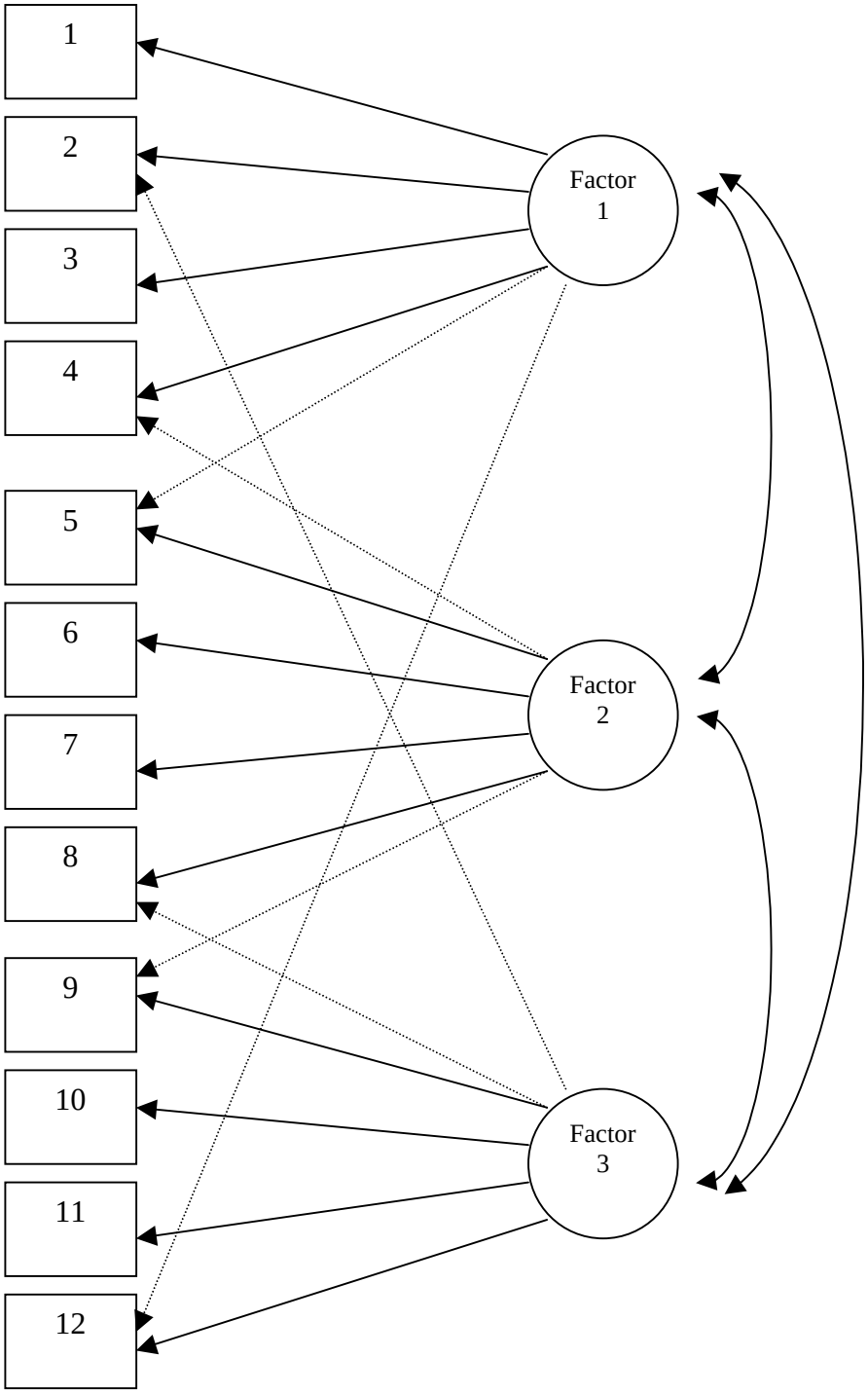
Appendix A

One-Factor Confirmatory Factor Analysis Model



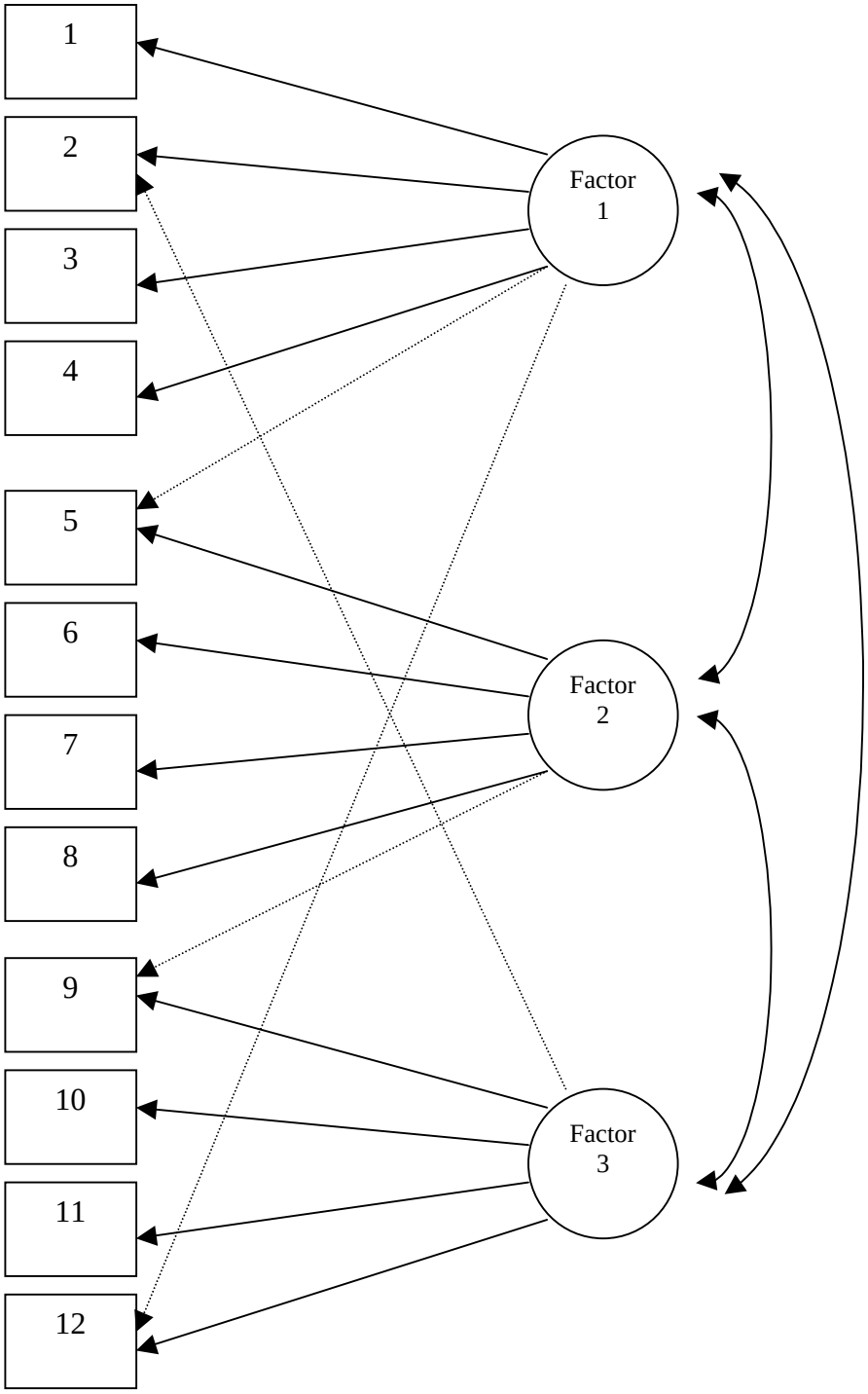
Appendix B

Model 1: Misspecified (Four Secondary Loadings Added to the Correct Model)



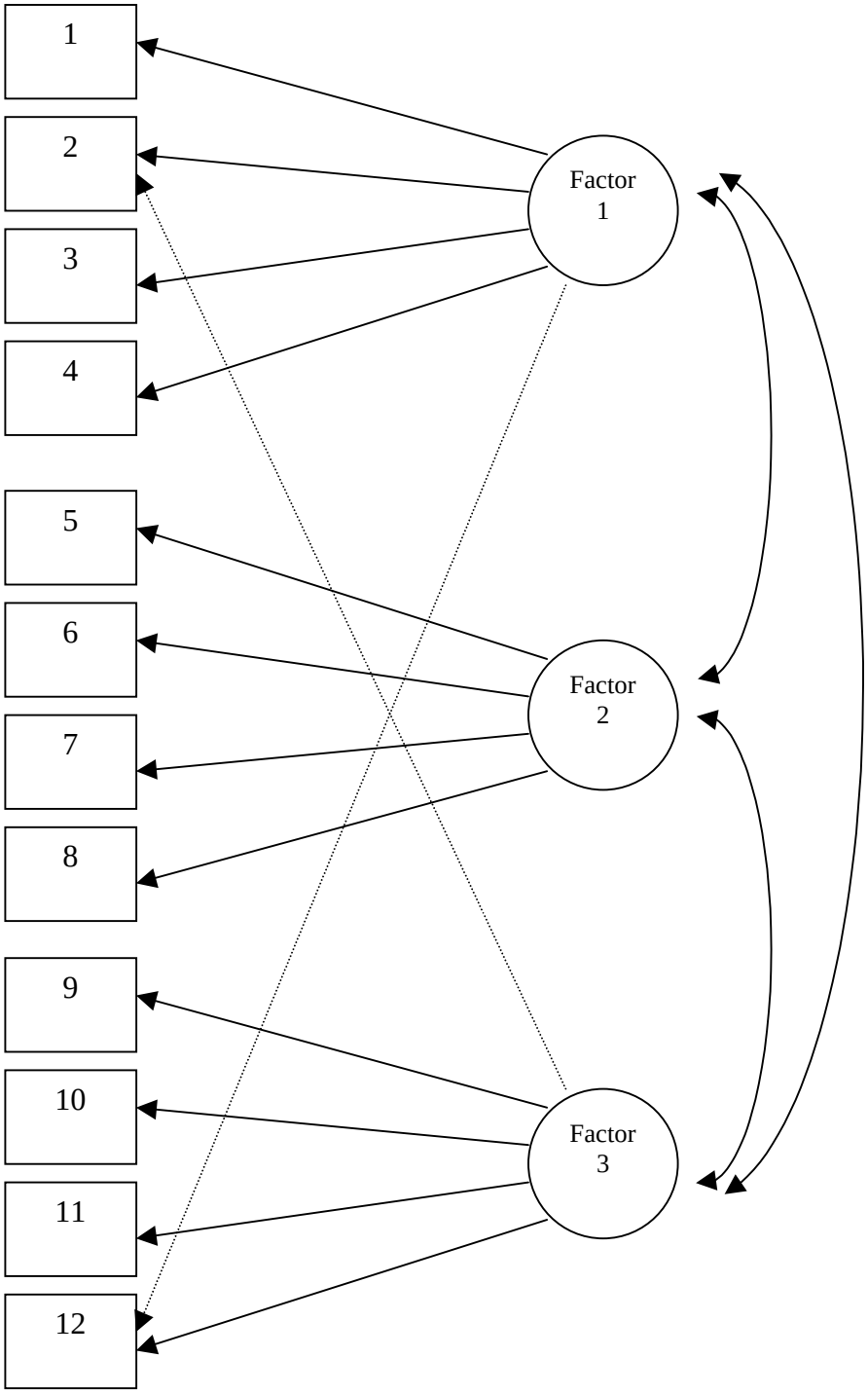
Appendix C

Model 2: Misspecified (Two Secondary Loadings Added to the Correct Model)



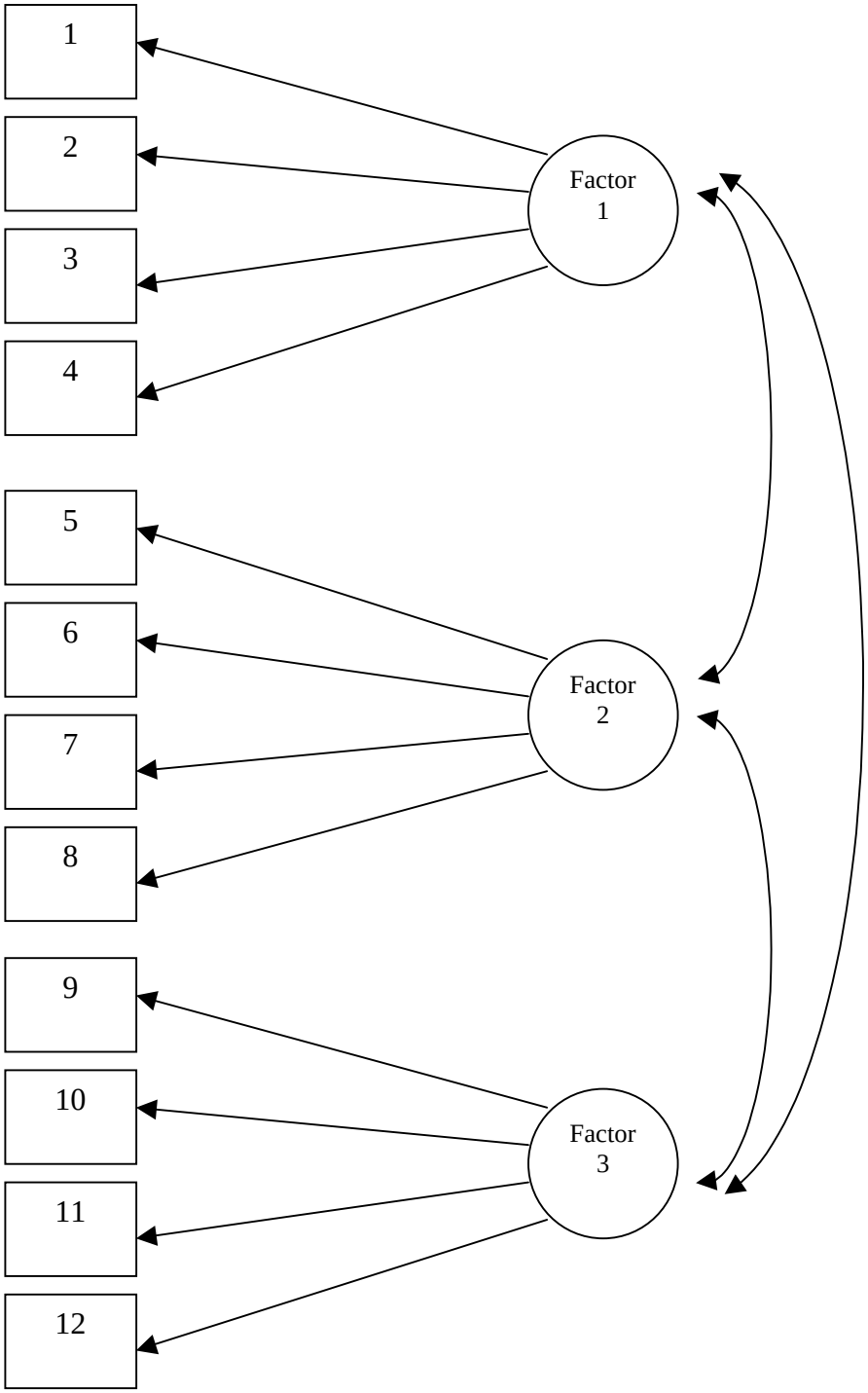
Appendix D

Model 3: Correct Factor Model (Two Variables with Secondary Loadings)



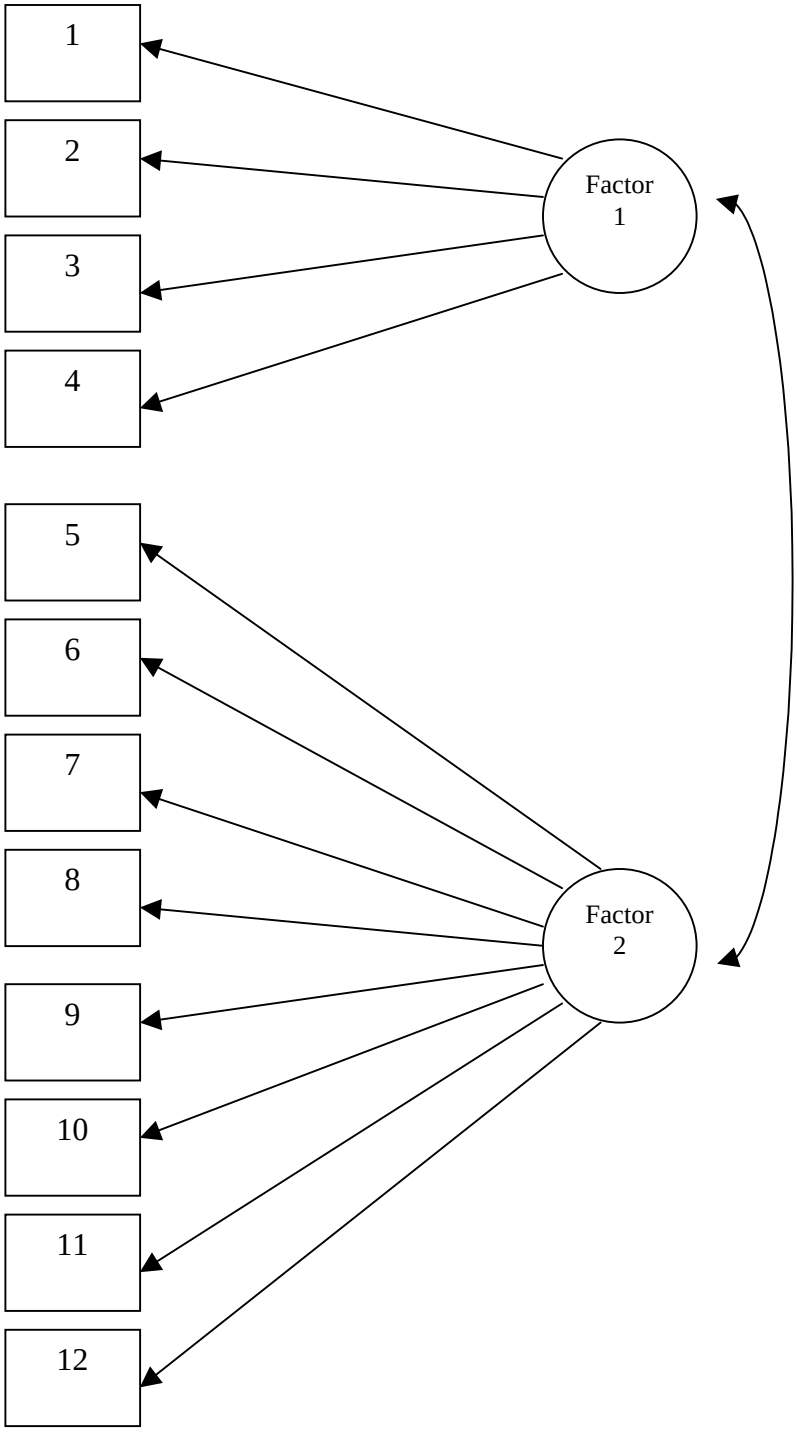
Appendix E

Model 4: Misspecified (Two Secondary Loadings Omitted from Correct Model)



Appendix F

Model 5: Misspecified (Correlation Between Factors 2 and 3 set to 1.0)



References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory.* Budapest: Akademiai Kiado.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics, 13*(3), 469-475.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics, 16*(1), 125-127.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173.

Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28*(3), 351-374.

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

Bentler, P. M. (1995). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Bollen, K. A. (1986). Sample size and Bentler and Bonnett's nonnormed fit index. *Psychometrika, 51,* 375-377.

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

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog and H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction.* (Part 1, pp. 149-173). Amsterdam: North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229-242.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107*(2), 260-273.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8,* 1-24.

Browne, M. W. (1975a). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology, 28,* 79-87.

Browne, M. W. (1975b). A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology, 28,* 112-120.

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

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*(4), 445-455.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*(2), 230-258.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows.* Thousand Oaks, CA: Sage Publications.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods and Research, 21*(1), 89-115.

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.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147-167.

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

De Gooijer, J. G., & Koopman, S. J. (1988). Cross-validation criteria and the analysis of covariance structures. In M. G. H. Jansen & W. H. van Schuur (Eds.), *The many faces of multivariate analysis: Vol. II. Proceedings of the SMABS-88 conference.* Groningen, The Netherlands: RION.

Efron, B. (1982). *The jacknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Eliason, S. R. (1993). *Maximum likelihood estimation.* Newberry Park, CA: Sage Publications.

Fleishman, A. I. (1978). A method for simulating nonnormal distributions. *Psychometrika, 43,* 521-532.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70*(350), 320-328.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research, 20,* 255-271.

Hamilton, L. C. (1992). *Regression with graphics.* Belmont, CA: Duxbury Press.

Henly, S. J. (1993). Robustness of some estimators for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 46,* 313-338.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research, 26*(3), 329-367.

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

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York, NY: The Guilford Press.

Loehlin, J. C. (1998). *Latent variable models* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics, 15*(4), 661-675.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519-530.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97-103.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247-255.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5-11.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

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

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

Nel, D. G. (1980). On matrix differentiation in statistics. *South African Statistical Journal, 14,* 137-193.

Neter, J., & Wasserman, W. (1974). *Applied linear statistical models.* Homewood, IL: Richard D. Irwin, Inc.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and Prediction* (2nd ed.). New York, NY: Holt, Rinehart and Winston.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*(387), 575-583.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye and C. C. Clogg (Eds.), *Latent variables analysis.* Thousand Oaks, CA: Sage Publications.

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics, 19*(4), 415-428.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465-471.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York, NY: John Wiley & Sons.

Weisberg, S. (1980). *Applied linear regression.* New York, NY: John Wiley & Sons.

VITA

Tiffany Ann Whittaker was born in San Antonio, Texas on September 10, 1971, the daughter of Harold Roger Whittaker and Sandra Sawyer Whittaker. After completing her work at Roosevelt High School, San Antonio, Texas, in 1989, she entered the University of Texas at San Antonio, majoring in Psychology and Criminal Justice. She received a Bachelor of Arts from the University of Texas at San Antonio in December 1995. In August 1996, she entered the Graduate School of the University of Texas at San Antonio, majoring in Psychology. She received a Master of Science from the University of Texas at San Antonio in May 1998. In August 1999, she entered the Graduate School of the University of Texas at Austin, majoring in Educational Psychology in the area of Quantitative Methods. Her publications include:

Whittaker, T. A., Fouladi, R. T., & Williams, N. J. (2002). Determining predictor importance in multiple regression under varied correlational and distributional conditions. *Journal of Modern Applied Statistical Methods, 1*(2), 354-366.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (in press). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*.

Permanent Address: 706 Stoneway, San Antonio, TX 78258

This dissertation was typed by the author.