2014

BOOTSTRAPPING WITH SMALL SAMPLES IN STRUCTURAL EQUATION MODELING: GOODNESS OF FIT AND CONFIDENCE INTERVALS

Craig Michael Krebsbach
University of Rhode Island, craigmk@my.uri.edu

BOOTSTRAPPING WITH SMALL SAMPLES IN STRUCTURAL EQUATION

MODELING: GOODNESS OF FIT AND CONFIDENCE INTERVALS

BY

CRAIG MICHAEL KREBSBACH

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2014

MASTER OF ARTS THESIS

OF

CRAIG MICHAEL KREBSBACH

APPROVED:

Thesis Committee:

Major Professor   Lisa L. Harlow

W. Grant Willis

Gavino Puggioni

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2014

ABSTRACT

Structural equation modeling (SEM) has become a regular staple of social science research, however very little is known about small sample size use. A sample size of 200 or larger for SEM models has been advocated (Boomsma, 1983; Kline, 2011) and the main test of model fit (χ^2 goodness-of-fit) is sample size dependent and performs optimally in a range of at least 200-400 (Kenny, 2012). Model complexity in SEM can vary, however, a simple model could hold potential benefits to a researcher without the ability to attain 200 observations. Thus research with models with less than 200 need to be considered more. Two manuscripts are presented, both stemming from a $3\,x\,3$ factorial simulation with varied sample sizes ($n = 50, 100, 200$), factor loadings ($\lambda = 0.60, 0.75, 0.90$), and bootstrap samples to the sample size n and a population sample of size $N = 400$. One study looks at SEM fit indices and independence from the χ^2 test as well as bootstrap extension potential. The second study analyzed the use and ease of bootstrap confidence intervals (CIs) for any of the fit indices used in tradition SEM publications, a much needed addition to the field.

DEDICATION

In loving memory of a great man who was so proud to see his son finally go back to college and get accepted to graduate school. It saddens me everyday to know he is gone and won't be able to see what I can accomplish, but I still aim to make him proud every day none the less. This thesis would never have been possible without the support of my dad as well as the rest of my family.

Dedicated to. . .

Kenneth L. Krebsbach

May 13, 1946 - May 9, 2012

PREFACE

Manuscript format is in use.

This is a manuscript based thesis submission for a masters degree in psychology at the University of Rhode Island. This thesis contains a total of two separate manuscripts both intended for separate submissions.

Manuscript 1

Title: SEM Fit Indices and Chi-Square Independence

Co-Author: Dr. Lisa L. Harlow

Intended Journal: *Structural Equation Modeling*

Status: Not yet submitted, but with the intention to submit by early 2014.
Manuscript is prepared for submission.

Manuscript 2

Title: Confidence Interval Estimation for SEM Fit Indices

Co-Author: Dr. Lisa L. Harlow

Intended Journal: *Psychological Methods*

Status: Not yet submitted, but with the intention to submit by early 2014.
Manuscript is prepared for submission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

MANUSCRIPT 1

SEM Fit Indices and Chi-Square Independence

Preface

Manuscript format is in use.

This is a manuscript based thesis submission for a masters degree in psychology
at the University of Rhode Island. This is the first manuscript of two for this
respective thesis.

Manuscript 1

Title: SEM Fit Indices and Chi-Square Independence

Co-Author: Dr. Lisa L. Harlow

Intended Journal: *Structural Equation Modeling*

Status: Not yet submitted, but with the intention to submit by early 2014.

Manuscript is prepared for submission.

Abstract

Structural equation model (SEM) fit indices have exponentially increased, but many indices are based on the sample-size-dependent χ^2 goodness-of-fit test and thus may have similar problems in assessing fit (Bollen, 1989; Kline, 2011). Debate over the use of SEM fit index usage has continued, but instead of a monochromatic approach, a synthesis of choosing several indices that behave well under many conditions and display a model-based lack of dependence on the χ^2 goodness-of-fit would be ideal. Use of bootstrapping has steadily become a common supplement to statistical parameter estimation as well as assisting potential small sample issues (Efron, 1979). Using a bootstrapping approach, the current research assesses models varying with small to moderate sample size $(50, 100, 200)$ and moderate to large factor loadings $(0.60, 0.75, 0.90)$ with the idea the small samples can compensate with large loadings (Guadagnoli & Velicer, 1988). Because of the highly correlated nature of fit indices, a partial least squares (PLS) regression was used to assess which indices showed the lowest association with the χ^2 goodness-of-fit under the multiple conditions. Ratios of the model fit between bootstrap replicates and the initial samples were also compared along with model non-conformity. The standardized root mean-square residual (SRMR) and McDonald's Centrality Index (MCI) may offer reasonable indices of fit as they were less related to the χ^2 goodness-of-fit test under the conditions presented and displayed the least variation across varied bootstrap replications.

SEM Fit Indices and Chi-Square Independence

The debate among structural equation modeling (SEM) practitioners is how to truly assess model fit. Since Jöreskog (1969) brought LISREL into the SEM forefront, the field has been burgeoning with fit indices. The catalog of indices has grown because of known sample size issues with the χ^2 goodness-of-fit test (Bentler & Bonett, 1980), which is the only true statistical test used to assess SEM model fit. Bentler (1985) continued the push forward with his SEM program EQS, which presented even more potential fit indices to help supplement the χ^2. An issue thus surfaces, however, that many of the fit indices regularly used in SEM are generally based on the χ^2 goodness-of-fit test which is exactly what the indices were intended to help remedy.

Past research has focused on specific indices and their relation to sample size, such as Tucker and Lewis (1973) who demonstrated that their Tucker-Lewis Index (TLI) was sample size independent. In a similar fashion, Bollen (1989) mathematically derived that his Incremental Fit Index (IFI) does not rely on sample size. However, a caveat is that traditionally SEM research involves large sample sizes, with the χ^2 behaving optimally for sample sizes of at least 200-400 (Boomsma, 1983; Kenny, 2012; Kline, 2011). Smaller sample sizes have become slightly taboo in the SEM literature, however the reality is that large sample sizes are not always possible. Other research has suggested that simplified designs can still be optimal with less than 200 observations, alluding to the fact that small samples may be possible with SEM when there are strong factor loadings (Guadagnoli & Velicer, 1988; Kline, 2011; Nevitt & Hancock, 2001).

Methodological researchers offer varying opinions about the choice of fit indices and the χ^2-test. Barrett (2007) inflexibly states that only the χ^2 should be

reported in SEM journals, however Markland (2007) rebuffs that by stating that fit indices have no "golden rules" and should be used as supplementary to the χ^2 index. Steiger (2007) also supports the notion that indices along with power and confidence intervals are necessary for a broader picture of the SEM model. What results is a mix of proponents and non-supporters of fit indices, and a catalog that is so large that any newcomer to SEM may become overwhelmed or simply confused (Beauducel & Wittmann, 2005; Gerbing & Anderson, 1992; Marsh, Balla, & McDonald 1988; Nevitt & Hancock, 2001).

Fit indices are known to be correlated quite often with sample size even though some have tried to mathematically deduce and analyze sample size independence (Bentler, 1990; Bollen, 1989; Ding, Velicer, & Harlow, 1995; Marsh et al., 1988). Strong correlations can still exist between indices and the χ^2 goodness-of-fit as well as amongst the other indices themselves. An analysis of actual model fit and their predictive qualities on the χ^2 goodness-of-fit are thus imperative to expand beyond just mathematical independence from sample size.

Several indices have garnered a good amount of use in publications; these include the root mean square of approximation (RMSEA), the standardized root mean-square residual (SRMR), the comparative fit index (CFI), the normed fit index (NFI), and McDonald's centrality index (MCI) (Bentler & Bonett, 1980; Kline, 2011; McDonald, 1989; Steiger & Lind, 1980). Most of the measures operate in a 0.00 to 1.00 range, with larger values ideal for the CFI, IFI, TLI, NFI, and MCI. Values closer to 0.00 are optimal for the RMSEA and SRMR indices. And finally, the χ^2 goodness-of-fit has a minimum value of 0.00 and should be similar to the degrees of freedom (df) for the model in a good fitting SEM analysis.

Whereas many SEM indices have been scrutinized for sample size independence, what could also be beneficial is a comparison of fit index independence

from the χ^2 goodness-of-fit test itself. By addressing sample size and χ^2 independence, we could perhaps begin to narrow down the list of fit indices to a select few that actually work well as a supplement to the χ^2 goodness-of-fit test. We should also keep in mind that extremely large sample sizes are rather uncommon in practice, so keeping perspective on rational sample sizes also seems worthwhile.

One option for small sample analysis is a process called bootstrapping, or sampling with replacement from an original data set to a desired sample size (Chernick, 1999; Efron, 1979). By taking an empirical sample of size n, we can randomly draw repeated (anywhere from 1,000 to 10,000 is common) samples with replacement to the same size n, with the goal that the repeated sampling will display a comprehensive picture of the sampling distribution (Efron & Tibshirani, 1986). Extensions can also be accomplished by bootstrapping, to not only draw random samples of size n but even larger sample sizes (i.e. $3n$, $5n$). In other research, Nevitt and Hancock (2001) have advised against using bootstrapping with data that have samples of size 100 or less, however offering the caveat that further research with simpler models could be possible and beneficial. Others have posited that strong factor loadings can amend a small sample size with simple structure (Guadagnoli & Velicer, 1988). A brief overview of bootstrapping is presented next to help illustrate this iterative procedure.

Bootstrapping Basics

Starting with a data matrix of observations such that n is the sample size with c columns each representing a recorded variable. Each row (r) of the matrix acts as a single observation. The $r \, x \, c$ matrix of parameters (p) is expected to be an adequate representation of the expected population parameters θ_p. The $r \, x \, c$ matrix for any p can be used in a bootstrapping framework such than each row of the matrix will be available for every replicate with replacement to create a

bootstrap based data matrix $\hat{\theta}_p$ of randomized rows from θ_p. Every row has a probability of $1/r$ chances of selection regardless of any previous selections. With thousands of bootstrap replicates (B), $\hat{\theta}_B$ will contain B samples of size n, with r randomly selected rows from θ_p with replacement. A rather thorough picture of the sampling distribution can thus be assumed for any of the parameters p in the B $\hat{\theta}_p$ matrices.

Efron (1979) notes that data from an unknown distribution F where $X_1, X_2, \ldots, X_n \sim F$ with replicates of $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ has an expected mean of

$$\bar{x} = \sum_1^n \frac{x_n}{n} \tag{1.1}$$

As pointed out by Efron (1979), $\bar{x}$ should approximate the expected mean of the sample $\mathrm{E}_F[X]$ because of the numerous replications representing a Taylor series expansion. The variance of the sampling distribution from the bootstrap replicates can thus be found by $\mathrm{E}_F[X^2] - \mathrm{E}_F[X]^2$. The standard error of the sampling distribution $(\hat{\sigma})$ can thus be found with

$$\hat{\sigma}(F) = \left(\frac{\mathrm{E}_F[X^2] - \mathrm{E}_F[X]^2}{n} \right)^{1/2} \tag{1.2}$$

Bootstrap estimates of the CI can be constructed via a non-parametric approach that finds the desired cutoff values of the empirically based probability mass function. Finding CI estimates from bootstrapping will be briefly described, shortly. Efron and Tibshirani (1986) state that a minimum of $B = 1000$ replicates are necessary for non-parametric CI estimation because of its complex nature. Teetor (2011) displays the use of the `quantile` function in R and the use of an even

larger number of replicates (9,999) to provide an empirical estimation of the CI.

In the case of SEM, the Bollen-Stine bootstrap (BSBS) transforms the data so that it is properly testing the χ^2 statistic (Bollen & Stine, 1992). Each $\hat{\theta}_p$ will represent a bootstrap sample from which a covariance matrix can be determined and the SEM model fit for every one of the thousands of replicates. The parameters of interest would typically be the B χ^2 values, however fit indices of each fitted model could also be estimated B times.

Modern computing power has made bootstrapping readily available, but the understanding of the process itself can perhaps seem daunting. Appendix A displays the structure of a typical bootstrap process, with the resulting $\hat{\theta}_B$ matrices allowing for a simulated mean or median value for starters, as well as a model based CI of any percentage. The `quantile` function in `R` can easily compute the cutoff points for a string of means or medians based on $\hat{\theta}_B$ at a given CI level.

As an example, a particular fit index of an SEM model, say the SRMR, could be estimated using $B = 2000$ replicates, or double the minimum B suggested previously (Efron & Tibshirani, 1986). The 90% CI could be found as such for the SRMR ($\hat{\theta}_{SRMR}$) which includes all of the model fit SRMR indices from 2000 replicates. To find a 90% CI, we would need to estimate the cutoff for the lower (0.05) and upper (0.95) proportion of scores, and the command in `R` with a vector of the 2000 SRMR estimates (`theta.srmr`) would look like:

```
quantile(theta.srmr, c(0.05, 0.95))
```

This function could be extended easily by adding a third quantile value (which is contained between the `c( )` in the command), that measures the median of the sample. The `quantile` function in `R` works similar to constructing a box-plot, but with the user having the ability to specify cutoff points for the CI. In this case, we could find the 90% CI and the median for $\hat{\theta}_{SRMR}$ as such:

```
quantile(theta.srmr, c(0.05, 0.50, 0.95))
```

A similar process can thus be done for any given fit index of a SEM model, including the χ^2. While tedious, a typical model would only need one set of bootstrap replicates conducted, thus bootstrapping is not as time consuming as one might perceive. Bootstrapping packages do exist in R and most commercial programs contain simulated bootstrap additions as well. Allowing the computer to store all of the $\hat{\theta}_p$ matrices and then compute the necessary information allows for a not so complicated bootstrap process that adds great depth to any researchers expertise. The resulting bootstrap process assumes that the large number of $\hat{\theta}_p$ replicates will adequately map the sampling distribution of θ_p. In doing so, difficult parameters (i.e., fit indices) can be estimated with CIs that are based on the data as opposed to complex mathematical processes.

Simple Bootstrapping Examples

A very simple bootstrap procedure could be done with something as simple as the roll of a die. Say we roll the same die repeatedly 20 times, the numbers would never change on the die, but the resulting values all have an equal chance of selection with replacement. Using the `quantile` function in R, we could find a basic 90% CI of the mean for a standard dice with recorded values stored in a variable we'll call `dice`.

```
> dice
[1] 6 3 3 5 3 2 2 5 1 3 1 1 3 5 1 4 5 3 3 4
> mean(dice)
[1] 3.15
> quantile(dice, c(0.05, 0.50, 0.95))
    5%    50%    95%
  1.00   3.00   5.05
```

While the sample of one die is generic, it gives the basic concept of what bootstrap replication accomplishes (i.e., a mean, median, CI). Now if we were to

roll five dice at a time, and take the mean of each roll 20 times, we'd be able to construct a CI around the mean of the five dice. Note that each roll can have a mix of any numbers, ranging anywhere from all different values to all the same.

```
> dice
[1] 4.4  3.2  3.4  3.2  2.8  3.4  2.8  4.2  2.8  2.8
    3.4  2.8  3.0  4.2  3.6  3.0  3.4  3.6  2.8  4.2
> mean(dice)
[1] 3.35
> quantile(dice, c(0.05, 0.50, 0.95))
   5%    50%    95%
 2.80   3.30   4.21
```

The reason for the difference in means for one die ($M = 3.15$) and five dice ($M = 3.35$) can be described by the varied nature of resampling. With one die, each replicate is only represented by a single die, and thus only 20 total values were included. With five dice, the replicates were the mean of five dice done 20 times, thus 100 total dice values comprised the 20 means. The median and CIs displayed the same pattern, however 20 replicates are certainly not enough and therefore results are slightly varied. Doing this same experiment over thousands of replicates would result in means, medians, and CIs all relatively similar since the compiled $\hat{\theta}_B$ will be a close approximation to θ_p. Any variable or variables in a data set can be done in this same simple bootstrap process (e.g., CFI, χ^2, RMSEA, etc). Instead of dice, one could let a computer randomly select r in each replicate so that $(\hat{\theta}_p)$ is equivalent in length to θ_p. Extending $\hat{\theta}_p$ by drawing a larger value of observations r than the data set is possible as well, but bootstrapping to the same size n as the sample data is usual.

Population Covariance Matrices

Matrix algebra can be used to construct a covariance matrix to generate data according to the varied parameters in a simulation context. Based on a two factor CFA model with three items each (Figure 1.1), a matrix would contain six manifest

variables, and thus the covariance matrix would be $6\,x\,6$, representing two factors with three measures each. To construct a population matrix for the CFA design in Figure 1.1, we need the use of several matrices: Σ represents a $6x6$ population matrix, Λ is a $6\,x\,2$ matrix of factor loadings, Φ is a $2\,x\,2$ matrix of the factor correlations, and Θ is a $6\,x\,6$ matrix of measurement errors which standardizes results in Σ for a diagonal of all ones. Using this, one can create a population sample of N to bootstrap samples of n from within.

SEM requires one item loading or variance per factor to be set to 1.00 before analysis, such that the constructed Λ matrices contain identical λ values within an acceptable range $(0.50, 1.00)$ and the minimum criteria of three items per factor to prevent model misfit was also met (Kline, 2011). Thus, the Λ loading matrix is comprised of:

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0.00 \\ \lambda_{21} & 0.00 \\ \lambda_{31} & 0.00 \\ 0.00 & \lambda_{42} \\ 0.00 & \lambda_{52} \\ 0.00 & \lambda_{62} \end{pmatrix} = \begin{pmatrix} 1.00 & 0.00 \\ \lambda_{21} & 0.00 \\ \lambda_{31} & 0.00 \\ 0.00 & 1.00 \\ 0.00 & \lambda_{52} \\ 0.00 & \lambda_{62} \end{pmatrix}$$

Based on the suggestion of Kline (2011), the factor correlation should be minimally 0.50 and therefore the factor correlation matrix Φ consists of:

10

$$\Phi = \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}$$

Finally, the error matrix Θ is constructed such that Σ ends up with all diagonals of 1.00. The computation necessary for the covariance matrix Σ needed for simulating the population based datasets of $N = 400$ is completed with:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \tag{1.3}$$

A unique aspect of bootstrapping within SEM is that because the χ^2 goodness-of-fit is testing against the null hypothesis, a readjustment needs to be made to the sample so that we are properly resampling the model fit within a χ^2 framework (Bollen & Stine, 1992; Chernick, 1999; Enders, 2001; Nevitt & Hancock 2001; Sharma & Kim, 2012). Thus a Bollen-Stine bootstrap (BSBS) requires a transformation to the data matrix so that the samples are reporting accurate simulated p-values (Figure 1.2). The resulting BSBS gives us a composite Bollen-Stine p-value (BSp) that is useful to help verify the model fit but also allows for examining fit index behavior under repeated samples. The transformation converts the covariance matrix used in the χ^2 goodness-of-fit test to accurately test the null by forming a new data set, Z, from the initial data when Y represents the centered raw data, S represents the sample covariance matrix of Y, and $\hat{\Sigma}$ is the implied covariance matrix such that:

$$Z = YS^{-1/2}\hat{\Sigma}^{1/2} \tag{1.4}$$

Bollen and Stine (1992) then mathematically deduce that the covariance matrix of Z being $Z'Z/(N-1)$ is now equal to the implied covariance matrix $\hat{\Sigma}$ and

boot-strapping from the transformed data in Z will accurately test the null hypothesis in all bootstrap samples.

Absolute Fit Indices

The absolute indices are based simply on the covariance matrix and are most notably represented by the χ^2 goodness-of-fit test and its comparison to a critical value for the given degrees of freedom (df). It is widely known that χ^2 is affected by sample size, however it is usually reported and is preferred to not be significant in SEM research (Barrett, 2007; Bentler & Bonett, 1980). Absolute fit indices in the current study will thus be represented by the χ^2 as well as another common absolute fit index, the SRMR. Suggested maximum values for the SRMR have generally been 0.08, but this value has been noted as being too liberal and that 0.06 might be better (Hu & Bentler, 1999; Kline, 2011).

Non-centrality Fit Indices

Non-centrality indices are those which measure distance between the covariance matrix and the matrix of the null model. The RMSEA (Steiger & Lind, 1980) has been reported very often in the literature, particularly because it assumes initially that the model will not have a perfect fit (Raykov, 1998). Beauducel and Wittmann (2005) state that similar to the SRMR, the RMSEA holds rather steady even with slightly non-normal data as is common in psychometrics. The use of RMSEA with a 90% CI has become common and has been recommended by several within SEM (Kenny, 2012; Steiger, 2007). Not unlike the RMSEA, the MCI (McDonald, 1989) has also been shown to lack dependence on sample size, and has performed well in simulations (Ding et al., 1995; Ding, 1996; Gerbing & Anderson 1992; Widaman & Thompson, 2003) and will be studied further in the current research. The CFI is a third non-centrality index that has shown some sample size independence and should be considered simply because of its common use in the

SEM field with values between 0.95 and 1.00 being preferred (Bentler, 1990; Ding, 1996).

Relative Fit Indices

The TLI has become one of the primary indices for relative fit since its introduction (Tucker & Lewis, 1973) and has performed well, but has shown some dependence on the interaction of sample size and factor loading (Kline, 2011; Sharma et al., 2005). The similarity of the TLI to the non-centrality based CFI has been noted with the advice that because of such high intercorrelation, only one should be reported and the TLI is sometimes preferred (Kenny, 2012). This research has opted to go with both the CFI (non-centrality) and TLI due to both their popularity and the need for further analysis of both. The IFI will be the second relative fit index tested particularly because of its reported independence from sample size (Bollen, 1989). The Normed Fit Index (NFI) (Bentler & Bonett, 1980) will also be included as a non-centrality index based in part to its popularity in use even with some advice against its use (Ding et al., 1995; Ding, 1996). While the IFI and TLI can minimally exceed values of 1.00, all three relative fit indices should be bounded between 0.00 and 1.00 with a preferred acceptable range between of 0.95 to 1.00 (Kline, 2011).

Thus, the goal of the current research is to address the dual-facetted question of χ^2 independence as well as the bootstrapping potential of select fit indices that have seemingly fared well in past research such as the MCI and IFI (Ding et al., 1995; Ding, 1996; Gerbing & Anderson, 1992; McDonald, 1989). Focus will be limited to small and moderate sample sizes of 50, 100, and 200 with moderate to good loadings of 0.60, 0.75, 0.90 in a simple two factor CFA. Comparison of bootstrapped samples can be compared to the initial sample as well as bootstrapped back to the population sample of $N = 400$ for direct comparison.

Further research has been needed for some time on the fit indices and their supplementation of the χ^2 for small n designs (Ding, 1996; Gerbing & Anderson, 1992; Nevitt & Hancock, 2001; Sharma et al., 2005; Tanaka, 1987). What is preferred are indices that can remain stable over varied model parameters and bootstrap replicates that can demonstrate a strong correlation to the original data set. Indices that function with minimal correlation to the χ^2 based on model fit would also be highly beneficial. Model non-convergence can also be an issue, particularly with samples below 100, and thus analysis of such will also be important (Boomsma, 1983; Guadagnoli & Velicer, 1988; Nevitt & Hancock, 2001).

Research Goals

First, based on past SEM simulation research, it is hypothesized that the MCI might be a prime candidate for use in small sample SEM designs (Gerbing & Anderson, 1992). Second, the TLI, CFI, and IFI tend to be strongly correlated (Kline, 2011) and thus it is also hypothesized that they will behave similarly across conditions and only one really needs to be reported. Third, it is also hypothesized that bootstrap replicates will highlight very inconsistent χ^2 values and that stable fit indices could help support the sample-size-influenced χ^2 test. Fourth, the MCI and SRMR will show strong bootstrap consistency between population (N) and samples (n) of varied conditions (Beauducel & Wittmann, 2005; McDonald, 1989). And finally, samples of $n = 50$ will likely see the largest amount of non-converging models, particularly with lower levels of λ (Guadagnoli & Velicer, 1988).

Methods

A simulation-based study was used to address the bootstrap based qualities of eight SEM fit indices with smaller sample sizes. A simple confirmatory factor analysis (CFA) model with two factors and three items each (Figure 1.1) was used based on the suggestions of past research (Nevitt & Hancock, 2001) with varied small samples sizes ($n = 50, 100, 200$) and moderate to strong factor loadings ($\lambda = 0.60, 0.75, 0.90$) which can help compensate for a small n (Guadagnoli & Velicer, 1988). The factor correlation (Φ) between the two factors was set at a modest 0.50, which is considered a lower bound for well-fitting models but could also help prevent over-powering the varied sizes of n and λ (Kline, 2011).

A population matrix based sample of $N = 400$ was generated for every condition and the samples of 50, 100, and 200 were drawn from N. Within the nine mixed conditions of n and λ (i.e., $n = 100$ & $\lambda = 0.90$; $n = 200$ & $\lambda = 0.60$), 30 total samples were drawn to increase the robustness of the study and to exceed the recommended five samples per population matrix (Guadagnoli & Velicer, 1988. Seeds were used to randomly draw the data in R (see Appendix B), however univariate and multivariate kurtosis were not checked to allow for varied types of model fit in each condition. The data was simulated on a 7-point integer based Likert scale, quite typical in social science research. Figure 1.3 depicts the simulation process used for this research.

All simulations were conducted using the `lavaan` package in R (R-Core-Team, 2013; Rosseel, 2012). A total of 270 simulated samples (i.e. 3 n's x 3 λ's x 30 seeds) were conducted with 2,000 bootstrap replicates taken in each scenario providing a total of 558,000 total models being fit. The initial fit for the population samples (N) as well as the samples drawn from the population (n) were stored as well as

the bootstrap replicated fit indices and CIs for the sample n bootstrapped to a size of n as well as to the population sample size for comparison ($N = 400$). R code for the simulation is presented in Appendix B.

Because of the high amount of correlation between most of the fit indices and the truncated nature of the zero to one boundaries, standard multiple regression would not be advised to assess findings. Thus, a partial least squares (PLS) method of regression will be used, which utilizes a principle components analysis approach and has been shown to be optimal with small data sets as well as with high intercorrelations (Sanchez, 2012; Tenenhaus et al., 2005). The goal is to use a PLS regression to find the weakest fit index predictors of the χ^2 value for each condition. By determining the indices that are the weakest predictors, one can presume that they are less associated with the χ^2 goodness-of-fit, however that does not mean there is no correlation between the fit index and the χ^2. Use of the `plsdepot` package in R was necessary for the PLS regression aspect to be employed (Sanchez, 2012).

Also recorded was the counts of model non-convergence for the three sizes of n and λ. The fit indices for the original population sample N was recorded as were the indices for the sample n randomly selected from N. A BSp was calculated and recorded for each of the 270 total samples bootstrapped, as were the quantile based 90% CI values for the bootstrapped models and a corresponding median value of the replicates.

<center>**Results**</center>

The correlation matrix of the indices to the χ^2 goodness-of-fit in the original population sample of N and sample of n as well as the bootstrap replicates of both sizes suggest varied results (see Table 1.1). The NFI has the lowest correlations with χ^2 regardless of sample size or whether bootstrapping was used or not. The TLI and CFI had equal correlations to χ^2 for the population samples ($r = -0.747$) and samples of n ($r = -0.562$). The IFI had correlations for population values and samples of n ($r = -0.751$, $r = -0.614$), almost similar to the TLI and CFI. The RMSEA and MCI both displayed the strongest correlations with χ^2 for N, n, and bootstrap values back to n and N.

Use of PLS regression for the 270 samples suggests that the TLI and CFI are the weakest predictor of χ^2 within the $N = 400$ population samples ($\beta = 0.002$) followed closely by the IFI, again supporting strong correlation between the three indices (see Table 1.2). The SRMR was the least significant indicator of χ^2 for the samples of size n ($\beta = 0.008$) followed by the correlated CFI and TLI ($\beta = -0.025$) having identical values. The IFI is the least associated with χ^2 for the bootstrap replicate models of size n ($\beta = -0.006$) whereas the RMSEA ($\beta = 0.485$) and MCI ($\beta = -0.405$) both display very large standardized weights with χ^2. In the replicates of size $N = 400$, the SRMR shows a very weak association with χ^2 ($\beta = -0.026$). Compiling all four of these conditions, a PLS regression was run for the indices with the TLI and CFI being the weakest in association to the χ^2 ($r = -0.009$). The SRMR also show a weak association in the overall PLS comparison to the χ^2-test ($r = -0.053$).

Correlations between bootstrapped (BS) models and the original samples of N and n for all the fit indices (see Table 1.3) posits that the NFI ($r = 0.724$),

<center>17</center>

and SRMR ($r = 0.664$) are strongly correlated between their values for N and the BS N replicates. The same pattern follows for the correlation of samples n and the BS values of size n respectively ($r = 0.834$, $r = 0.710$). The NFI ($r = 0.545$) and SRMR ($r = 0.443$) have the highest correlation between population samples N and samples n drawn from N. The use of Cronbach's alpha (α) to address inter-correlation between the values of N, n, BS to $n = 400$, and BS to $n = n$ suggests that the NFI ($\alpha = 0.759$) and SRMR ($\alpha = 0.729$) have the strongest item correlation as seen in Table 1.3. The RMSEA had the weakest overall inter-correlation of the varied conditions in Table 1.3 ($\alpha = 0.090$).

Table 1.4 displays the means of the ratios of index values in $N = 400$ over the bootstrap replicates back to size $n = 400$ from the various sample sizes of n. The MCI had the mean ratio closest to 1 for the population based ratios in Table 1.3 ($M = 0.999$, $SD = 0.005$). The MCI also had the ratios closest to 1 in the sample size n of 50, 100, or 200 bootstrapped back to the same size of n ($M = 0.995$, $SD = 0.023$). The TLI, IFI, NFI, and CFI also preformed rather consistently with correlations between 0.99 and 1.00 as well, again a reflection of their covariation. The χ^2 ($M = 1.212$, $SD = 0.611$) and RMSEA ($M = 1.623$, $SD = 1.674$) both had rather large mean ratios of population samples to bootstrap of $n = 400$. For the sample n to bootstrap n ratios, the χ^2 ($M = 1.262$, $SD = 0.635$) and RMSEA ($M = 1.683$, $SD = 1.645$) both preformed poorly.

Non-conformity within the $n = 50$ bootstrap replicates was influenced by λ, with loadings of 0.60 ($n = 8879$), 0.75 ($n = 847$), and 0.90 ($n = 2$) (see Table 1.5). The replicates to size $n = 400$ for $n = 50$ only resulted with non-conforming models for $\lambda = 0.60$ ($n = 1970$), while $\lambda = 0.75$ and 0.90 had no issues with conformity. Samples of 100 ($n = 2752$) and 200 ($n = 107$) for the $\lambda = 0.60$ subgroup in the

replicates to size n as well as for 100 ($n = 1309$) and 200 ($n = 5$) in the replicates to $n = 400$ had varied conformity results. Table 1.5 also shows the replicates in the $\lambda = 0.75$ and $\lambda = 0.90$ conditions all had 0 non-conforming models for BS replicates from n back to the size of 400.

Figure 1.4 shows an interaction effect between the Bollen-Stine p-values (BSp) between the varied levels of n and λ. When $\lambda = 0.90$, the BSp holds relatively stable as sample size increases. Conversely, the means for the BSp drop drastically for $\lambda = 0.60$ and $\lambda = 0.75$.

Interactions are evident for the χ^2 goodness-of-fit for n and λ for the initial samples drawn from N (Figure 1.5). An interaction is not evident for BS replicates of χ^2 with size n (Figure 1.6), but an interaction is observed for the BS replicates of size n (Figure 1.7). The RMSEA has an interaction effect for the samples of size n (Figure 1.8), but not for the bootstrap replicates of size n (Figure 1.9). There is an interaction again for the RMSEA in Figure 1.10 where n is BS replicated to size N. The MCI has an interaction of n and λ for samples of n (Figure 1.11), but not for the BS replicates of size n (Figure 1.12). Figure 1.13 shows an interaction for n and λ of the MCI in replicates to size $N = 400$. The interaction effects for these indices suggest a joint dependency for each index in relation to sample size and loading when bootstrap sample sizes are extended to 400 as well as the initial samples. The varied values of n are each effected by λ as well.

Discussion

The correlational analysis of the fit indices from this research portray the RMSEA as highly correlated with the χ^2-test and also has the strongest association with the χ^2 in a PLS form as well. The RMSEA may thus be okay for use as a supplement, but its strong association with the χ^2 for model fit in the context of small samples might suggest it as not a primary index for use with SEM model fit. According to the PLS analysis in this research (see Table 1.2), the IFI seems to show the weakest correlation with χ^2 while the SRMR excelled in the BS replicates to the population size of $N = 400$. The NFI appeared weakly correlated to the χ^2, particularly for the initial population fits, however its advice against use should be kept in mind as well (Ding et al., 1995; Ding, 1996).

The NFI and SRMR also showed strong inter-correlation between their initial values and the BS replicate values. The RMSEA on the hand had very little inter-correlation within its values and again supports the notion that the RMSEA might not be an optimal primary fit index because of its high correlation with χ^2 and potential for interaction effects. Because of these findings, the RMSEA could be advised against use in bootstrap extensions within SEM because of its poor inter-correlation with samples and bootstrap replicates.

The CFI, IFI, and TLI acted in similar fashions for their ratios of initial population and sample fit measures compared to their bootstrap replicated fit measures regardless of sample size. Because of the similar correlations and ratios, it may only be necessary to select only one of the three indices since they have negligible differences. The χ^2 and RMSEA not surprisingly had varied and rather large overall ratios of non-bootstrapped initial fit measures to BS fit measures for both sizes of n and N. The MCI however seemed to hold a ratio closest to one

for its initial fit in comparison to its BS replicate fits and could be an ideal stable index for use in small n models and bootstrapping in general.

The issues for non-conformity for sizes of n and N were quite evident when $\lambda = 0.60$, but does improve as n increases. The combination of $\lambda = 0.60$ and $n = 50$ seems to be very inconsistent with regards to conforming models and would thus be ill-advised for bootstrapping use. Extending the bootstrap size to N did seem to alleviate some of the non-conformity issues, particularly for $\lambda = 0.75$. The overall non-conforming models for $\lambda = 0.60$ did decrease considerably with the larger bootstrap size, so this option could potentially assist a very small n in assessing its model fit. The BSp values did not show a difference with regards to λ, but did for n as would be expected with the idea that χ^2 is influenced by n. The interaction of n and λ for the BSp (see Figure 1.3) does allude to the BSp being much smoother for $\lambda = 0.90$ versus the other λ's. Again, larger loadings could help with consistency when n is small and even for bootstrap purposes.

The hypothesis that the MCI could be an optimal index for small samples was met as the MCI displayed a very consistent ratio of initial and bootstrap values across sample sizes. The hypothesis that the MCI and SRMR would be optimally consistent for small n models was only partly accepted in that the MCI didn't have a very strong correlation between n and N replicates, however the SRMR did fair well. The hypotheses that the χ^2-fit would be sporadic with the varied conditions and that the samples of $n = 50$ would have the most non-conforming models were both fully accepted. And finally, the TLI, CFI, and IFI did prove to be rather correlated as expected, with the IFI potentially being slightly better with regards to χ^2 independence.

Simulation studies always have the limitation of non-real data being used. While the samples used in this simulation were based strictly on 1-7 Likert scale

interval based results, it was still artificial. The samples used were not specifically checked for univariate and multivariate skewness and kurtosis as is advised, simply to allow the simulated data to be varied more than just selecting only very good fitting models. Empirical data should ideally meet univariate normality for skewness and kurtosis as well as multivariate measures of normality such as that provided by Mardia's test for multivariate normality to allow for proper estimation by maximum likelihood (Gao, Mokhtarian, & Johnston, 2008; Harlow, 1985; Kline, 2011; Mardia, 1970).

Future research could certainly aim at a bootstrap approach to larger samples as well. Fit indices could certainly vary with larger n designs, especially since many are based on the χ^2 which is influenced by n. The use of real empirical data sets with this bootstrap approach to small n designs could certainly help add credence to the reliability of the BSBS for SEM models. More research with varied factor loadings could certainly display some different patterns as well, and thus a larger set of conditions could help to delineate the difference between fit indices and λ.

What seems to be reasonable for fit indices is to act as a supplement to the χ^2 goodness-of-fit. Instead of focusing on whether fit indices should even be used (Barrett, 2007; Markland, 2007; Steiger, 2007), maybe the approach should be taken of fit indices being used as what they were intended for, a supportive case along with the χ^2. Fit index independence based on sample size alone is an issue of its own, but when determining what fit indices are optimal for reporting in SEM literature, looking at the correlations among the fit indices and particularly with the χ^2 may be more important. Including two or three fit indices that function differently from one another and yet supplement the χ^2 goodness-of-fit regardless of sample size could perhaps be the most beneficial use of indices in SEM.

Keeping in mind that indices are not true statistical tests and that there

really are no set cutoff limits, it becomes the researchers' role to provide specific interpretation relative to their study. However, cherry-picking indices that support ones model the best within a single study is quite counterintuitive to the ideal of using a fit index in the first place and shouldnt have a place in SEM research. Therefore, a primary goal for fit indices is to thin out the massive catalog of them and find the several that are least correlated with one another and with the χ^2 goodness-of-fit test. The current research suggests the CFI, TLI, and IFI as potentials, but again their high intercorrelation should lead researchers to utilize only one of the indices. Highly correlated index measures such as the RMSEA may not provide enough non-redundant information as they are mimicking the χ^2 goodness-of-fit too closely. Thus, more research on this index is also needed.

Further work is needed in the area of fit index correlation and χ^2 goodness-of-fit independence beyond the initial presentation in this research to make solid claims on which indices are truly optimal. Model specifications and complexity will ultimately weigh a huge burden on optimal fit index usage, so continued research is needed. However, within a small n framework, it appears that the IFI and SRMR might provide good indices because of their lack of χ^2 association and are most optimal in terms of their relatively low correlation with chi-square. The MCI also faired rather well in this small sample bootstrap framework, particularly in keeping a very consistent ratio of actual fit to bootstrap fit ratios. The SRMR and MCI could both be strongly advised for small n models as well as when bootstrapping a model becomes a reasonable option.

The use of the χ^2 goodness-of-fit is usually presented for covariance modeling, but the fit indices used need further clarification. This research seems to suggest that the χ^2 mixed with the SRMR and MCI as well as either the IFI, TLI, or CFI would be most optimal. The BSp should also be included for any SEM research,

especially when the BSBS is a readily available option. More work is needed in the area of using bootstrap replication to expand a small sample size issue in SEM, however there seems to be some potential strengths within some indices such as the MCI, SRMR, and IFI. While BS sampling can never replace real data, it can also help to assess the sampling distribution used and could posit an opportunity for research where samples of at least 200 might be hard to achieve.

Bibliography

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in cfa based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41–75.

Bentler, P. M. (1985). Theory and implementation of EQS, a structural equations program. Los Angeles, CA: BMDP Statistical Software.

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

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in covariance structures. *Psychological Bulletin, 88*, 588–606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*(3), 303–316.

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

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Phd thesis, University of Groningen.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide.* John Wiley and Sons Inc: New York.

Ding, L. (1996). *Evaluation of structural modeling fit indices.* PhD thesis, University of Rhode Island, Kingston, RI.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2*(2), 119–143.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard error, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54–77.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352–370.

Gao, S., Mokhtarian, P. L., & Johnston, R. A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record: Journal of the Transportation Research Board, 2082*, 116–124.

Gerbing, D. W., & Anderson, J. C. (1992). Monte carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research, 21*, 132–160.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275.

Harlow, L. L. (1985). *Behavior of some elliptical theory estimators with non-normality data in a covariance structures framework: A Monte Carlo study.* Phd dissertation, University of California, Los Angeles.

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.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202.

Kenny, D. A. (2012). Measuring model fit. Retrieved electronically October 22, 2012 from http://www.davidakenny.net/cm/fit.htm.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* The Guilford Press: New York. 3rd edition.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519–530.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on paul barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences, 42*, 851–858.

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.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*(3), 353–377.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences, 24*(2), 291–293.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Sanchez, G. (2012). R package plsdepot: PLS regression 1. Retrieved electronically May 5, 2013 from http://gastonsanchez.com/plsdepot.

Sharma, P. N., & Kim, K. H. (2012). *A comparison of PLS and ML bootstrapping techniques in SEM: A Monte Carlo study.* Presentation at the 7th International Conference on Partial Least Squares and Related Methods, Houston, TX. Retrieved electronically on October 1, 2012 from http://www.pitt.edu/

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for model fit in covariance structure models. *Journal of Business Research, 58*, 935–943.

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. IN the annual meeting of the Psychometric Society, Iowa City, IA (Vol. 758).

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*, 893–898.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit

in structural equation models with latent variables. *Child Development, 58*(1), 134–146.

Teetor, P. (2011). *R cookbook: Proven recipes for data analysis, statistics, and graphics.* O'Reilly Publishing: Beijing. 1st Edition.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis, 48*(1), 159–205.

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

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*(1), 16–37.

Table 1.1. Correlation matrix of fit indices in comparison to the χ^2 goodness-of-fit test within the population samples ($N = 400$), sample size n from the $N = 400$ sample, the bootstrap replicates of size n, and the bootstrap replicates of size $n = 400$.

Index	$N = 400$	$n = n$	BS to n	BS to N
SRMR	0.815	0.454	0.237	0.068
TLI	-0.747	-0.562	-0.225	-0.204
IFI	-0.751	-0.614	-0.276	-0.208
NFI	-0.620	-0.287	-0.048	-0.018
RMSEA	0.951	0.868	0.765	0.998
CFI	-0.747	-0.562	-0.232	-0.204
MCI	-0.937	-0.778	-0.692	-0.984

Note: BS = bootstrap

Table 1.2. Partial least squares regression based β coefficients for fit index association to the χ^2 within population samples ($N = 400$), samples drawn from N of size n, bootstrap replicates of the CFA model to size n, replicates to size N, and all four conditions combined into one PLS regression. Two principal components used for each PLS regression.

	$N = 400$	$n = n$	BS to n	BS to N	All 4
SRMR	0.23138	0.00828	-0.05056	-0.02631	-0.05256
TLI	0.00216	-0.02470	0.03303	-0.03007	-0.00910
IFI	-0.00285	-0.08341	-0.00566	-0.03227	-0.07703
NFI	0.09660	0.19337	0.19571	0.06230	0.23659
RMSEA	0.43022	0.49019	0.48500	0.49482	0.57359
CFI	0.00216	-0.02470	0.02663	-0.03007	-0.00997
MCI	-0.42763	-0.34486	-0.40518	-0.48563	-0.29438

Table 1.3. Correlations between the population samples of $N = 400$, the samples of size n drawn from N, and the bootstrap (BS) replicated models of size n and $n = 400$. Cronbach's α is also included for each index between the samples of $N = 400$ and $n = n$ as well as the bootstrap replicates of $n = n$ and $n = 400$.

Index	$r(N,\text{BS } N)$	$r(n, \text{BS } n)$	$r(n, N)$	Cronbach α
χ^2	-0.12610	-0.12606	0.13372	0.10086
SRMR	0.66393	0.70990	0.44345	0.72886
TLI	0.36846	0.26615	0.18224	0.31488
IFI	0.37071	0.23409	0.18105	0.30665
NFI	0.72437	0.83407	0.54498	0.75893
RMSEA	-0.11996	0.04744	0.09630	0.08954
CFI	0.36846	0.25156	0.18224	0.30726
MCI	-0.11913	0.15791	0.02305	0.12302

Table 1.4. Means and SD for the ratios of population sample size $N = 400$ to its corresponding bootstrap (BS) values of $n = 400$ as well as means and SD for ratios of samples of size n to the bootstrap (BS) replicates of size n.

| | $N = 400$ / BS $n = 400$ | | $n = n$ / BS $n = n$ | |
Index	Mean	SD	Mean	SD
χ^2	1.21179	0.61063	1.26199	0.63477
SRMR	1.09297	0.32703	1.14169	0.35451
TLI	0.99446	0.01703	0.98233	0.09379
IFI	0.99712	0.00880	0.99139	0.03997
NFI	0.99532	0.01265	0.98270	0.04737
RMSEA	1.62328	1.67355	1.68310	1.64499
CFI	0.99706	0.00904	0.99057	0.04801
MCI	0.99860	0.00509	0.99499	0.02339

Table 1.5. Means and SD for the ratios of population sample size $N = 400$ to its corresponding bootstrap (BS) values of $n = 400$ as well as means and SD for ratios of samples of size n to the bootstrap (BS) replicates of size n.

Condition	$n = 50$	$n = 100$	$n = 200$	Total
BS to $n = n$				
$\lambda = 0.60$	8879	2752	107	11,738
$\lambda = 0.75$	847	10	0	857
$\lambda = 0.90$	2	0	0	2
BS to $n = 400$				
$\lambda = 0.60$	1970	1309	5	3284
$\lambda = 0.75$	0	0	0	0
$\lambda = 0.90$	0	0	0	0
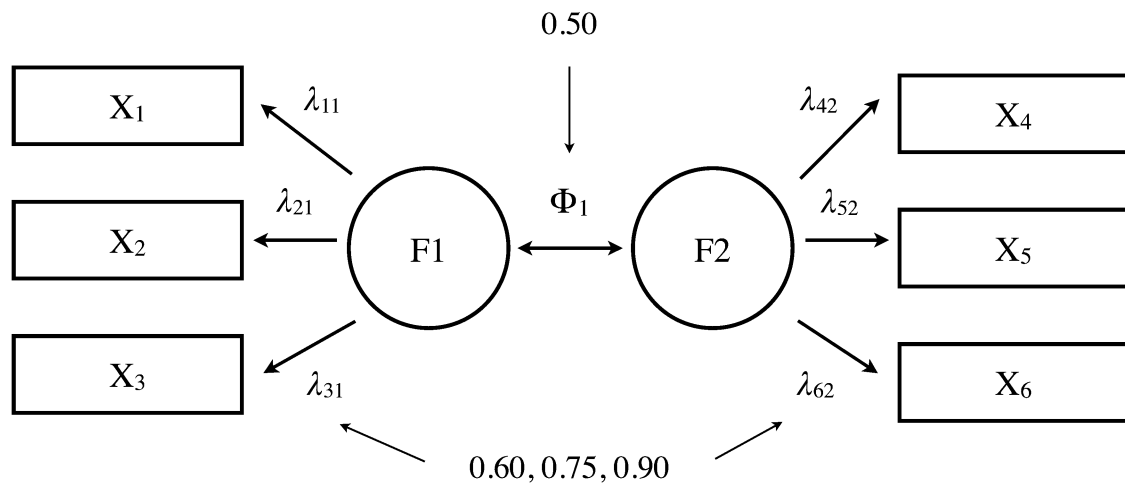
Figure 1.1. Two factor model utilized for the presented simulation study. Factor loadings (λ) are varied between 0.60, 0.75, and 0.90 where as the correlation between the two factors (F1 and F2) is set to 0.50.
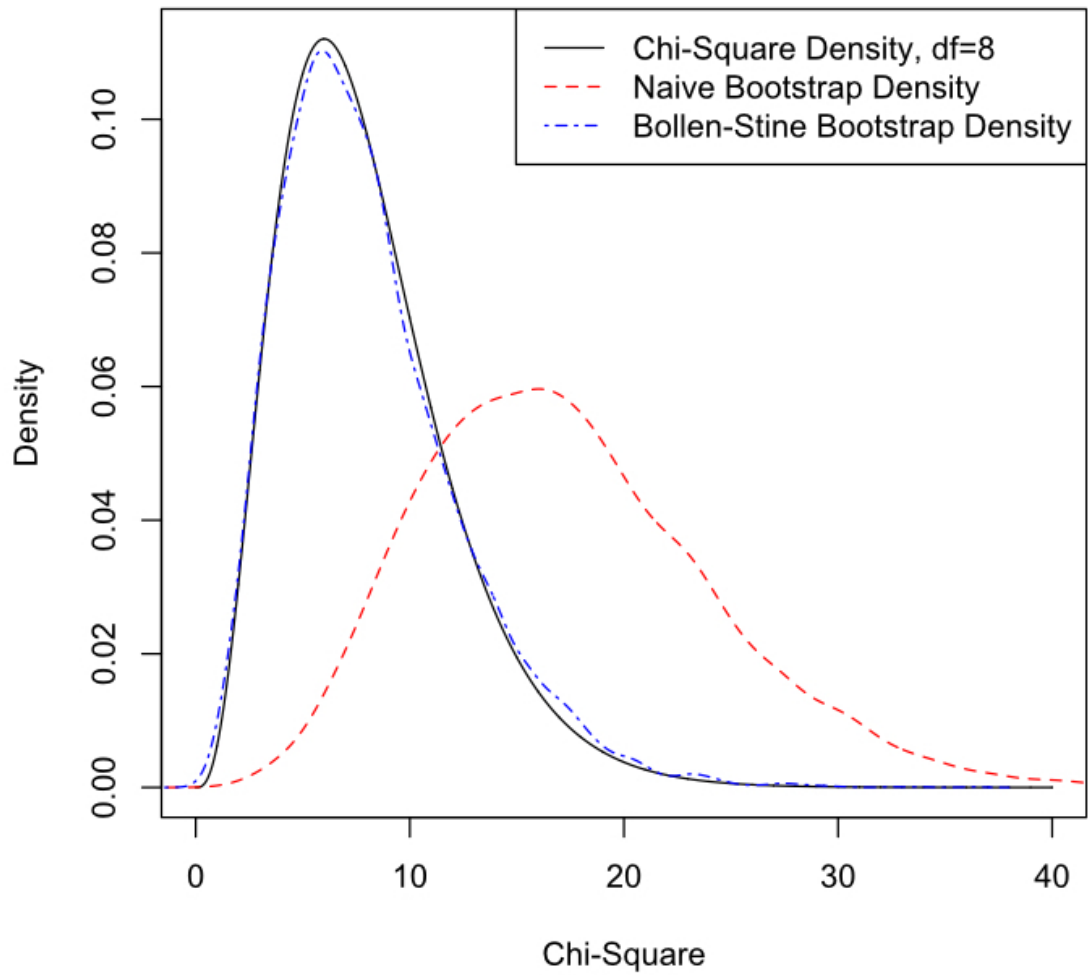
Figure 1.2. Density of the χ^2 distribution with $df = 8$ as well as the standard bootstrap and the Bollen-Stine bootstrap. The Bollen-Stine transformation is a necessary step in χ^2 resampling.
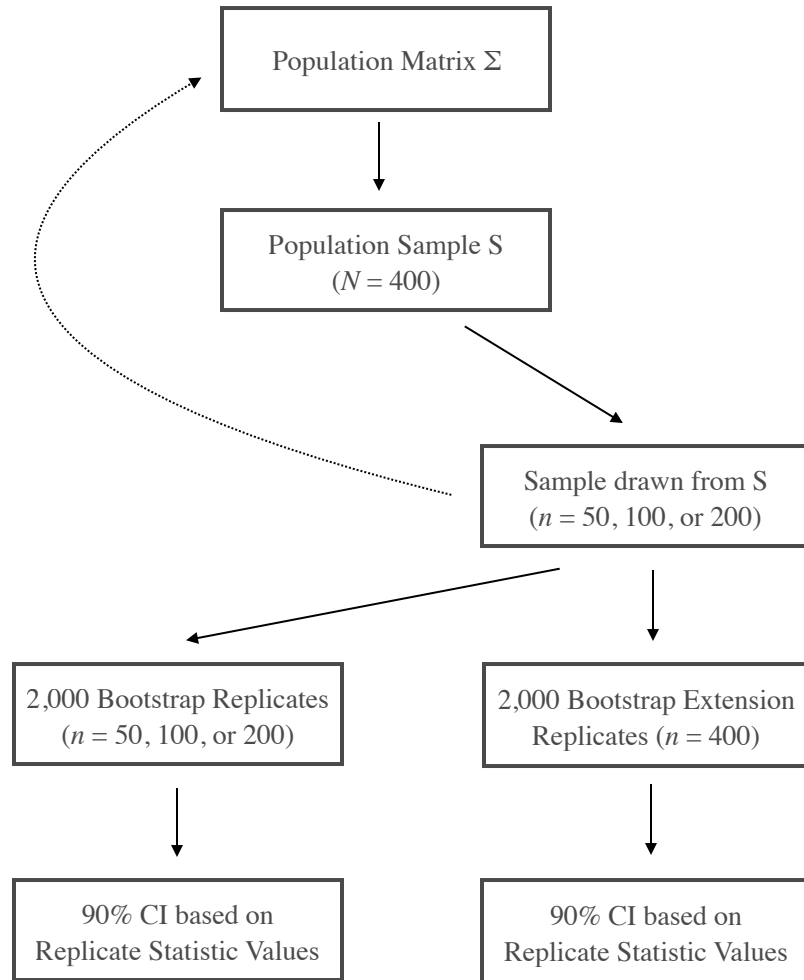
Figure 1.3. Flowchart of the simulation based design used.
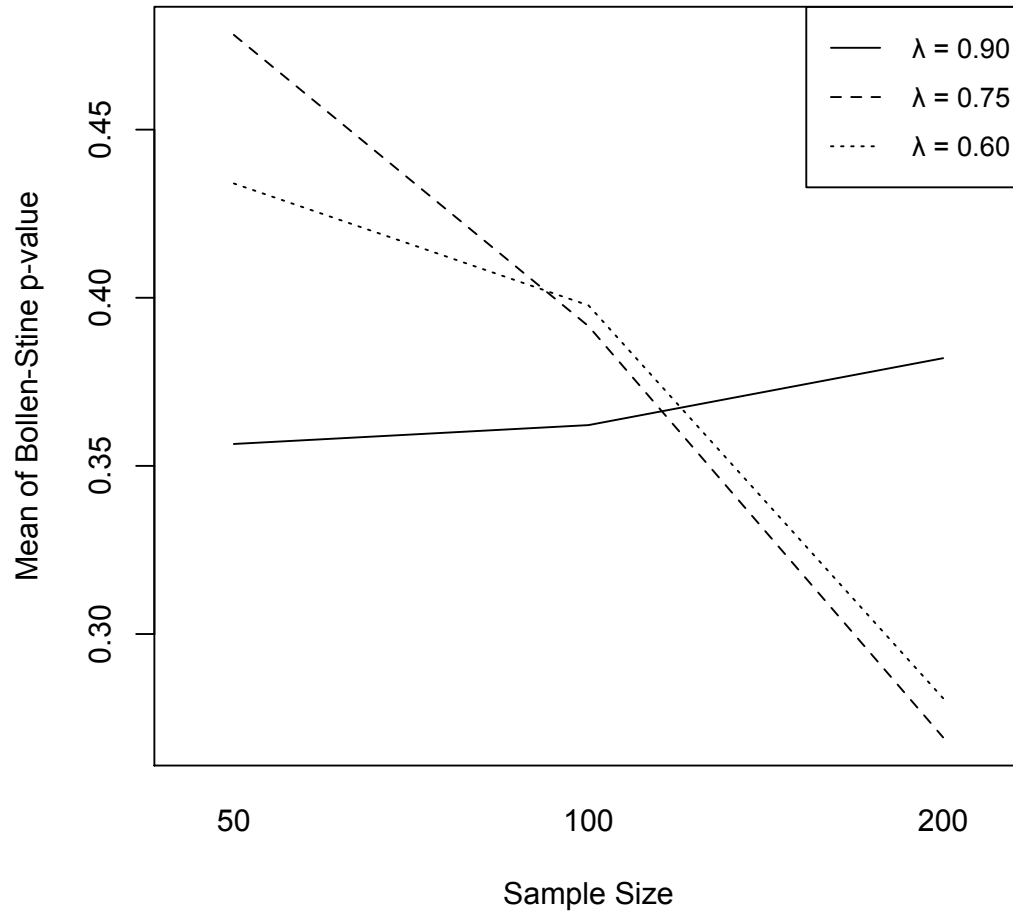
Figure 1.4. Interaction plot of Bollen-Stine p-values (BSp) for the initial sample sizes of $n = 50, 100, 200$.

Figure 1.5. Interaction plot of χ^2 values for the initial sample sizes (n).
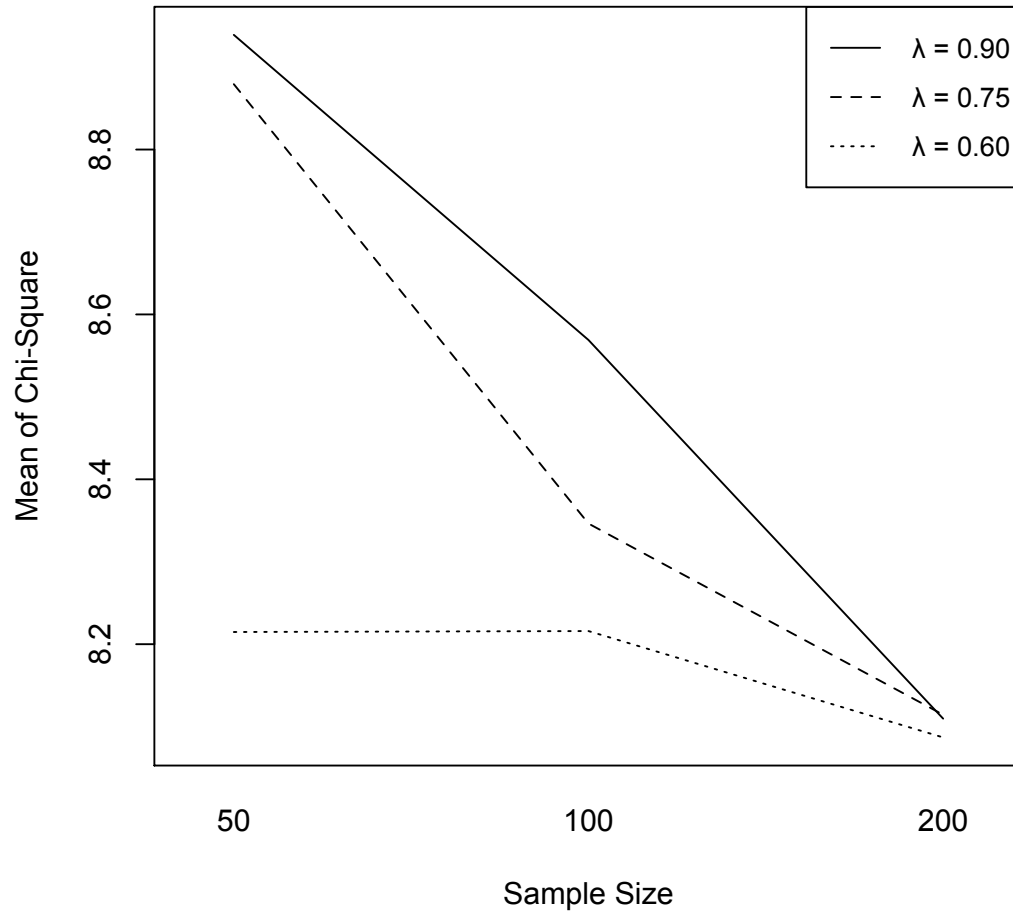
Figure 1.6. Interaction plot of the χ^2 values for the samples sizes of n bootstrapped to the same size as n.
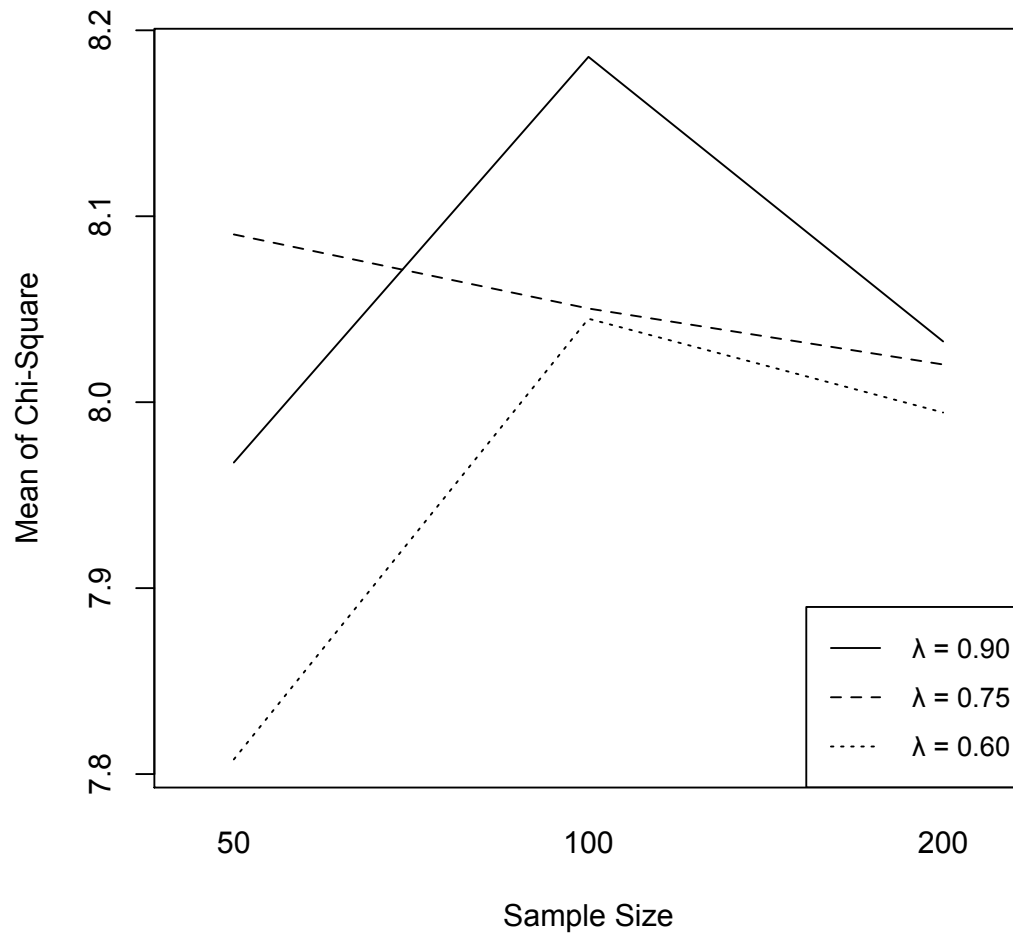
Figure 1.7. Interaction plot of χ^2 values for the bootstrap replicate extensions of the sample sizes n to the population sample size ($N = 400$).
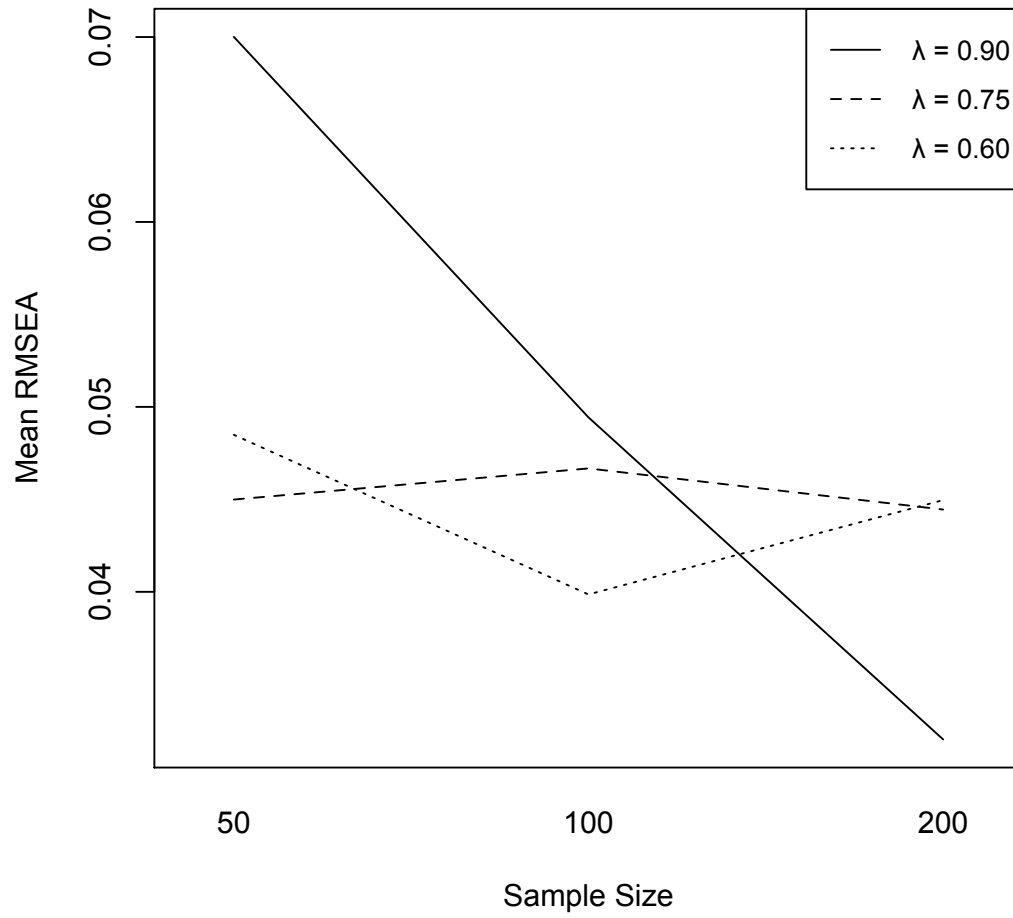
Figure 1.8. Interaction plot of RMSEA values for the initial sample sizes (n).
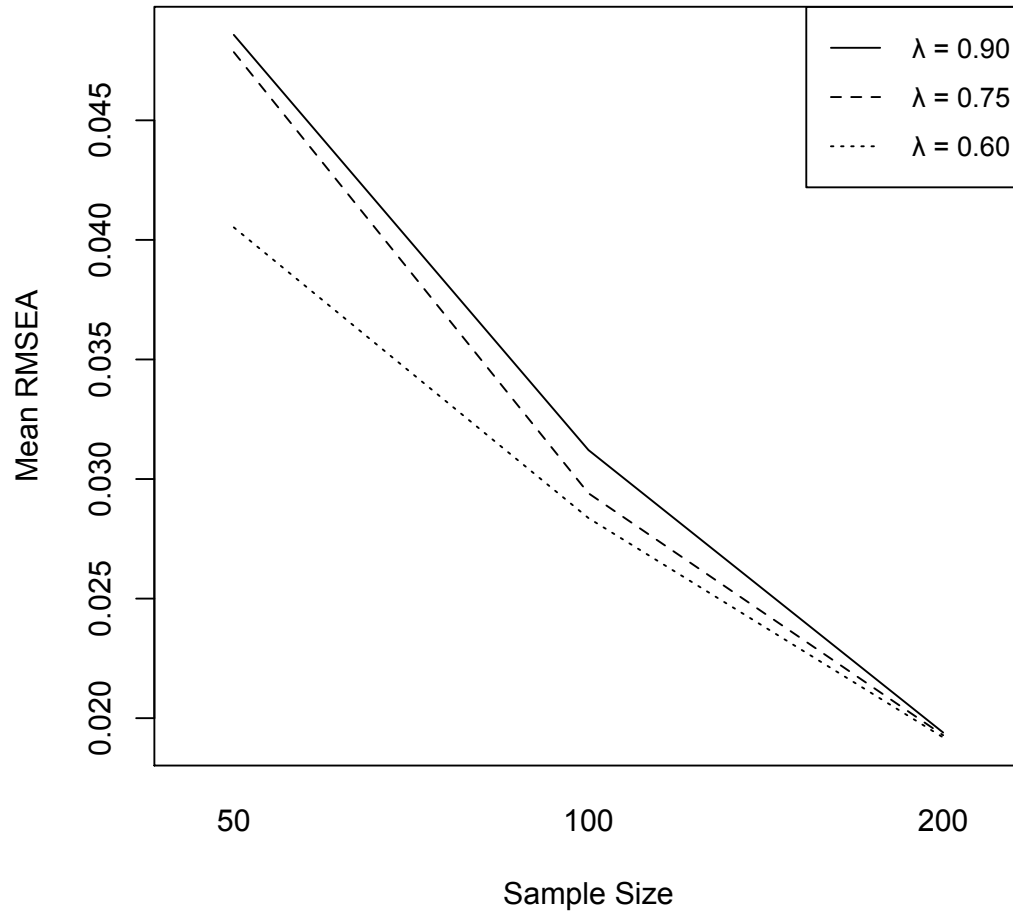
Figure 1.9. Interaction plot of the RMSEA values for the samples sizes of n boostrapped to the same size as n.
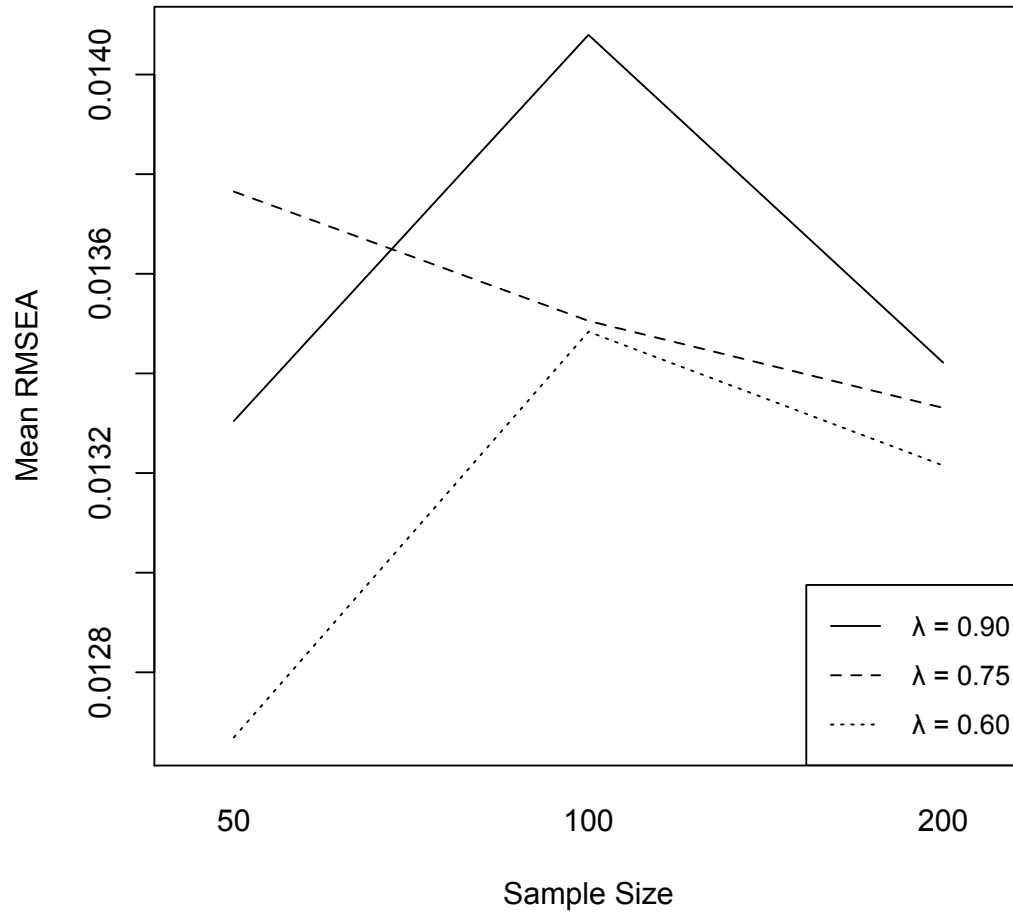
Figure 1.10. Interaction plot of RMSEA values for the bootstrap replicate extensions of the sample sizes n to the population sample size ($N = 400$).

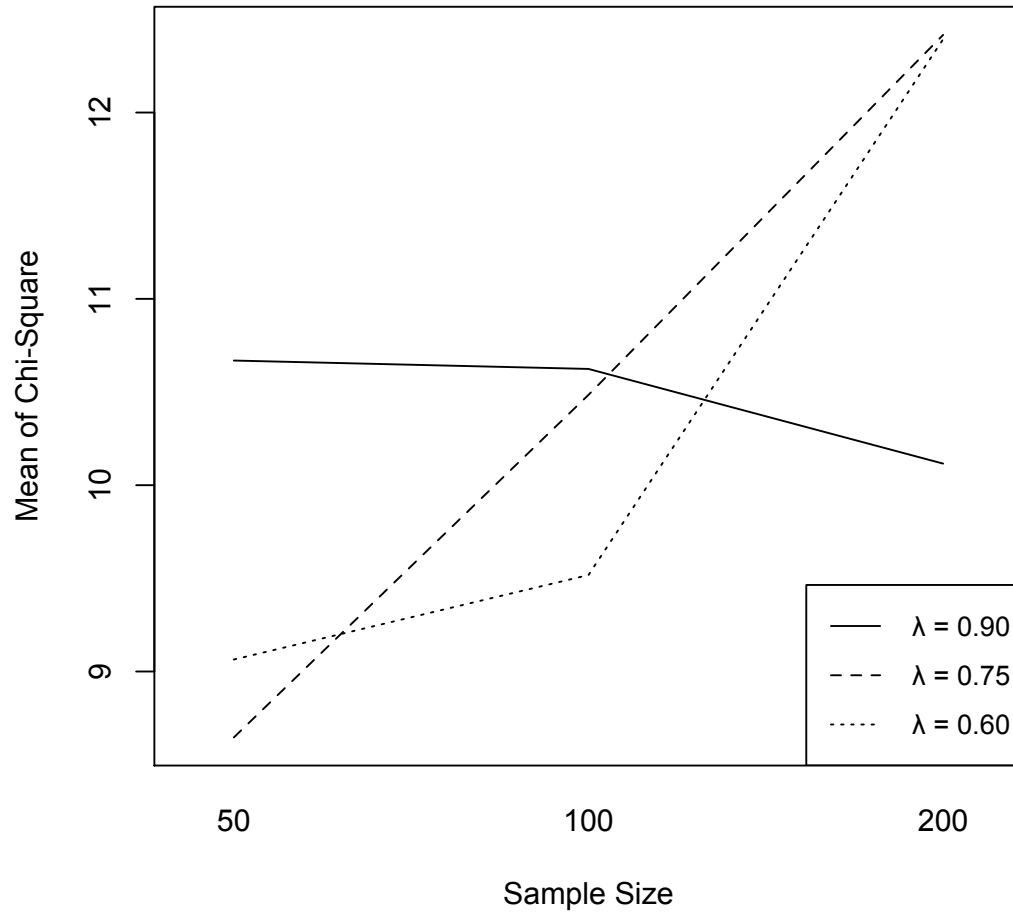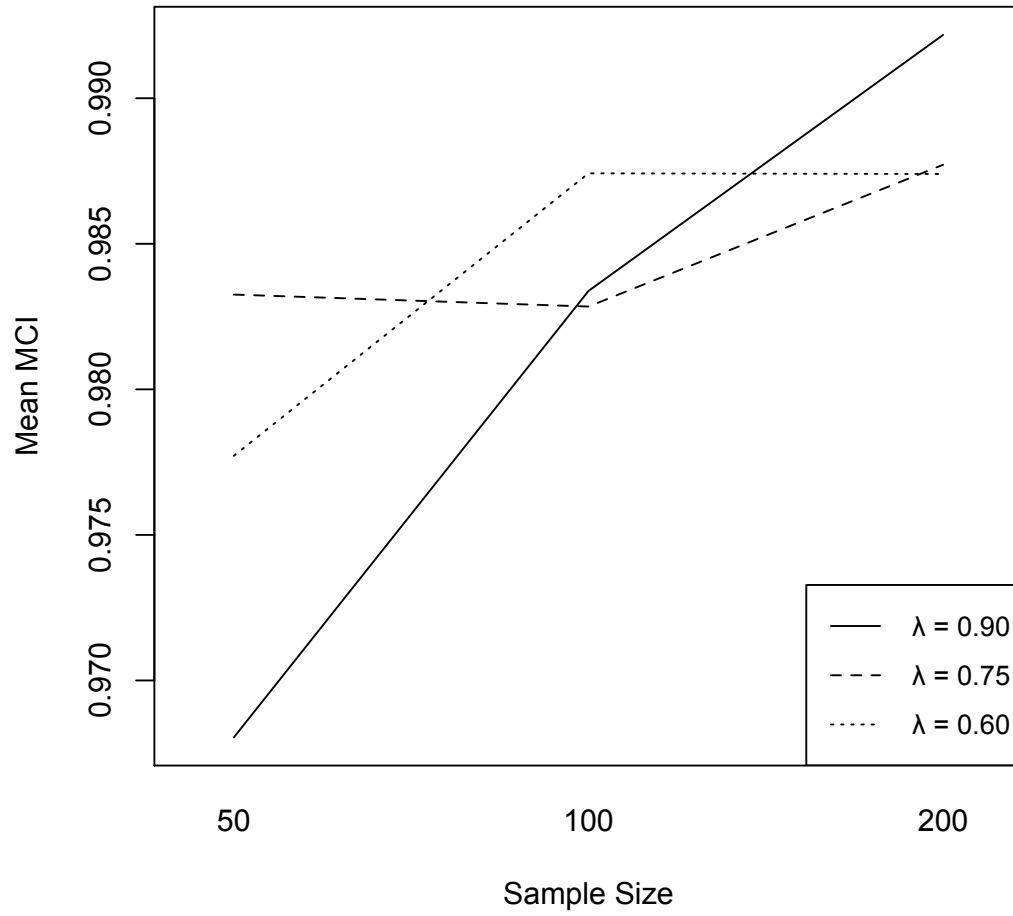Figure 1.11. Interaction plot of MCI values for the initial sample sizes (n).
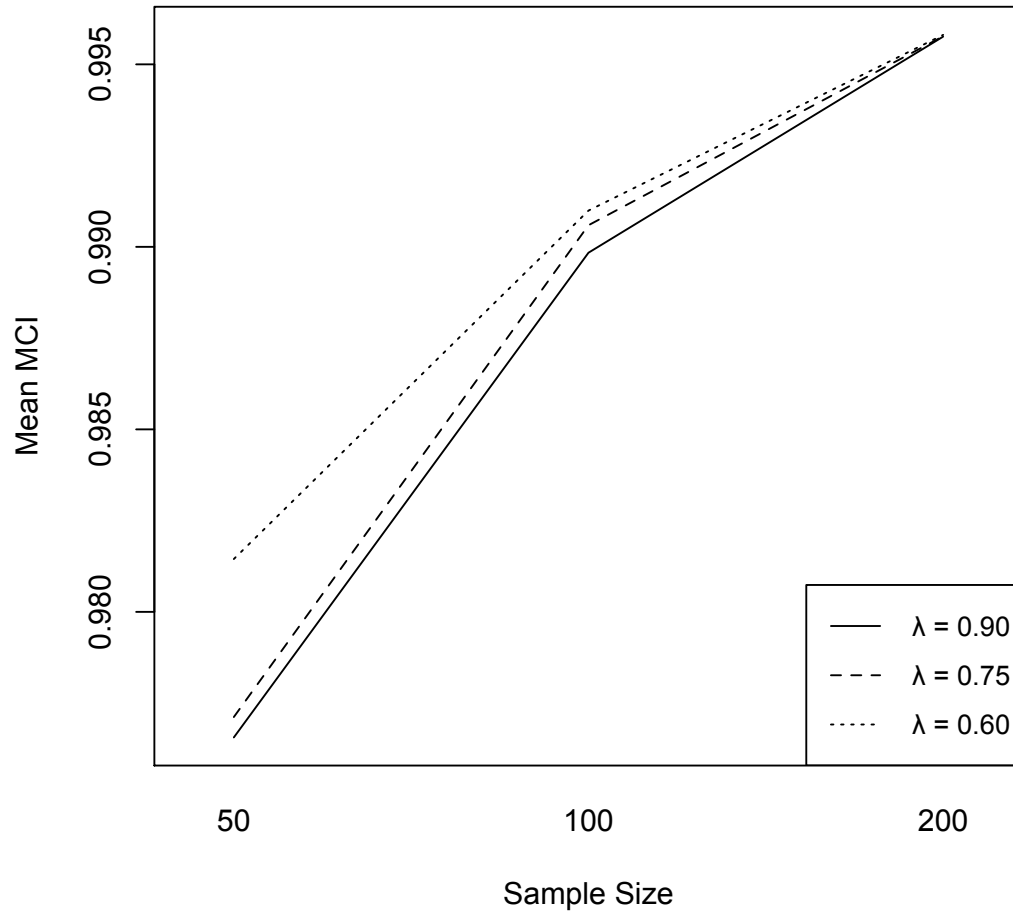
Figure 1.12. Interaction plot of the MCI values for the samples sizes of n boost-rapped to the same size as n.

Figure 1.13. Interaction plot of MCI values for the bootstrap replicate extensions of the sample sizes n to the population sample size ($N = 400$).

MANUSCRIPT 2

Confidence Interval Estimation for SEM Fit Indices

Preface

Manuscript format is in use.

This is a manuscript based thesis submission for a masters degree in psychology at the University of Rhode Island. This is the second manuscript of two for this respective thesis.

Manuscript 2

Title: Confidence Interval Estimation for SEM Fit Indices

Co-Author: Dr. Lisa L. Harlow

Intended Journal: *Psychological Methods*

Status: Not yet submitted, but with the intention to submit by early 2014.

Manuscript is prepared for submission.

Abstract

Many scientific journals now require confidence intervals (CIs) of a mean or median to be reported in any publication. Social science has been slow to follow on this trend, particularly with the very informative yet inconsistently reported field of structural equation modeling (SEM). However, because of the nature of the χ^2 goodness-of-fit test and SEM testing against a null model, a Bollen-Stine transformation is necessary for bootstrapping processes and may have hindered some researchers from doing so (Bollen & Stine, 1992). Many commercial programs and `R` packages now include this feature to bootstrap SEM models. Thus, empirically based CIs and medians can easily be deduced for any index much like the Bollen-Stine p-value which is based on bootstrap resampling of a SEM model to assess its empirical likelihood of fit. This research aims to show the ease of simulating CIs with the `lavaan` package in `R`, along with trying to extrapolate some SEM indices that display narrower and more accurate CIs in small-sample SEM models. Results suggest the standardized root mean-square residual (SRMR) and McDonald's Centrality Index (MCI) have consistently smaller CI ranges across the simulated conditions and could thus be rather consistent and not as sample size dependent as other indices presented and their CI ranges. The SRMR also has a large amount of sample fit indices not contained within the estimated CI for each condition, an indication of sensitivity to model misfit in small n CFA models.

Confidence Interval Estimation for SEM Fit Indices

Confidence intervals (CIs) have been a widely accepted and in many cases required practice of modern day scientific research (Cumming, 2012). The vastly growing field of structural equation modeling (SEM) has, however, been slow to present viable CIs for its vast aspects of fit besides the root mean square error of approximation (RMSEA). This deficit has led some to conclude that fit indices may not be appropriate. For example, Barrett (2007) is a big proponent for the elimination of fit indices whereas Steiger (2007) has countered that with the suggestion that power and confidence intervals be more readily available in SEM research. The RMSEA (Steiger & Lind, 1980) is a fit index that is commonly presented with a 90% CI in SEM results, however it is not empirically based nor have any other indices adopted CIs. Markland (2007) takes notice that there are no "golden" cutoff values for indexes, and therefore CIs could offer a more complete picture.

The potential problem for SEM CIs is that estimation is best done by bootstrapping, a procedure that takes constant resampling with replacement from an empirical set of data to estimate a parameter's sampling distribution as well as its CI (Chernick, 1999; Efron, 1979). A simulation bootstrap process within the SEM framework requires a transformation to the data matrix or else the null hypothesis is not being tested correctly with the χ^2 goodness of fit (Figure 2.1) and will thus lead to erroneous results without a proper shift of the χ^2 distribution. The problem of resampling error in SEM models has been corrected by the Bollen-Stine bootstrap (BSBS) which allows for an adjusted simulated p-value of an empirical SEM model (Bollen & Stine, 1992). This transformation also allows for fit indices to have sampling distributions constructed based on empirical data, and thus 90% CI estimation is possible not only for the RMSEA or χ^2 goodness-of-fit test, but

also for any fit index a research might choose to utilize. The BSBS has proven quite useful to SEM yet it has not become commonplace knowledge in the field yet.

A naive bootstrap, or selecting repeated observations with replacement from the empirical sample, is the simplest bootstrapping approach and generally works with most data. However, the use of naive bootstrapping in SEM has been repeatedly shown as inaccurate (Bollen & Stine, 1992; Chernick, 1999; Enders, 2001; Nevitt & Hancock, 2001). Thus, the Bollen-Stine transformation must be performed so the covariance matrix becomes consistent with the null hypothesis. That is, one can find the resulting Bollen-Stine p-value (BSp) by finding the total number of bootstrapped χ^2 values larger than the χ^2 statistic of the actual empirical dataset used to bootstrap the SEM model divided by the number of replicates. In the following equation, Bollen and Stine (1992) state the transformation for the new data set Z from the initial data where Y represents the centered raw data, S represents the sample covariance matrix of Y, and $\hat{\Sigma}$ is the implied covariance matrix:

$$Z = YS^{-1/2}\hat{\Sigma}^{1/2} \tag{2.1}$$

Bollen and Stine (1992) then mathematically deduce that the covariance matrix of Z being $Z'Z/(N-1)$ is now equal to the implied covariance matrix $\hat{\Sigma}$ and bootstrapping from the transformed data in Z will accurately test the null hypothesis in all bootstrap samples. While this process may seem tedious, the `lavaan` package in `R` comes readily installed with the capability to do BSBS replications and thus the ease for CI estimation is nowhere near as difficult as it might be presumed (R-Core-Team, 2013; Rosseel, 2012). Presented below, is some preliminary information on bootstrapping.

Bootstrapping Basics

Starting with a data matrix of observations such that n is the sample size with c columns each representing a recorded variable. Each row (r) of the matrix acts as a single observation. The $r\,x\,c$ matrix of parameters (p) is expected to be an adequate representation of the expected population parameters θ_p. The $r\,x\,c$ matrix for any p can be used in a bootstrapping framework such than each row of the matrix will be available for every replicate with replacement to create a bootstrap based data matrix $\hat{\theta}_p$ of randomized rows from θ_p. Every row has a probability of $1/r$ chances of selection regardless of any previous selections. With thousands of bootstrap replicates (B), $\hat{\theta}_B$ will contain B samples of size n, with r randomly selected rows from θ_p with replacement. A rather thorough picture of the sampling distribution can thus be assumed for any of the parameters p in the B $\hat{\theta}_p$ matrices.

Efron (1979) notes that data from an unknown distribution F where $X_1, X_2, \ldots, X_n \sim F$ with replicates of $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ has an expected mean of

$$\bar{x} = \sum_{1}^{n} \frac{x_n}{n} \tag{2.2}$$

As pointed out by Efron (1979), $\bar{x}$ should approximate the expected mean of the sample $\mathrm{E}_F[X]$ because of the numerous replications representing a Taylor series expansion. The variance of the sampling distribution from the bootstrap replicates can thus be found by $\mathrm{E}_F[X^2] - \mathrm{E}_F[X]^2$. The standard error of the sampling distribution $(\hat{\sigma})$ can thus be found with

$$\hat{\sigma}(F) = \left(\frac{\mathrm{E}_F[X^2] - \mathrm{E}_F[X]^2}{n} \right)^{1/2} \tag{2.3}$$

Bootstrap estimates of the CI can be constructed via a non-parametric approach that finds the desired cutoff values of the empirically based probability mass function. Finding CI estimates from bootstrapping will be briefly described, shortly. Efron and Tibshirani (1986) state that a minimum of $B = 1000$ replicates are necessary for non-parametric CI estimation because of its complex nature. Teetor (2011) displays the use of the `quantile` function in R and the use of an even larger number of replicates (9,999) to provide an empirical estimation of the CI.

In the case of SEM, the Bollen-Stine bootstrap (BSBS) transforms the data so that it is properly testing the χ^2 statistic (Bollen & Stine, 1992). Each $\hat{\theta}_p$ will represent a bootstrap sample from which a covariance matrix can be determined and the SEM model fit for every one of the thousands of replicates. The parameters of interest would typically be the B χ^2 values, however fit indices of each fitted model could also be estimated B times.

Modern computing power has made bootstrapping readily available, but the understanding of the process itself can perhaps seem daunting. Appendix A displays the structure of a typical bootstrap process, with the resulting $\hat{\theta}_B$ matrices allowing for a simulated mean or median value for starters, as well as a model based CI of any percentage. The `quantile` function in `R` can easily compute the cutoff points for a string of means or medians based on $\hat{\theta}_B$ at a given CI level.

As an example, a particular fit index of an SEM model, say the SRMR, could be estimated using $B = 2000$ replicates, or double the minimum B suggested previously (Efron & Tibshirani, 1986). The 90% CI could be found as such for the SRMR ($\hat{\theta}_{SRMR}$) which includes all of the model fit SRMR indices from 2000 replicates. To find a 90% CI, we would need to estimate the cutoff for the lower (0.05) and upper (0.95) proportion of scores, and the command in `R` with a vector of the 2000 SRMR estimates (`theta.srmr`) would look like:

```
quantile(theta.srmr, c(0.05, 0.95))
```

This function could be extended easily by adding a third quantile value (which is contained between the `c( )` in the command), that measures the median of the sample. The `quantile` function in `R` works similar to constructing a box-plot, but with the user having the ability to specify cutoff points for the CI. In this case, we could find the 90% CI and the median for $\hat{\theta}_{SRMR}$ as such:

```
quantile(theta.srmr, c(0.05, 0.50, 0.95))
```

A similar process can thus be done for any given fit index of a SEM model, including the χ^2. While tedious, a typical model would only need one set of bootstrap replicates conducted, thus bootstrapping is not as time consuming as one might perceive. Bootstrapping packages do exist in `R` and most commercial programs contain simulated bootstrap additions as well. Allowing the computer to store all of the $\hat{\theta}_p$ matrices and then compute the necessary information allows for a not so complicated bootstrap process that adds great depth to any researchers expertise. The resulting bootstrap process assumes that the large number of $\hat{\theta}_p$ replicates will adequately map the sampling distribution of θ_p. In doing so, difficult parameters (i.e., fit indices) can be estimated with CIs that are based on the data as opposed to complex mathematical processes.

Simple Bootstrapping Examples

A very simple bootstrap procedure could be done with something as simple as the roll of a die. Say we roll the same die repeatedly 20 times, the numbers would never change on the die, but the resulting values all have an equal chance of selection with replacement. Using the `quantile` function in `R`, we could find a basic 90% CI of the mean for a standard dice with recorded values stored in a variable we'll call `dice`.

```
> dice
[1] 6 3 3 5 3 2 2 5 1 3 1 1 3 5 1 4 5 3 3 4
> mean(dice)
[1] 3.15
> quantile(dice, c(0.05, 0.50, 0.95))
   5%    50%    95%
 1.00   3.00   5.05
```

While the sample of one die is generic, it gives the basic concept of what bootstrap replication accomplishes (i.e., a mean, median, CI). Now if we were to roll five dice at a time, and take the mean of each roll 20 times, we'd be able to construct a CI around the mean of the five dice. Note that each roll can have a mix of any numbers, ranging anywhere from all different values to all the same.

```
> dice
[1] 4.4  3.2  3.4  3.2  2.8  3.4  2.8  4.2  2.8  2.8
    3.4  2.8  3.0  4.2  3.6  3.0  3.4  3.6  2.8  4.2
> mean(dice)
[1] 3.35
> quantile(dice, c(0.05, 0.50, 0.95))
  5%    50%    95%
2.80   3.30   4.21
```

The reason for the difference in means for one die ($M = 3.15$) and five dice ($M = 3.35$) can be described by the varied nature of resampling. With one die, each replicate is only represented by a single die, and thus only 20 total values were included. With five dice, the replicates were the mean of five dice done 20 times, thus 100 total dice values comprised the 20 means. The median and CIs displayed the same pattern, however 20 replicates are certainly not enough and therefore results are slightly varied. Doing this same experiment over thousands of replicates would result in means, medians, and CIs all relatively similar since the compiled $\hat{\theta}_B$ will be a close approximation to θ_p. Any variable or variables in a data set can be done in this same simple bootstrap process (e.g., CFI, χ^2, RMSEA, etc). Instead of dice, one could let a computer randomly select r in each replicate

55

so that $(\hat{\theta}_p)$ is equivalent in length to θ_p. Extending $\hat{\theta}_p$ by drawing a larger value of observations r than the data set is possible as well, but bootstrapping to the same size n as the sample data is usual.

Applying Bootstrap CIs to SEM Fit Indices

To find the 90% CI of any size of bootstrapped index or parameter, one must find the interval from the lowest 5% of replicated values to the largest 5%, which can be done with the `quantile` function in `R` (Teetor, 2011). The `quantile` limits thus provide is the likelihood that the median value will fall within these values 90% of the time, and therefore the smaller the CI the better presumption that the simulated parameter is accurate. It should also then be presumed that the empirical index values should also fall within this CI range or the index might not be well representative of the model. Some researchers warn against bootstrapping SEM models of less than sample size 100 (Nevitt & Hancock, 2001), and others state SEM sample size should be at least 200 for accuracy reasons (Kenny, 2012; Kline, 2011). However, work with small n models with simple designs could be possible and needs further research (Nevitt & Hancock, 2001). In essence, small sample size fit indices could be verified or rejected by CI estimation using the BSBS procedure, and therefore empirical CIs with a median could be readily presented for any one of the fit indices in SEM.

It is thus expected that CI construction will be relatively easy when the appropriate BSBS process is used. From the sampling distributions of various models, 90% CIs can be constructed along with a median index value to supplement the actual index values. Because of the minimal amount of CI research in SEM, this research is more about the process and an exploratory procedure to determine what fit indices seemingly behave well in small n SEM models as it relates to CI estimation.

Methods

A simulation study varying several small to moderate sample sizes ($n = 50, 100, 200$), small to moderate factor loadings ($\lambda = 0.60, 0.75, 0.90$), and bootstrap replicated of size n as well as back to the initial population matrix derived size of 400 was used in a $3\,x\,3$ factorial design. Because of the minimal research with small samples in SEM, it seems important to start here with these different conditions for CI estimation and then move forward. Because of the small n constraint, a simple six-item and two-factor confirmatory factor analysis (CFA) model will be used (Figure 2.2).

A selected set of commonly used fit indices will be monitored with a 90% CI similar to that commonly presented with the RMSEA, but with an empirical bootstrap basis. Two absolute fit indices (χ^2 goodness-of-fit; and standardized root mean-square residual, SRMR) will be used, which are generally covariance matrix based indices (Kline, 2011). Three relative fit indices (Tucker-Lewis Index, TLI; Bollen's incremental fit index, IFI; and normed fit index, NFI) (Bentler & Bonett, 1980; Bollen, 1989; Tucker & Lewis, 1973) will be used with the caveat that the TLI, IFI, and the comparative fit index (CFI) show heavy correlation with one another (Kenny, 2012). And finally, three non-centrality indices (McDonald's centrality index, MCI, CFI, and RMSEA) (Bentler, 1990; McDonald, 1989; Steiger & Lind, 1980) which measure the distance of the proposed model from a null model will be used.

For robustness, 30 multiple random samples in each of the nine simulated conditions were chosen. All data were simulated to form integer based 7-point Likert scales similar to those used in much of the SEM literature. The samples for the nine conditions were pulled from a population matrix based sample created from

the varied levels of n and λ. Population matrices were constructed randomly and a set sample of 400 was drawn as a population sample for the sake of comparison. The samples of size n were drawn from the population samples of size 400 and were also randomly generated for all 30 samples in each of the nine conditions, for a total of 270 simulated models. Figure 2.3 illustrates the organization of the methods used in this simulation.

Efron and Tibshirani (1986) make note that bootstrap replication for CI estimation should be a minimal of $B = 1,000$ replicates because of the complex nature of CIs. Thus $B = 2,000$ replicates was used for each initial sample along with three λ levels and two bootstrap sizes using the `lavaan` package in R (R-Core-Team, 2013; Rosseel, 2012). The sample generation code is presented in appendix B. A simulated BSp as well as simulated values for nine fit indices were examined for all of the 270 simulated models comprised by the nine combinations of n and λ used in the study, as well as for the bootstrap replicates to size n and back to the population sample size of $N = 400$.

The range of the CI is of importance and should be analyzed, as well as the likelihood of the fit indices from the model actually being contained within the simulated 90% CI. Any size CI can be constructed, but this research opted for a 90% interval since it is commonly used with the RMSEA. The standard RMSEA is mathematically deduced, so using the non-parametric BSBS to get a RMSEA CI is also of interest.

Results

Tabulations of the model fit indices encapsulated by their respective 90% CIs for the varied conditions of n and λ can highlight subtle differences in the indices (Table 2.1). CIs for when $n = 50$ failed to contain the fit index values of the samples quite often, with the exception of the χ^2 and SRMR (Table 2.1). The NFI had the lowest rate of indices contained by their CI for $n = 50$, with only 3 for $\lambda = 0.60$, 1 model for $\lambda = 0.75$ and no model indices contained for $\lambda = 0.90$. The TLI, CFI, and IFI had only modest performance and only 3 CIs containing the indices for all three when $n = 50$ and $\lambda = 0.90$.

The smallest CI ranges (Table 2.2) in the sample size n 90% CI bootstrap estimates were observed in the SRMR and MCI indices whereas the RMSEA and TLI indices had the largest CI regions for the 0 to 1 bound indices. The χ^2 goodness-of-fit had a median value of 8.347 for simulated samples, which is relatively close to the 8 df's of the model. The 90% CI range for the median lower and upper confidence levels for the χ^2 was 14.138.

Table 2.3 displays the ratios of the simulated medians of the bootstrap replicates for all models compared to the actual model index values as well as a 90% CI for the ratios of the 270 bootstrapped models in relation to their original values. The TLI ($Med = 1.0004$), IFI ($Med = 1.0002$), and CFI ($Med = 1.0001$) were all very close to an ideal value of 1, or identically similar bootstrap values to original sample index values. The RMSEA ($Med = 1.492$) and χ^2 ($Med = 1.145$) had simulated model indices farther from their original sample fit indices. Table 2.3 also shows the RMSEA with by far the largest CI range of ratio values for the replicates compared to original model indices $[0.000, 4.710]$. The MCI had the smallest 90% CI range of simulated to original index values $[0.948, 1.026]$ as well

as a median ratio of the bootstrap to sample values ($Med = 1.0009$).

The dispersion of the indices with a zero to one boundary (Figure 2.4) suggests very little difference in the spread of the indices at the bootstrap level of sample size n. Furthermore, the SRMR, MCI, and RMSEA are without outliers in the box-plots (Figure 2.4), where as the TLI has the largest spread of values and outliers. Plotting the difference between the lower confidence limit (LCL) and the median compared to the upper confidence limit (UCL) and the median (Figure 2.5) for several indices with small CI ranges shows the SRMR and MCI both have minimal amounts of variation in their CIs. The CIs for the SRMR and MCI have the smallest differences between the limits and the median as seen in Figure 2.5 as well.

Discussion

Fit index CIs definitely have a strong potential of use for SEM researchers, as the levels of confidence can add great depth to any fit index. The use of the `lavaan` package in `R` is one example of simplifying the BSBS procedure to achieve not only a BSp-value but also 90% CIs for any or all of the fit indices available through the `bootstrapLavaan` function (see Appendix B). It would be helpful to have bootstrap fit CI values become regularly available to all SEM researchers in an easy to execute manner.

Because of the ease and added bonus fit index CIs can present to SEM research (i.e., providing more insight into the degree of uncertainty), simulated index values and CIs would be helpful to include with the actual index values selected by the researcher. In this presented research with smaller sample sizes, the SRMR and MCI appear to be preferable because of their minimal CI size variance regardless of n or λ. The SRMR also seems to be rather consistent with including the actual fit index value in its bootstrapped CI, which could be beneficial in regards to small n SEM models.

With a simple CFA model of sample sizes of 50, 100, and 200 the SRMR and MCI seemed to behave with the least amount of variance. Across all sizes of n and λ, the minimal variances and differences in means between the MCI and SRMR appear to be least sensitive to parameter variation. The RMSEA similarly showed patterns of being effective across all nine conditions, but resulted with slightly larger CI ranges compared to the SRMR and MCI (Figure 2.5). The TLI, IFI, NFI, and CFI behaved similarly across conditions which was likely due to their high intercorrelations, and all four have the widest variance of CI ranges for the varied sizes of n and λ.

Many of the indices were only moderate at simulating CIs that contained their respect fit index value from the sample itself (Table 2.1). The NFI most notably was only adequate for when $n = 200$ and thus its use in small sample SEM could be ill-advised, further supporting the notion that the NFI is over-sensitive to parameters (Ding et al., 1995; Ding, 1996. The SRMR managed to stay consistent regardless of n and λ and could be an ideal fit index for small samples as well as bootstrap CIs.

The MCI and the SRMR both had the smallest CI ranges based on the median values of the 270 bootstrapped intervals. Since a small CI range is ideal, the MCI and SRMR seem viable options for empirically based CIs. The CFI and IFI also had rather small ranges for the 270 simulated CIs (Table 2.2). The ratios of the median values for the bootstrap replicates in comparison to the actual sample fit indices again suggests the MCI as one of the closest to an ideal ratio of 1:1 from fit index to bootstrap fit index as well as having a minimal 90% CI range based on the these values (see Table 2.3). The IFI and CFI had slightly closer ratios to 1 than the MCI, but had 90% CIs for their respective 270 ratios as almost double the size of the MCI CI range.

The addition of a χ^2 goodness-of-fit 90% CI would be valuable to any researcher within SEM, and should therefore be considered for usage whenever possible. The benefits of an added CI to the χ^2-test would undoubtedly add some increased depth to the research presented in the field. Research presented with an empirical CI for the χ^2 of a SEM model lends credence to how strong the model may or may not be and the model becomes clearer to the reader and even allows for added ability of power and effect size estimation for current and future use.

Within the eight indices presented in this research, the SRMR has a strength of bootstrap stability (i.e. the fit index being contained in the 90% CI) as well as

having a minimal amount of variance within the CI ranges themselves. The SRMR thus seems to have an edge over the other indices in its minimal CI ranges and stability of the sample SRMR value being contained within its bootstrapped CI. The MCI displayed small CI ranges as well regardless of n and λ and also added a strong ratio of fit index to bootstrap fix index median being close to 1:1 and having minimal variance in the ratios. The IFI and CFI both have benefits in these areas as well, and could be worth adding to a SEM CI estimation.

The use of several indices as opposed to just one in general is advised (Kline, 2011) and therefore the SRMR may very well be a good option for any researcher to use or continue using as a supplement to the χ^2 goodness-of-fit. The MCI could make a good compliment to the research as well as the SRMR and χ^2, and mixed with empirical CIs for all three would add strength to SEM research. Due to the high correlation between the IFI and CFI, either one could make a viable addition as well to the fit indices used with bootstrap CIs, but certainly both are not necessary.

A limitation of this research is that it was conducted with only one CFA model based on several possible conditions (e.g., three sample sizes and three factor loading sizes). Future research could examine additional conditions including different SEM models types and sizes. Further, larger sample sizes and varied λ levels could be investigated in future research. Another limitation is that the artificiality of any simulation may not always generalize to real-world research. However, the constructed data sets in this research were made to be integer based and similar to what a real Likert scale study would look like. Still, true empirical data would obviously lend more credit to this CI estimation method and would be complementary to the current findings.

Future research could also include statistical power analysis as a factor to

further the information that can be presented with a given SEM model. Power in SEM is a much more difficult task than CI estimation, and thus may be better to address on its own. However, both power and CI estimation would offer more clarity and depth to SEM research, as suggested by Steiger (2007). Although many of the common and well-supported SEM indices were included in the current research, other indices can also be examined with the procedures presented in this study.

Other research could also focus directly on the expansion of small sample sizes to larger ones with bootstrapping. Analysis investigating how large the bootstrap extensions should be (i.e., to what extended sample size) and what fit indices would be best to utilize. Another possible bootstrapping approach to SEM could focus on one model with varied subsamples of different sizes. Overall model fit might be good, but comparison of the model to specific groups within the overall sample (i.e., gender or race) may not have balanced subsample sizes and thus comparison could be slightly difficult. The issue to address then becomes, should a researcher bootstrap smaller samples to the size of the larger samples, or could smaller replicated bootstrap samples of the larger subsample sizes be more adequate?

The need for CIs in any research has become obvious and in many cases a requirement (e.g., Cumming, 2012). Taleb (2010) sums it up best by suggesting that no measure of central tendency should ever be trusted without a range of confidence. SEM has been surprisingly absent of CIs other than for the RMSEA, and thus the goal of this research is to help illustrate the necessity and the rather non-complex nature of adding 90% CIs to all published models. Modern computing has opened many new gateways and it has become rather easy in the 21st century to conduct repeated iterations of any model in a bootstrap fashion, especially since many programs and packages already allow replication. Researchers

need to become more familiar with these methods and journals need to provide more guidelines for SEM publication requirements, which should include CIs and adequately justified fit indices *a priori*.

Bibliography

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

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

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide.* John Wiley and Sons Inc: New York.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge: New York.

Ding, L., Velicer, W. F., and Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2*(2), 119–143.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352–370.

Kenny, D. A. (2012). Measuring model fit. Retrieved electronically October 22, 2012 from http://www.davidakenny.net/cm/fit.htm.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* The Guilford Press: New York. 3rd edition.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on paul barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences, 42*, 851–858.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*(3), 353–377.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Personality and Individual Differences, 24*(2), 291–293.

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. IN the annual meeting of the Psychometric Society, Iowa City, IA (Vol. 758).

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*, 893–898.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable.* Random House Trade Paperbacks. 2nd edition.

Teetor, P. (2011). *R cookbook: Proven recipes for data analysis, statistics, and graphics.* O'Reilly Publishing: Beijing. 1st Edition.

Table 2.1. Summation of model fit index values contained within the simulated 90% CI broken down by sample size (n) and factor loading (λ). Each cell has a max of 30 samples.

Index	n	$\lambda = 0.60$	$\lambda = 0.75$	$\lambda = 0.90$
	$n = 50$	30	26	27
χ^2	$n = 100$	30	27	27
	$n = 200$	30	25	26
	$n = 50$	27	25	26
SRMR	$n = 100$	27	25	26
	$n = 200$	27	24	24
	$n = 50$	11	10	3
TLI	$n = 100$	18	14	9
	$n = 200$	19	19	14
	$n = 50$	12	10	3
IFI	$n = 100$	18	14	9
	$n = 200$	19	19	14
	$n = 50$	3	1	0
NFI	$n = 100$	12	9	7
	$n = 200$	30	25	25
	$n = 50$	14	15	19
RMSEA	$n = 100$	17	16	19
	$n = 200$	20	19	15
	$n = 50$	11	10	3
CFI	$n = 100$	18	14	9
	$n = 200$	19	19	14
	$n = 50$	20	21	11
MCI	$n = 100$	19	18	12
	$n = 200$	20	19	15

Table 2.2. Median values for lower and upper 90% confidence limits as well as for the fit index for bootstrap samples. The CI range for each index based on the median LCL and UCL values is also included.

Index	LCL	Median	UCL	CI Range
χ^2	2.68060	8.34702	16.81871	14.13811
SRMR	0.02159	0.04025	0.06308	0.04149
TLI	0.89678	0.97772	1.00000	0.10322
IFI	0.94743	0.98871	1.00000	0.05257
NFI	0.90387	0.95212	0.98501	0.08114
RMSEA	0.00000	0.02972	0.10619	0.10619
CFI	0.94495	0.98812	1.00000	0.05505
MCI	0.95547	0.99035	1.00000	0.04453

Note: LCL = Lower 90% Confidence Limit

UCL = Upper 90% Confidence Limit

Table 2.3. Median and 90% confidence levels for the ratio of the sample index to the bootstrap sample median.

Index	LCL	Ratio	UCL	CI Range
χ^2	0.44135	1.14503	2.48610	2.04475
SRMR	0.65666	1.09250	1.81646	1.15980
TLI	0.84365	1.00038	1.07877	0.23512
IFI	0.92192	1.00020	1.03461	0.11269
NFI	0.90623	0.99161	1.03836	0.13213
RMSEA	0.00000	1.49188	4.71028	4.71028
CFI	0.91740	1.00006	1.03954	0.12214
MCI	0.94834	1.00086	1.02622	0.07788

Note: LCL = Lower 90% Confidence Limit
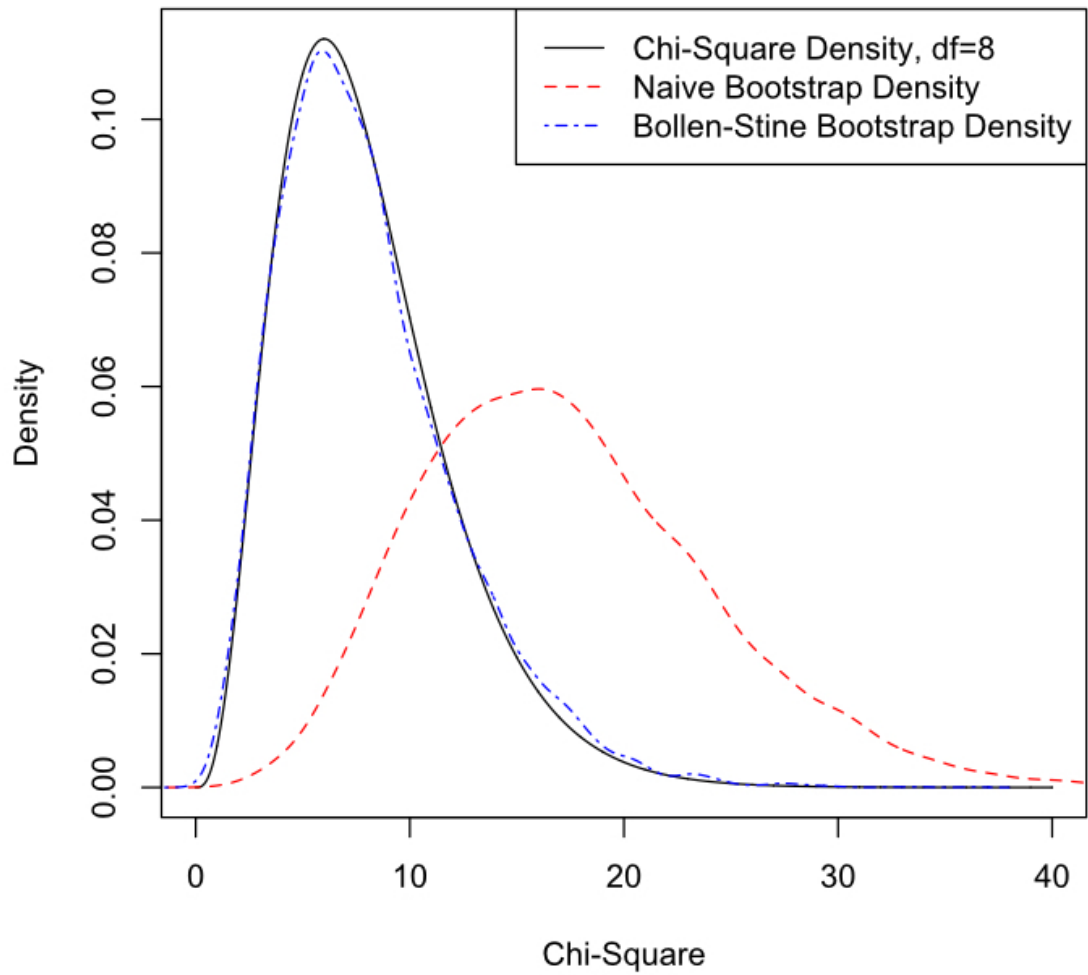
UCL = Upper 90% Confidence Limit

Figure 2.1. Density of the χ^2 distribution with $df = 8$ as well as the standard bootstrap and the Bollen-Stine bootstrap. The Bollen-Stine transformation is a necessary step in χ^2 resampling.
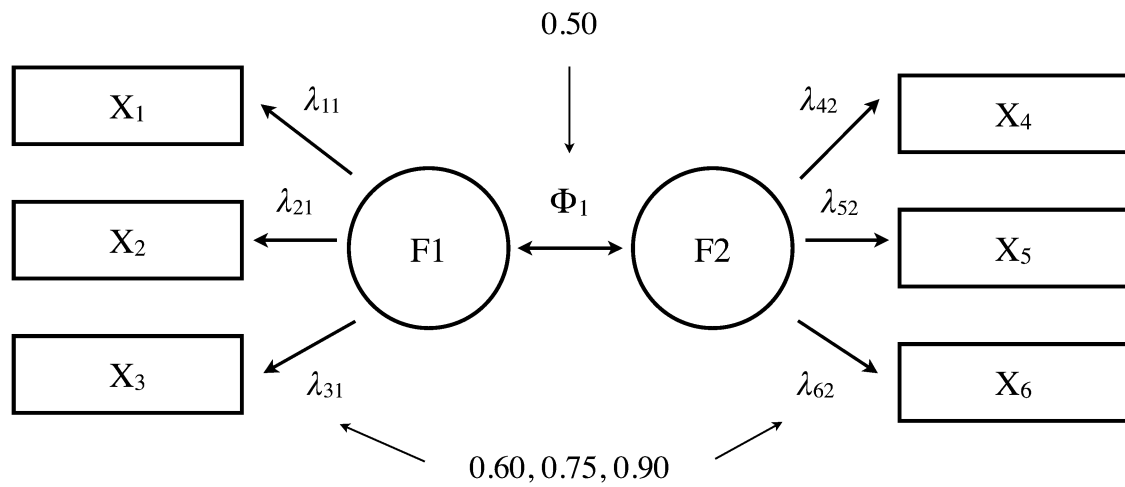
Figure 2.2. Two factor model utilized for the presented simulation study. Factor loadings (λ) are varied between 0.60, 0.75, and 0.90 where as the correlation between the two factors (F1 and F2) is set to 0.50.
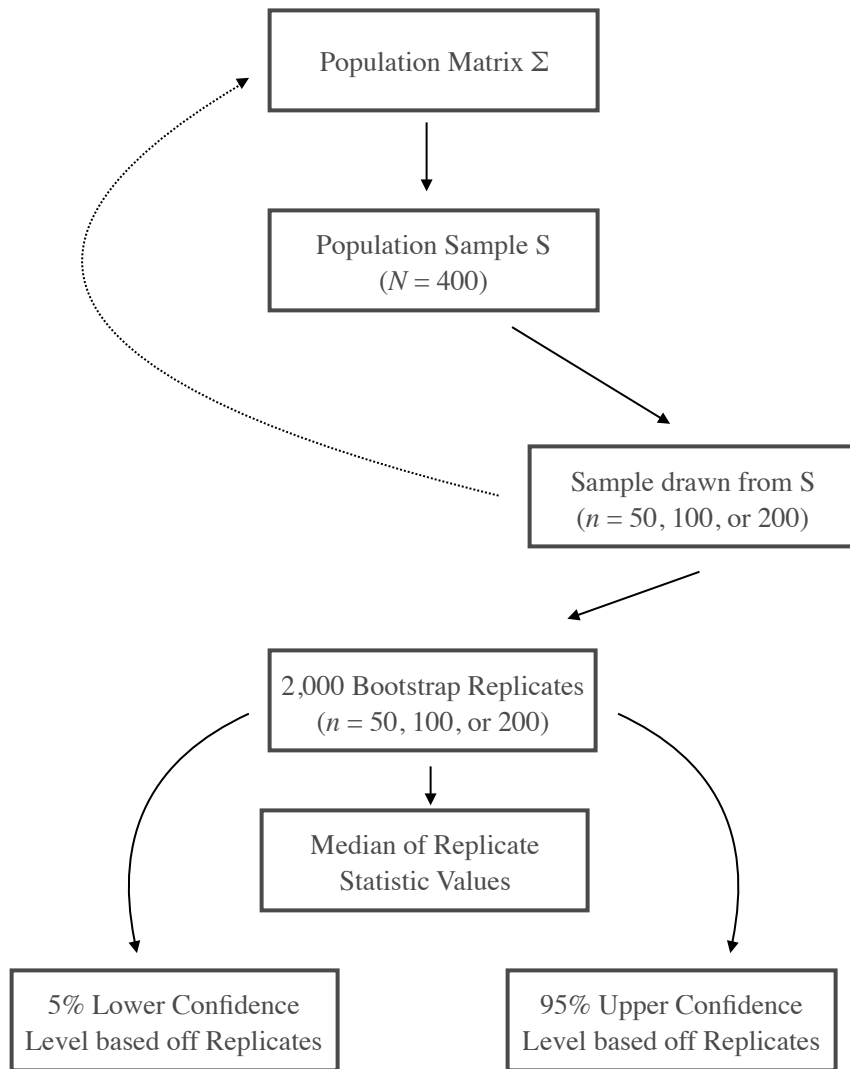
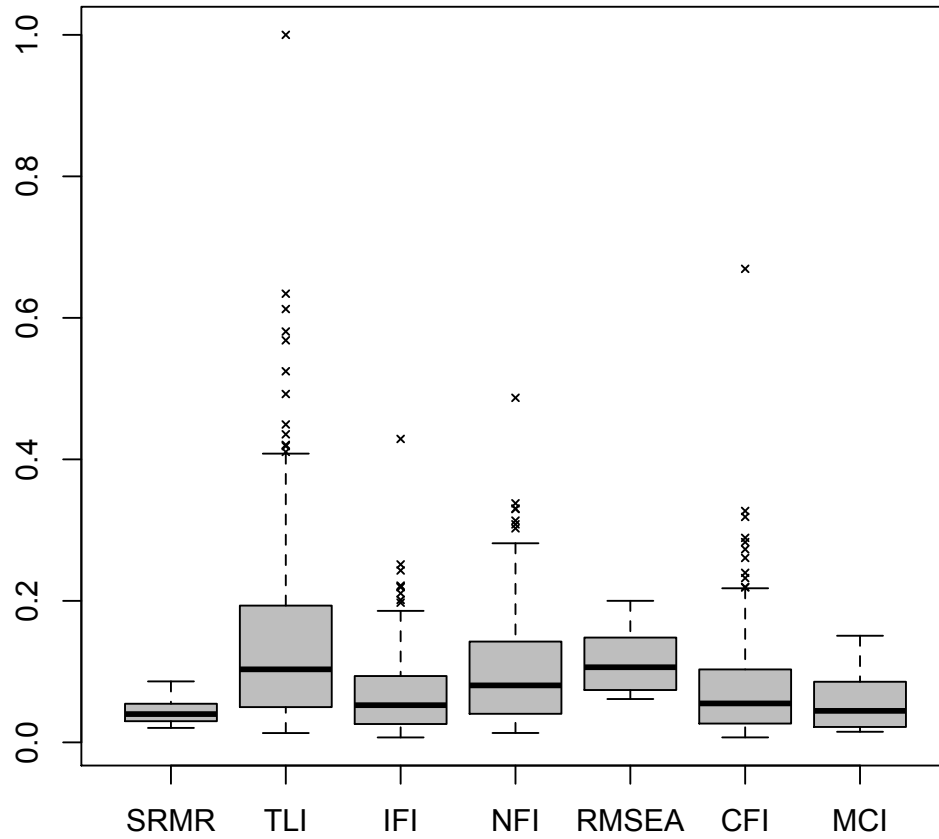Figure 2.3. Flowchart of the simulation based design used.

Figure 2.4. Box-plots of the fit indices with 0 to 1 boundaries based on there CI ranges when bootstrapped back to the sample size of n (50, 100, 200). All three simulated factor loadings (0.60, 0.75, 0.90) are accounted for.
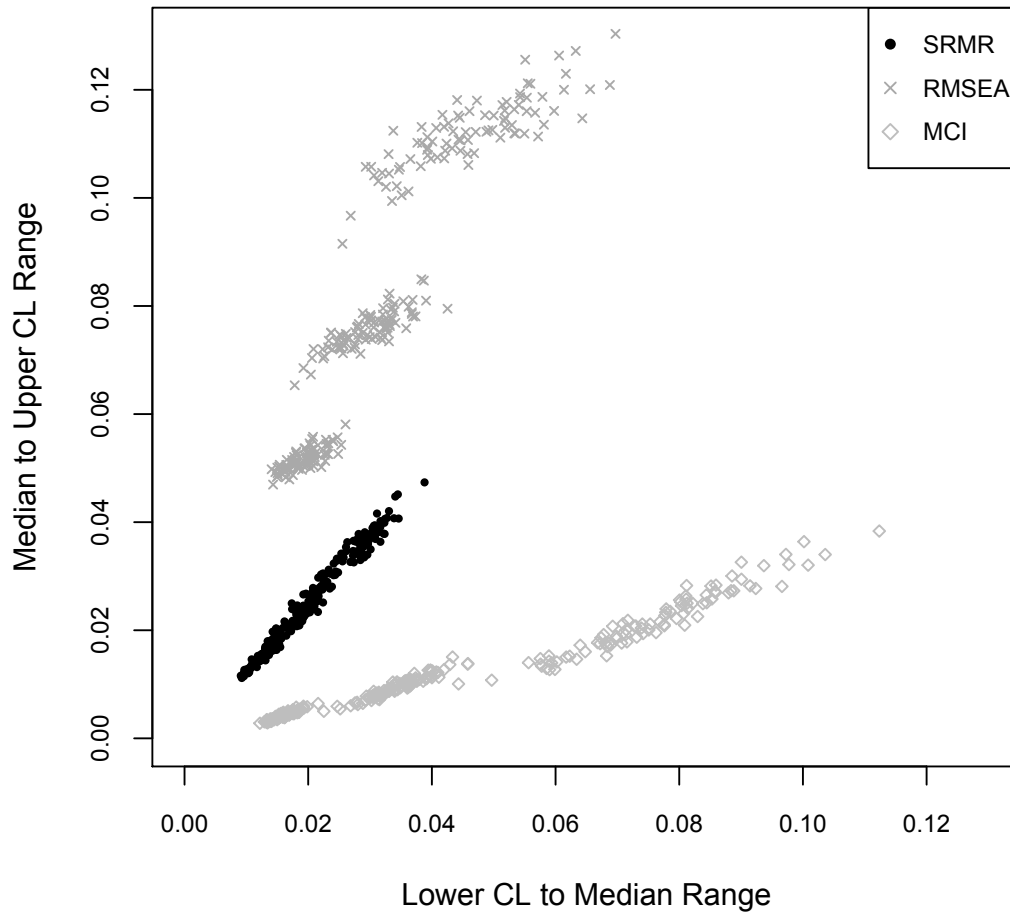
Figure 2.5. Scatterplots of several fit indices (SRMR, RMSEA, MCI) and the dispersion of their CI ranges. Comparison is made between the distance from the lower confidence level to the median for all 270 bootstrap results on the x-axis and the difference between median and upper confidence level on the y-axis.
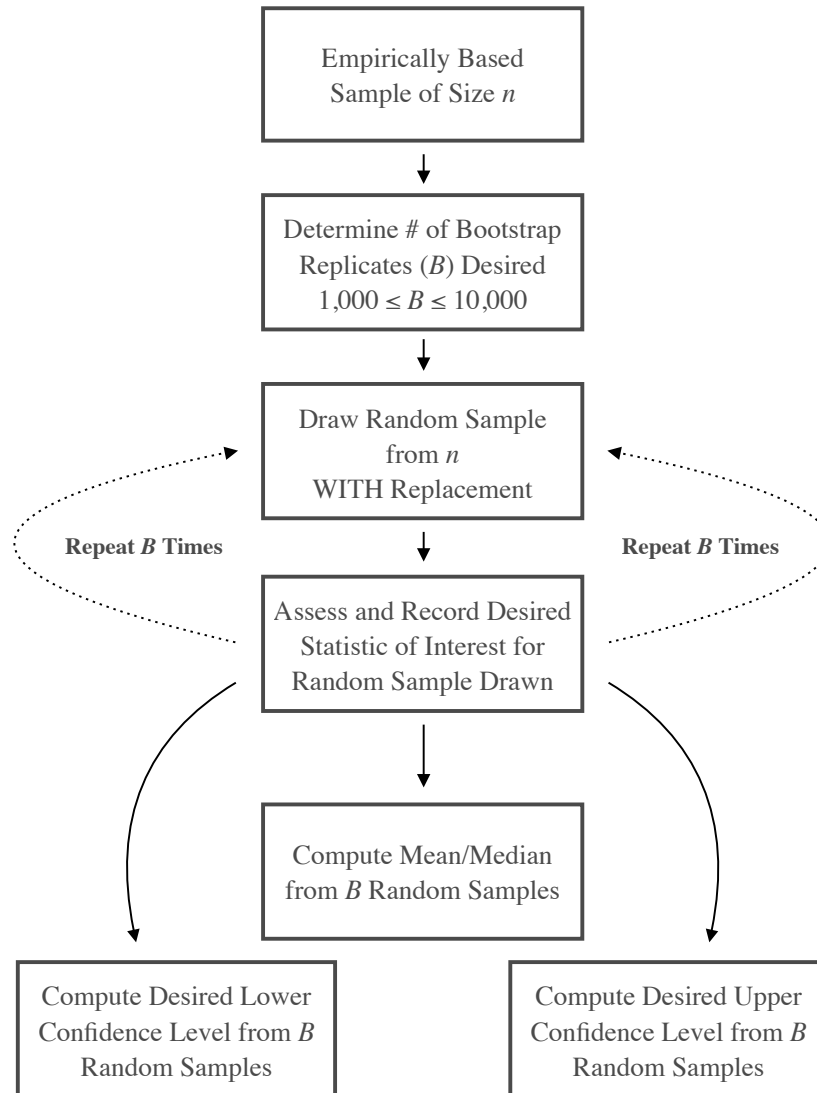
APPENDIX A

Basic Bootstrap Flowchart

Figure A.1. General flowchart of the basic bootstrapping process based on the work of Efron (1979). This design can be used for parameter estimation, standard error estimation, confidence limits, or any combination of these.

APPENDIX B

Simulation Code for Use in R

```
overall  <- NULL
library(lavaan)
library(tmvtnorm)
library(stringr)

latCor <- 0.50    # Fixed Correlation between Latent Variables

for(seed in 1001:1031)
{
n <- 50    # Varied Sample Size n (50,100,200)
lam <- 0.60    # Varied Lambda Loading (0.6,0.75,0.9)

# lavaan Model Syntax
Adv02Fact  <- ' A =~ 1*X1 + X2 + X3
     B =~ 1*X4 + X5 + X6
     A ~~ B '

meanie <- c(3.5,3.9,3.7,4.5,4.1,3.7)    # Selected Variable Means

# Create Population Matrix according to Parameters
a  <- matrix(c(lam,lam,lam,0,0,0,0,0,0,lam,lam,lam),ncol=2)
b  <- matrix(c(1,latCor,latCor,1),ncol=2)
pop  <- a %*% b %*% t(a)
diag(pop)  <- rep(1,6)
dimnames(pop)  <- list(c("X1","X2","X3","X4","X5","X6"),
                       c("X1","X2","X3","X4","X5","X6"))

set.seed(seed)

f0@Fit@converged <- FALSE
while(f0@Fit@converged==FALSE)
{
   # Initial Sample and Fit
   s0 <- data.frame(trunc(rtmvnorm(400, mean=meanie, sigma=pop,
                    lower=rep(1,6),upper=rep(8,6))))
   f0 <- cfa(Adv02Fact, data=s0)
}
f1@Fit@converged <- FALSE
while(f1@Fit@converged==FALSE)
```

```
{
# Sample and fit for size n
loc <- sample(1:400,n,replace=TRUE)
s1  <- s0[loc,]
f1  <- cfa(Adv02Fact, data=s1)
}
# Fits Model with Bollen-Stine Bootstraps
b1 <- bootstrapLavaan(f1, R=2000,type="bollen.stine",verbose=T,
                        FUN=fitMeasures,fit.measures=c("pvalue",
                        "chisq","srmr","aic","tli","ifi","nfi",
                        "rmsea","cfi"))

# Need to Calculate the MCI by hand
d  <- ((b1[,"chisq"]/(n-1))-8/(n-1))
mci  <- exp(-.5 * d)
b1 <- cbind(b1,mci)

# Adjust Indices to assure they are bounded by 0 and 1
b1[which(b1[,"tli"]> 1),"tli"]  <- 1
b1[which(b1[,"ifi"]> 1),"ifi"] <- 1
b1[which(b1[,"mci"]> 1),"mci"] <- 1
b1[which(b1[,"tli"]< 0),"tli"] <- 0
b1[which(b1[,"ifi"]< 0),"tli"] <- 0
b1[which(b1[,"mci"]< 0),"mci"] <- 0

# Bootstrap n back to 400
s2 <- s1
for(i in 2:(400/n))
{
   s2 <- rbind(s2,s1)
}
f2 <- cfa(Adv02Fact, data=s2)

# Fits Model with Bollen-Stine Bootstraps
b2 <- bootstrapLavaan(f2, R=2000, type="bollen.stine",verbose=T,
                        FUN=fitMeasures,fit.measures= c("pvalue",
                        "chisq","srmr","aic","tli","ifi","nfi",
                        "rmsea","cfi"))

# Need to Calculate the MCI by hand
d  <- ((b2[,"chisq"]/(360-1))-8/(360-1))
mci  <- exp(-.5 * d)
b2 <- cbind(b2,mci)
```

```r
# Adjust Indices to assure they are bounded by 0 and 1
b2[which(b2[,"tli"]> 1),"tli"]  <- 1
b2[which(b2[,"ifi"]> 1),"ifi"] <- 1
b2[which(b2[,"mci"]> 1),"mci"] <- 1
b2[which(b2[,"tli"]< 0),"tli"] <- 0
b2[which(b2[,"ifi"]< 0),"tli"] <- 0
b2[which(b2[,"mci"]< 0),"mci"] <- 0

# BSBS Calculation
bs1 <- sum(b1[,"chisq"]>fitMeasures(f1,"chisq"))/length(b1[,1])
bs2 <- sum(b2[,"chisq"]>fitMeasures(f1,"chisq"))/length(b2[,1])

# Non-Convergent Models
nc1 <- 2000-length(b1[,1])
nc2 <- 2000-length(b2[,1])

# Create .csv Based on Parameters
data <- str_c(n,"to",dim(s1)[1],"L",lam*100,"raw",seed,".csv")
write.csv(b1, data)
data <- str_c(n,"to",dim(s2)[1],"L",lam*100,"raw",seed,".csv")
write.csv(b2, data)
```

Bibliography

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42:815–824.

Beauducel, A. and Wittmann, W. W. (2005). Simulation study on fit indexes in cfa based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1):41–75.

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

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in covariance structures. *Psychological Bulletin*, 88:588–606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research*, 17(3):303–316.

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

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Phd thesis, University of Groningen.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide.* John Wiley and Sons Inc, New York.

Ding, L. (1996). *Evaluation of structural modeling fit indices.* PhD thesis, University of Rhode Island, Kingston, RI.

Ding, L., Velicer, W. F., and Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, 2(2):119–143.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4):352–370.

Gao, S., Mokhtarian, P. L., and Johnston, R. A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record: Journal of the Transportation Research Board*, 2082:116–124.

Gerbing, D. W. and Anderson, J. C. (1992). Monte carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research*, 21:132–160.

Guadagnoli, E. and Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103:265–275.

Harlow, L. L. (1985). *Behavior of some elliptical theory estimators with non-normality data in a covariance structures framework: A Monte Carlo study.* Phd dissertation, University of California, Los Angeles.

Hu, L. and 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.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202.

Kenny, D. A. (2012). Measuring model fit. Retrieved electronically October 22, 2012 from http://www.davidakenny.net/cm/fit.htm.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* The Guilford Press, New York, 3rd edition.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on paul barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences*, 42:851–858.

Marsh, H. W., Balla, J. R., and 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.

Nevitt, J. and Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8(3):353–377.

R-Core-Team (2013). *R: A language and environment for statistical computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing. http://www.R-project.org/, Vienna, Austria.

Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences*, 24(2):291–293.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Sanchez, G. (2012). R package plsdepot: PLS regression 1. Retrieved electronically May 5, 2013 from http://gastonsanchez.com/plsdepot.

Sharma, P. N. and Kim, K. H. (2012). *A comparison of PLS and ML bootstrapping techniques in SEM: A Monte Carlo study.* Presentation at the 7th International Conference on Partial Least Squares and Related Methods, Houston, TX. Retrieved electronically on October 1, 2012 from http://www.pitt.edu/

Sharma, S., Mukherjee, S., Kumar, A., and Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for model fit in covariance structure models. *Journal of Business Research*, 58:935–943.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42:893–898.

Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable.* Random House Trade Paperbacks, New York, 2nd edition.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, 58(1):134–146.

Teetor, P. (2011). *R cookbook: Proven recipes for data analysis, statistics, and graphics.* O'Reilly Publishing, Beijing, 1st edition.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., and Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, 48(1):159–205.

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

Widaman, K. F. and Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1):16–37.