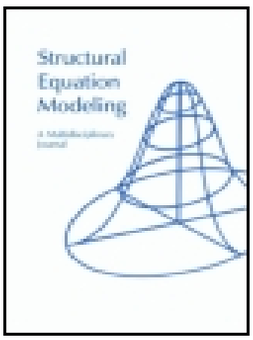

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Assessing Model Fit in Structural Equation Modeling Using Appropriate Test Statistics

Katerina M. Marcoulides, Njål Foldnes & Steffen Grønneberg

To cite this article: Katerina M. Marcoulides, Njål Foldnes & Steffen Grønneberg (2019): Assessing Model Fit in Structural Equation Modeling Using Appropriate Test Statistics, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2019.1647785

To link to this article: https://doi.org/10.1080/10705511.2019.1647785

Published online: 24 Sep 2019.

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–11, 2019

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2019.1647785

Assessing Model Fit in Structural Equation Modeling Using Appropriate Test Statistics

Katerina M. Marcoulides,[1] Njål Foldnes,[2] and Steffen Grønneberg[2]
[1]*University of Florida*
[2]*BI Norwegian Business School*

The assessment of model fit has received widespread interest by researchers in the structural equation modeling literature for many years. Various model fit test statistics have been suggested for conducting this assessment. Selecting an appropriate test statistic in order to evaluate model fit, however, can be difficult as the selection depends on the distributional characteristics of the sampled data, the magnitude of the sample size, and/or the proposed model features. The purpose of this paper is to present a selection procedure that can be used to algorithmically identify the best test statistic and simplify the whole assessment process. The procedure is illustrated using empirical data along with an easy to use computerized implementation.

Keywords: Structural equation modeling, model fit, bootstrapping

Assessing the fit of a proposed model in structural equation modeling (SEM) applications is of paramount importance to researchers in the social, behavioral, business, educational, and medical sciences. This is because any elaboration concerning the parameter estimates or the connecting relationships among examined variables is conditional upon establishing support for the proposed model. The assessment of model fit is typically carried out on the basis of a test statistic that uses the likelihood ratio statistic T_{ML} based on the normality assumption (Lawley & Maxwell, 1971). Overall model fit evaluation is in effect conducted by comparing this likelihood ratio statistic against the nominal chi-square distribution or other descriptive or alternative fit indices such as the root mean square error of approximation (RMSEA, Steiger & Lind, 1980) or the comparative fit index (CFI, Bentler, 1990). Most currently available SEM computer programs (e.g., AMOS, EQS, LISREL, *lavaan*, M*plus*, & *OpenMx*) automatically supply these and many other fit indices as standard output.

When sampled data are normally distributed and the number of observations is sufficiently large, the likelihood ratio statistic T_{ML} compared to a chi-square distribution is expected to perform well (Amemiya & Anderson, 1990). With non-normal data and smaller sample sizes, however, the likelihood ratio can significantly deviate from the chi-square distribution, including in settings with either complete or missing data (Enders, 2001; Hu, Bentler, & Kano, 1992). Of course, in realistic research situations sampled data generally tend to be non-normally distributed (Micceri, 1989). To complicate matters further, when T_{ML} follows a chi-square distribution in large samples, it is difficult to determine precisely how large a sample size is needed, particularly with tested models having relatively large numbers of variables (Deng, Yang, & Marcoulides, 2018). In circumstances where the distribution of T_{ML} is likely not well approximated by a chi-square, one may choose rescaled and adjusted test statistics for overall model evaluation (e.g., rescaled statistic T_{SB}, developed by Satorra & Bentler, 1994), as these are expected to perform robustly under violations of normality assumptions (Hu et al., 1992; Satorra & Bentler, 1988, 1994). Depending of course on the severity of the assumption violations, conditions might still exist that can interfere with the accurate applications of these test statistics (Yuan, Yang, & Jiang, 2017). In such situations, it is likely that none of these test statistics can be

Correspondence should be addressed to Katerina M. Marcoulides, Email: kmarcoul@umn.edu Department of Psychology, University of Minnesota, Elliott Hall 75 East River Rd., Minneapolis, MN 55455, USA.

Katerina M. Marcoulides has changed affiliations from University of Florida to University of Minnesota since the submission of this article.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hsem.

well described by a chi-square distribution (Bentler & Yuan, 1999; Foldnes & Grønneberg, 2018; Wu, 2018; Wu & Lin, 2016).

Assessing model fit in SEM is unquestionably quite complicated as the selection of the test statistic depends on the various distributional characteristics of the sampled data, the sample size, and the proposed model characteristics. To make matters worse, no definitive guidance has been given in the extant literature on which test statistic to adopt as the best option across various model, sample size, and distributional conditions. The only actuality that past research has supported is that there does not appear to be a single test statistic that will universally outperform others (Foldnes & Grønneberg, 2018; Foldnes & Olsson, 2015; Grønneberg & Foldnes, 2019; Wu, 2018; Wu & Lin, 2016; Yuan et al., 2017). Given the plethora of available test statistics, researchers undoubtedly need access to a process that can objectively guide them to make the difficult choice of selecting a statistic to use as a basis for model fit evaluation.

This article seeks to address this selection activity and contribute to the literature by presenting an innovative approach with practical strategies that researchers can use whenever they are faced with assessing the fit of a proposed model in SEM applications. The approach relies upon a resampling selection procedure that, as will be shown in subsequent sections, enables the approximate determination of the objectively most suitable test statistic. We advocate that this approach essentially removes researchers from having to scrutinize the various distributional characteristics of the sampled data, the magnitude of the sample size, and/or the proposed model characteristics in order to determine the appropriate test statistic. This selection activity is performed algorithmically and fits nicely within the general framework of bootstrapping techniques where the focus is on methods capable of reconstructing the sampling distribution associated with the original studied population (Bollen & Stine, 1993; Efron & Tibshirani, 1993; Marcoulides, 1990).

The remainder of the paper is organized in the following way. The next section presents a brief description of four commonly used model evaluation rescaled and adjusted test statistics. This is followed by a presentation of a resampling method that can be used to approximately identify the best performing test statistic for the given data and model conditions. Next, an illustrative empirical example in which a simple confirmatory factor analytic model is considered and a description of results obtained from analyses of the example data is presented. Finally, a discussion of the implications of the methodology and findings is provided.

TEST STATISTICS

Let us consider $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ to be a simple random sample from a population of interest represented by $\mathbf{x}$, with means $E(\mathbf{x_i}) = \mu$ and covariances $Cov(\mathbf{x_i}) = \mathbf{\Sigma}$. Assuming that to model the covariances we use a model $\mathbf{\Sigma(\theta)}$ and estimate the parameter $\boldsymbol{\theta}$ by minimizing the following discrepancy function

$$F_{ML}(\theta) = \mathrm{tr}\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\Sigma^{-1}(\theta)\right| - p \qquad (1)$$

where S is the sample covariance matrix and $\mathbf{\Sigma(\theta)}$ is the proposed model[1]. We seek to apply an inference procedure in order to determine whether the null hypothesis H_o: $\mathbf{\Sigma = \Sigma(\theta)}$ holds at some critical value. Assessment of model fit based on the normality assumption is then carried out using a likelihood ratio test statistic defined as

$$T_{ML} = (N - 1)\, F_{ML} \qquad (2)$$

and compared against the chi-square distribution χ^2_{df} corresponding to the selected probability value for statistical inference. Comparing the obtained test statistic to a critical value at a given level of significance is a key phase for assessing model fit (Marcoulides & Yuan, 2017). We note that presently T_{ML} is the default test statistic invoked in most available SEM computer packages.

It is a general and well-known phenomenon that the likelihood ratio test statistic T_{ML} tends to reject the correct model more often than expected when the sample size is not sufficiently large. As a consequence, numerous other test statistics have been developed in an attempt to correct the likelihood ratio statistic. These various developments have also addressed situations with non-normally distributed data. The corrections essentially attempt to regulate potentially problematic situations by simply replacing the $(N - 1)$ in the formulation of the T_{ML} test statistic. These corrections are variations of the formula originally proposed by Bartlett (1950) to correct the behavior of this test statistic in exploratory factor analysis with questionable sample sizes. The formula proposed by Bartlett (1950) to test for the number of factor (m) replaced the $(N - 1)$ value with $N_{Bartlett} = N - p/3 - 2m/3 - 11/6$. Because research established that the original Bartlett correction maintained satisfactory Type I error rates in SEM models with non-normal data and small sample sizes, different formulations of corrections to the T_{ML} test statistic have been deemed worthy of further investigation and application (Fouladi, 2000; Nevitt & Hancock, 2004; Yang, Jiang, & Yuan, 2018).

Two of the most widely used test statistics that take into consideration the potential non-normality of the data are the rescaled statistics T_{SB} and T_{YB} (Satorra & Bentler, 1994;

[1] Although mean structure is an important aspect of SEM, for ease of presentation in this article we focus mainly on covariance structure models.

Yuan & Bentler, 2000). The first rescaled test statistic is more commonly referred to as the Satorra-Bentler test statistic and is defined as

$$T_{SB} = \frac{d}{tr\left(\widehat{U}\widehat{\Gamma}\right)} T_{ML} \tag{3}$$

where d = degrees of freedom, $\widehat{U}$denotes an estimate of a matrix dependent on the model, and $\widehat{\Gamma}$is an estimate of the asymptotic covariance of the empirical covariance matrix. Model fit is then assessed by comparing T_{SB} against the chi-square distribution χ^2_{df} with d degrees of freedom corresponding to the selected probability level for statistical inference (for complete details, see Satorra & Bentler, 1988). Research has shown that this test statistic improves Type I error rates over the likelihood ratio statistic when the data are non-normally distributed, but the test can still over reject correct models when the sample size is not large enough (Foldnes & Olsson, 2015; Nevitt & Hancock, 2004; Yang et al., 2018). The T_{YB} is a closely related test statistic that was proposed by Yuan and Bentler (2005, Equation 20), in which the correction factor is calculated using a method that accommodates missing data.

Another popular correction formulation was also provided by Asparouhov and Muthén (2010) that both scales and shifts the value of T_{ML}. For this reason, the test statistic is commonly denoted as the "scaled-and-shifted" test using T_{SS} and is defined as

$$T_{SS} = \sqrt{\frac{d}{tr\left[\left(\widehat{U}\widehat{\Gamma}\right)^2\right]}} \cdot T_{ML} + d - \sqrt{\frac{d \cdot \left[tr\left(\widehat{U}\widehat{\Gamma}\right)\right]^2}{tr\left[\left(\widehat{U}\widehat{\Gamma}\right)^2\right]}} \tag{4}$$

Model fit is similarly assessed by comparing T_{SS} against the chi-square distribution χ^2_{df} with d degrees of freedom corresponding to the selected probability level for statistical inference. It should be noted that all of the above test statistics can be obtained in most available software by using the appropriate estimator command option. For example, in the R package *lavaan* (Rosseel, 2012) the command "*test = satorra.bentler*" would supply the T_{SB} test statistic, whereas in M*plus* (Muthén & Muthén, 2018) the command to obtain this test statistic is "*estimator = mlm*".

TEST STATISTIC IDENTIFICATION USING RESAMPLING METHODS

Ideally, a statistical hypothesis test with a simple null-hypothesis should produce p-values that are uniformly distributed when one repeatedly samples from a population in which the null hypothesis holds true. In theory, one can use this property to identify the most appropriate test statistic. Of course, in a practical situation the population is not known and so there is no way to resample from it. However, resampling methods can at least be used to emulate such a situation by resampling instead from the sample. In such a situation one can thereby investigate whether a test statistic is in fact able to produce uniform p-values in a condition that emulates the population condition. Statistical inference using resampling methods was first popularized by Efron (1979) in his monumental paper on bootstrap approaches. With bootstrapping, statistical inferences could be made by intensive computations and provide solutions to problems that would otherwise be intractable (Efron & Tibshirani, 1993). Within the field of SEM, applications of bootstrap methods originate from the work of Beran and Srivastava (1985) and subsequently popularized by Bollen and Stine (1992) as well as Yung and Bentler (1996), and more recently by Grønneberg and Foldnes (2019). The bootstrap method starts by transforming the sample observations $\mathbf{x_i}$ into $\mathbf{x_i}$* using the following formulation

$$\mathbf{x}_i* = \Sigma\left(\hat{\theta}\right)^{\frac{1}{2}} S^{\frac{-1}{2}} \mathbf{x}_i \tag{5}$$

for i = 1, 2, …, n, where S and $\Sigma\left(\hat{\theta}\right)$ are respectively the sample and model implied covariance matrices. Noting next that the proposed model holds exactly in the transformed sample data, it is subsequently assumed that the transformed sample data can serve as a proxy of the population from which the original sample was drawn, provided the model holds true in the population. As a result, the ideal test statistic will produce p-values that are uniformly distributed and that by repeated bootstrap sampling the distribution of the p-values can be approximated and evaluated.

An application of the standard bootstrap method was recently suggested by Grønneberg and Foldnes (2019) for testing model fit in SEM. Through a bootstrap selection mechanism, the method identifies the test statistic among any set of possible candidates that exhibits the best sampling distribution of the p-values for the observed data and model conditions. It does this by selecting the best test statistic that most closely follows a uniform distribution through an evaluation of a Kolmogorov-Smirnov distance metric. Despite the promising performance of this method demonstrated by Grønneberg and Foldnes (2019), some researchers have criticized the Kolmogorov-Smirnov distance metric for not being sensitive in establishing distances between analyzed distributions and suggested that instead other distance metrics should be preferred (Babu & Rao, 2004; Stephens, 1974). Grønneberg and Foldnes (2019) also recognized the potential limitations of the Kolmogorov-Smirnov distance metric, advocating the

examination of other criteria to be an important topic for future research.

Following this call, the current work makes use of an alternative metric for selecting the best test statistic in SEM applications that is based on the Anderson-Darling method. This metric is motivated by the original contributions of Anderson and Darling (0, 1952)) in which they showed that the metric has several advantages over the Kolmogorov-Smirnov distance metric in terms of overall sensitivity and gives more weight to the tails of the examined distribution. The alternative Anderson-Darling metric was also selected in the current work because past research has shown that it is the best and most powerful option among the different available criteria, under a variety of data conditions (Arshad, Rasool, & Ahmad, 2003; Lilliefors, 1967; Razali & Wah, 2011; Yazici & Yolacan, 2007). To determine the Anderson-Darling (AD) metric, the following formula can be applied

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[lnF(w_i) + \ln(1 - F(w_{n-i+1}))] \quad (6)$$

where n = sample size, $F(w_i)$ = cumulative distribution function of the specified distribution to be tested, and w_i = the ordered data (see also D'Agostino & Stephens, 1986).

The perspective taken in this work is that the selection of the appropriate test statistic using the AD metric as a guide can be automated so that, once a proposed model is formulated, it can be objectively tested against observed data and the approximately most reliable test statistic be identified through an estimation procedure based on bootstrapping. Note that the method will not guarantee that the most reliable statistic is chosen, since this depends on the true distribution of the data, which is unknown. Instead, we select the most reliable test statistic based on the distribution of the transformed bootstrap distribution. If the model holds, this is a good approximation to the actual unknown distribution, and the selected method will therefore also be approximately the most reliable statistic available.

This process also fits nicely within the general framework of unsupervised data mining algorithms where the focus is on developing and applying algorithms capable of finding solutions based exclusively on the inspected data (which is the reason they are frequently referred to as machine learning algorithms; Marcoulides & Falk, 2018). The proposed algorithm is implemented in the publicly available software program R (R Development Core Team, 2010) with all models fit using the *lavaan* R package (Rosseel, 2012). The choice of *lavaan* was primarily based on its flexibility, ease of use, native integration within R, availability, and popularity. Appendix A presents the necessary R code needed to implement the approach using the illustrative data examples described in the next section. It is important to note that although the programming code can be easily modified to evaluate any number of candidate model test statistics, in this work we restrict our selection to those test statistics described above and currently available in the *lavaan* package (Rosseel, 2012).

AN ILLUSTRATIVE EXAMPLE

For the purpose of demonstrating the proposed approach, a simple confirmatory factor analysis (CFA) example is applied in which a two-factor model is tested against data. Two different data settings are considered. The first data set analyzed is based on a random sample of n = 194 complete cases out of an original data set containing 2,800 observations, which is included in the R package *psych* (Revelle, 2015b). The second data set analyzed is based on a random sample of n = 775 cases out of the same original data set. The proposed two-factor model investigated in both samples examines the structure of the latent variables "Agreeableness" and "Conscientiousness", each measured with 5 indicators. The indicators correspond to 10 Likert type self-report items from the 'International Personality Item Pool'. The empirical data set was selected for illustrative purposes because it is freely and readily available (the data can also be downloaded directly from "ipi.orig.org").

Testing next the proposed two-factor model using the n = 194 complete cases selected, the overall model fit information based on maximum likelihood estimation presented in Table 1 was produced by the *lavaan* program. Based on this model fit information, one might immediately conclude that the likelihood ratio test statistic T_{ML} is suggesting rejection of the proposed model. But is this the correct decision or should a different test statistic have been considered when assessing the fit of the proposed model? Table 2 provides values of four different model fit test statistics that could be obtained by using the *lavaan* program. Based on the computed test statistics and their corresponding p-values, a decision must now be made regarding which of these statistics to use as a basis for evaluating the fit of the proposed factor model. Undoubtedly the only two test statistics that somewhat favor the proposed model at α = .01 are the T_{SB} and T_{SS} statistics. But are these the best test statistics to use given the distribution of the sampled data, the sample size, and the proposed model characteristics? Which of these two test statistics is the one that best emulates the ideal test statistic and can ultimately be relied on to evaluate model fit?

To address this question our method first transforms the n = 194 sampled observations using Equation (5) so that the empirical covariance matrix of the transformed data is identical to the model-implied covariance matrix from the original sample, resulting in perfect model fit. Using then the transformed sample, 5000 bootstrapped samples are

TABLE 1
Two-Factor Model Fit Information for $N = 194$ Based on ML Estimation

Estimator	*ML*
Minimum Function Test Statistic	61.807
Degrees of freedom	34
P-value (Chi-square)	0.002
Model test baseline model:	
Minimum Function Test Statistic	449.608
Degrees of freedom	45
P-value	0.000
User model versus baseline model:	
Comparative Fit Index (CFI)	0.931
Tucker-Lewis Index (TLI)	0.909
Loglikelihood and Information Criteria:	
Loglikelihood user model (H0)	−2693.850
Loglikelihood unrestricted model (H1)	−2662.947
Number of free parameters	21
Akaike (AIC)	5429.700
Bayesian (BIC)	5495.052
Sample-size adjusted Bayesian (BIC)	5428.564
Root Mean Square Error of Approximation:	
RMSEA	0.070
90 Percent Confidence Interval	0.041 0.098
P-value RMSEA ≤ 0.05	0.114
Standardized Root Mean Square Residual:	
SRMR	0.060

drawn and the test statistics given in Equations (2) through (4) along with their corresponding p-values are computed for each. Next, to choose the best test statistic, the Anderson-Darling metric is calculated using Equation (6) and assessed relative to the calculated distance between the observed distributional approximation of p-values and the ideal uniform distribution. The obtained values for the Anderson-Darling metric are also presented in Table 2 and clearly indicate that the smallest distance value is achieved by the T_{SS} test statistic. Based on these findings, we conclude that the T_{SS} test statistic is the most reliable among the four examined test statistics and proceed to report the p-value of correct model specification to be equal to .043. For ease of interpretation we also present

TABLE 2
Computed Model Fit Test Statistics and P-values for $N = 194$

Test statistic	T_{ML}	T_{SB}	T_{SS}	T_{YB}
p-value	0.002	0.017	0.043	0.007
AD values	157.01	31.13	19.28	56.35

T_{ML} = maximum likelihood; T_{SB} = Satorra-Bentler; T_{SS} = Scaled and Shifted; T_{YB} = Yuan-Bentler. AD = Anderson-Darling distance metric.

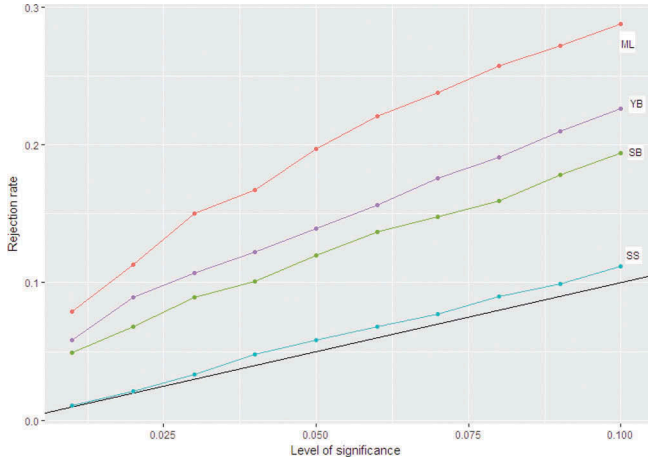

FIGURE 1 Plot of each test statistic against the nominal rejection rates for $n = 194$.
Note. ML = maximum likelihood; SB = Satorra-Bentler; SS = Scaled and Shifted; YB = Yuan-Bentler.

in Figure 1 trace plots for each of the examined test statistics against the nominal rejections rates. As can be seen by examining these plots, other than the T_{SS} test statistic, the other three test statistics deviate substantially from the nominal rejection rates. We note that, as expected, the T_{ML} test statistic deviates substantially from the nominal rejection rates whereas the other two test statistics seem to be closer related in their departure from the rejection rates.

Testing next the proposed two-factor model using the $n = 775$ cases sampled, the overall model fit information provided in Table 3 and four model fit test statistics in Table 4 were produced by the *lavaan* program. Based on the computed test statistics and their corresponding p-values, a decision must again be made regarding which of these statistics to use as a basis for evaluating the fit of the proposed factor model. Following the same procedure as described before, the Anderson-Darling metric is calculated and assessed relative to the calculated distance between the observed distributional approximation of p-values and the ideal uniform distribution. The obtained values for the Anderson-Darling metric are also presented in Table 4 and clearly indicate that the smallest distance value is now achieved by the T_{SB} test statistic. Based on these findings, in the current sampling setting, we conclude that the T_{SB} test statistic is the most reliable among the four examined test statistics and proceed to report the p-value of correct model specification to be equal to 0.000000064. For ease of interpretation we similarly present in Figure 2 the trace plots for each of the examined test statistics against the nominal rejections rates. As can be seen by examining these displayed plots, the T_{ML} test statistic again deviates substantially from the nominal rejection rates whereas the other test statistics seem to be closer related in their

TABLE 3
Two-Factor Model Fit Information for $N = 775$ Based on ML Estimation

Estimator	*ML*
Minimum Function Test Statistic	174.001
Degrees of freedom	34
P-value (Chi-square)	0.000
Model test baseline model:	
Minimum Function Test Statistic	1411.776
Degrees of freedom	45
P-value	0.000
User model versus baseline model:	
Comparative Fit Index (CFI)	0.898
Tucker-Lewis Index (TLI)	0.864
Loglikelihood and Information Criteria:	
Loglikelihood user model (H0)	−11898.659
Loglikelihood unrestricted model (H1)	−11811.659
Number of free parameters	21
Akaike (AIC)	23839.318
Bayesian (BIC)	23935.801
Sample-size adjusted Bayesian (BIC)	23869.119
Root Mean Square Error of Approximation:	
RMSEA	0.075
90 Percent Confidence Interval	0.064 0.086
P-value RMSEA ≤ 0.05	0.000
Standardized Root Mean Square Residual:	
SRMR	0.055

TABLE 4
Computed Model Fit Test Statistics and *P*-values for $N = 775$

Test statistic	T_{ML}	T_{SB}	T_{SS}	T_{YB}
p-value	0.000000052	0.000000064	0.00000016	0.0000000084
AD values	243.81	1.09	5.70	3.19

T_{ML} = maximum likelihood; T_{SB} = Satorra-Bentler; T_{SS} = Scaled and Shifted; T_{YB} = Yuan-Bentler. AD = Anderson-Darling distance metric.

departure from the rejection rates, although the T_{SB} test statistic appears to stray by the least amount.

CONCLUDING REMARKS

This paper discussed an algorithmic approach that can be used to approximately identify the best performing test statistic for the given data and model conditions examined. The approach can be used to evaluate a variety of sampled data exhibiting diverse distributional characteristics, sample size, and proposed model features. The approach was illustrated using empirical data obtained from a study of personality attributes. The results demonstrated the valuable capabilities of the procedure for automating the entire process of assessing model fit in structural equation modeling applications.

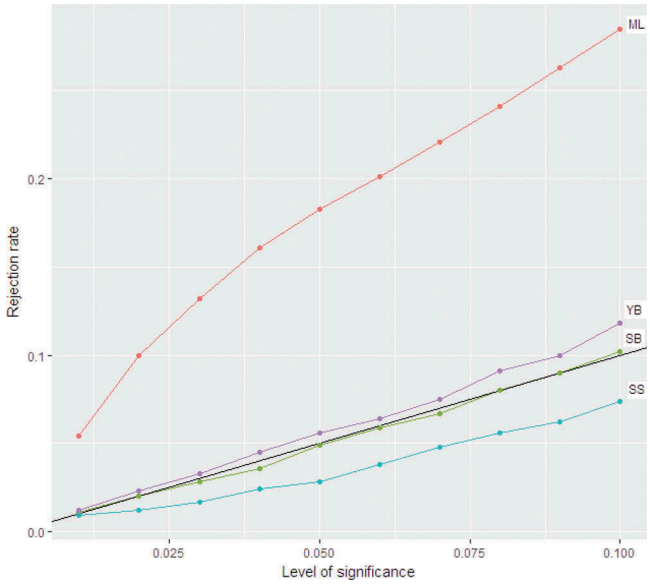

FIGURE 2 Plot of each test statistic against the nominal rejection rates for $n = 775$.
Note. ML = maximum likelihood; SB = Satorra-Bentler; SS = Scaled and Shifted; YB = Yuan-Bentler.

A major problem with many of the highly popular test statistics used when evaluating model fit in SEM is that it is not always apparent which should be preferred. In such instances, selecting the wrong one can have adverse consequences on assessing model fit. The approach introduced in this paper is simply an alternative tool that can help researchers better evaluate overall model fit. Once a proposed theoretical model is formulated it can be objectively tested against observed data so that the most reliable test statistic can be identified. Although the approach was illustrated using a simple factor analytic model, its application can be readily generalized to all types of modeling situations, including studies with multiple group settings.

The approach can also be extended beyond the four test statistics examined in this article. Many choices exist for assessing the fit of a proposed model in structural equation modeling applications and the bootstrap selection mechanism is a valuable approach that will help researchers with this difficult task. Using this approach, we feel that researchers can objectively evaluate model fit without forcing them to select a test statistic based on assumptions alone. At the same time, we acknowledge that more work needs to be done regarding different candidate test statistics and selection criteria, and look forward to continued research in this area.

REFERENCES

Amemiya, Y., & Anderson, T. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, *18*, 1453–1463. doi:10.1214/aos/1176347760

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, *23*, 193–212. doi:10.1214/aoms/1177729437

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49*, 765–769. doi:10.1080/01621459.1954.10501232

Arshad, M., Rasool, M. T., & Ahmad, M. I. (2003). Anderson Darling and modified Anderson Darling tests for generalized Pareto distributions. *Pakistan Journal of Applied Sciences*, *3*, 85–88. doi:10.3923/jas.2003.85.88

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction (Unpublished manuscript). Retrieved from www.statmodel.com/download/WLSMV_new_chi21.pdf

Babu, G. J., & Rao, C. R. (2004). Goodness-of-fit test when parameters are estimated. *Sankhya*, *66*, 63–74.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*, 77–85. doi:10.1111/j.2044-8317.1950.tb00285.x

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

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

Beran, R., & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, *13*, 95–115. doi:10.1214/aos/1176346579

Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness of fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.

Bollen, K. A., & Stine, R. (1992). Bootstrapping goodness of fit measures in structural equation models. *Sociological Methods & Research*, *21*, 205–209. doi:10.1177/0049124192021002004

D'Agostino, R., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York, NY: Marcel Decker.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). SEM with many variables: Issues and developments. *Quantitative Psychology and Measurement Section, Frontiers in Psychology*. doi:10.3389/fpsyg.2018.00580

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26. doi:10.1214/aos/1176344552

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

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

Foldnes, N., & Grønneberg, S. (2018). Approximating test statistics using eigenvalue block averaging. *Structural Equation Modeling*, *25*, 101–114. doi:10.1080/10705511.2017.1373021

Foldnes, N., & Olsson, U. H. (2015). Correcting too much or too little? The performance of three chi-square corrections. *Multivariate Behavioral Research*, *50*, 533–543. doi:10.1080/00273171.2015.1036964

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356–410. doi:10.1207/S15328007SEM0703_2

Grønneberg, S., & Foldnes, N. (2019). Testing Model Fit by Bootstrap Selection. *Structural Equation Modeling*, *26*, 182–190. doi:10.1080/10705511.2018.1503543

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

Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method (2nd ed.). London: Butterworths.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402. doi:10.1080/01621459.1967.10482916

Marcoulides, G. A. (1990). Evaluation of confirmatory factor analytic and structural equation models using goodness-of-fit indices. *Psychological Reports*, *67*, 669–670. doi:10.2466/PR0.67.6.669-670

Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling*, *25*, 484–491. doi:10.1080/10705511.2017.1409074

Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling*, *24*, 148–153. doi:10.1080/10705511.2016.1225260

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. doi:10.1037/0033-2909.105.1.156

Muthén, L. K., & Muthén, B. O. (2018). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*, 439–478. doi:10.1207/S15327906MBR3903_3

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Razali, N. W., & Wah, Y. B. (2011). Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*, 22–33.

Revelle, W. (2015b) Package "psych." Retrieved from http://org/r/psych-manual.pdf

Rosseel, Y. (2012). *Lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. doi:10.18637/jss.v048.i02

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

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

Steiger, J. H., & Lind, J. C. (1980, May 30). *Statistically-based tests for the number of common factors*. Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City.

Stephens, M. A. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of the American Statistical Association*, *69*, 730–737. doi:10.1080/01621459.1974.10480196

Wu, H. (2018). Approximations to the distribution of a test statistic in covariance structure analysis: A comprehensive study. *British Journal of Mathematical and Statistical Psychology*, *17*, 334–362. doi:10.1111/bmsp.12123

Wu, H., & Lin, J. (2016). A scaled F distribution as an approximation to the distribution of test statistics in covariance structure analysis. *Structural Equation Modeling*, *23*, 409–421. doi:10.1080/10705511.2015.1057733

Yang, M., Jiang, G., & Yuan, K. H. (2018). The performance of ten modified rescaled statistics as the number of variables increases. *Structural Equation Modeling*, *25*, 414–438. doi:10.1080/10705511.2017.1389612

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*, 175–183. doi:10.1080/10629360600678310

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. doi:10.1111/0081-1750.00078

Yuan, K. H., Yang, M., & Jiang, G. (2017). Empirically corrected rescaled statistics for SEM with small N and large p. *Multivariate Behavioral Research*, *52*, 673–698. doi:10.1080/00273171.2017.1354759

Yuan, K.-H., & Bentler, P. M. (2005). Asymptotic robustness of the normal theory likelihood ratio statistic for two-level covariance structure models. *Journal of Multivariate Analysis*, *94*, 328–343.

Yung, Y. F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ: Lawrence Erlbaum Associates.

APPENDIX A

R Source Code for Model Fitting Using the Illustrative Data Examples

```
# Procedure for the Anderson-Darling test.
rm(list=ls())
library(psych)
library(reshape2)
library(lavaan)
library(ggplot2)
library(goftest)

# Only candidate statistics already implemented in lavaan are used.

data(bfi)# Illustrative BFI data - using a random samples from the dataset.
set.seed(1)
orig.sample = bfi[ sample(1:nrow(bfi), size=175, replace=F),
                c(paste0("A", 1:5), paste0("C",1:5))]
orig.sample <- orig.sample[ complete.cases(orig.sample),]

model ="A =~ A1+A2+A3+A4+A5; C =~ C1+C2+C3+C4+C5"

## obtain p-values from original sample.
# 1. ML - Maximum Likelihood.
f=sem(model, data=orig.sample, test="standard")
pml.orig <- fitmeasures(f,"pvalue")

# 2 SB - Satorra-Bentler.
f=sem(model, data=orig.sample, test="Satorra.Bentler")
psb.orig <- fitmeasures(f,"pvalue.scaled")

# 3 SS Scaled-and-Shifted (asparouhov/muthen 2010).
f=sem(model, data=orig.sample, test="scaled.shifted")
pss.orig <- fitmeasures(f,"pvalue.scaled")

# 4 YB - Yuan-Bentler
f=sem(model, data=orig.sample, test="Yuan.Bentler")
pyb.orig <- fitmeasures(f,"pvalue.scaled")

cat("Which of these p-values is most trust worthy?\n ML: ",
     round(pml.orig, 3), "SB: ",
     round(psb.orig,3), "SS: ", round(pss.orig,3),
    "YB: ", round(pyb.orig,3), "\n")
#####
# Transform the sample data.
sigma.hat <- lavInspect(f, "sigma.hat")
sigma.sqrt <- lav_matrix_symmetric_sqrt(sigma.hat)
s.inv.sqrt <- lav_matrix_symmetric_sqrt(f@SampleStats@icov[[1]])
sample.transformed <- data.frame(as.matrix(orig.sample) %*% s.inv.sqrt %*% sigma.sqrt)
colnames(sample.transformed) <- colnames(orig.sample)
#####
#worker function. Easy to put in parallel
run.bootstrap <- function(X, b.reps, sample.transformed){
```

```
  set.seed(X)
  pb <- txtProgressBar(min = 0, max = b.reps, style = 3)
  pml=NULL; psb=NULL; pss=NULL; pyb=NULL
  for(b in 1:b.reps){
    setTxtProgressBar(pb, b)
    boot.sample = sample.transformed[ sample(1:nrow(sample.transformed), replace=T),]
    f.sb = tryCatch(sem(model,boot.sample, test="Satorra.Bentler", start=f), error=
      function (w) { TRUE} )
    f.ss = tryCatch(sem(model, boot.sample, test="scaled.shifted", start=f.sb),error=
      function (w) { TRUE} )
    f.yb = tryCatch(sem(model,boot.sample, test="Yuan.Bentler", start=f.sb), error=
      function(w) { TRUE} )

    while(isTRUE(f.sb) | isTRUE(f.ss) | isTRUE(f.yb)){
      boot.sample = sample.transformed[ sample(1:nrow(sample.transformed), replace=T),]
      f.sb = tryCatch(sem(model,boot.sample, test="Satorra.Bentler", start=f), error=
        function(w) { TRUE} )
      f.ss = tryCatch(sem(model, boot.sample, test="scaled.shifted", start=f.sb),error=
        function(w) { TRUE} )
      f.yb = tryCatch(sem(model,boot.sample, test="Yuan.Bentler", start=f.sb), error=
        function(w) { TRUE} )
    }

    pml <- c(pml,tryCatch(fitmeasures(f.sb,"pvalue"), error=function(w) { TRUE} ))
    psb <- c(psb,tryCatch(fitmeasures(f.sb,"pvalue.scaled"), error=function(w) { TRUE} ))
    pss <- c(pss,tryCatch(fitmeasures(f.ss,"pvalue.scaled"), error=function(w) { TRUE} ))
    pyb <- c(pyb,tryCatch(fitmeasures(f.yb,"pvalue.scaled"), error=function(w) { TRUE} ))
  }
  return(data.frame(pml,psb,pss, pyb))
}

## bootstrap reps
b.reps=1000
## run bootstrap
res.df <- run.bootstrap(X=1, b.reps=b.reps, sample.transformed=sample.transformed)

colnames(res.df)<- c("ML", "SB", "SS", "YB")
rownames(res.df) = NULL
melted = melt(res.df)
melted$statistic <- melted$variable

#plot of the p-values
ggplot(melted, aes(value, fill=statistic))+geom_histogram()+facet_wrap(~statistic)

#Anderson-Darling distance metric.
distUniform= as.vector(sapply(res.df,function(x) ad.test(x, null="punif")$statistic))
distUniform

cat("Anderson-Darling test selects: ", colnames(res.df)[ which.min(distUniform)] )
# Look at the rejection rates for significance levels between 0.01 and 0.1:
x=seq(0.01, 0.1, length.out=10)
y <- t(sapply(x, function(x) colMeans(res.df<x)))
m <- melt(data.frame(cbind(x, y)), id.vars="x"); m$test <- m$variable
ggplot(m, aes(x, value, color=test))+geom_line()+geom_abline(slope=1, intercept=0)
```

```
+ylab("Rejection rate")+
xlab("Level of significance")+geom_point()
```

Note. A symbol # denotes a comment.