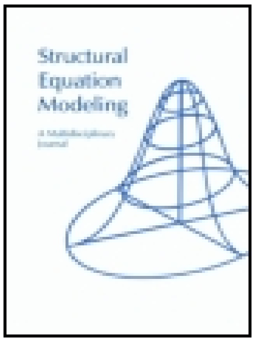

Structural Equation Modeling: A Multidisciplinary Journal

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

Approximating Test Statistics Using Eigenvalue Block Averaging

Njål Foldnes & Steffen Grønneberg

To cite this article: Njål Foldnes & Steffen Grønneberg (2017): Approximating Test Statistics Using Eigenvalue Block Averaging, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2017.1373021

To link to this article: http://dx.doi.org/10.1080/10705511.2017.1373021

Published online: 20 Oct 2017.

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–14, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2017.1373021

Approximating Test Statistics Using Eigenvalue Block Averaging

Njål Foldnes and Steffen Grønneberg
BI Norwegian Business School

We introduce and evaluate a new class of approximations to common test statistics in structural equation modeling. Such test statistics asymptotically follow the distribution of a weighted sum of i.i.d. chi-square variates, where the weights are eigenvalues of a certain matrix. The proposed eigenvalue block averaging (EBA) method involves creating blocks of these eigenvalues and replacing them within each block with the block average. The Satorra–Bentler scaling procedure is a special case of this framework, using one single block. The proposed procedure applies also to difference testing among nested models. We investigate the EBA procedure both theoretically in the asymptotic case, and with simulation studies for the finite-sample case, under both maximum likelihood and diagonally weighted least squares estimation. Comparison is made with 3 established approximations: Satorra–Bentler, the scaled and shifted, and the scaled F tests.

Keywords: fit statistics, nonnormal data, Satorra–Bentler, structural equation modeling

In general, test statistics for moment structural models converge in law to the distribution of a weighted sum of independent chi-squares, under the null hypothesis of correct model specification. More precisely, a test statistic T_n based on n observations will obey (Satorra, 1989; Shapiro, 1983)

$$T_n \underset{n\to\infty}{\overset{D}{\to}} \sum_{j=1}^{d} \lambda_j Z_j^2, \qquad Z_1, \ldots, Z_d \sim N(0,1)\text{IID}, \tag{1}$$

where the weights $\lambda = (\lambda_1, \ldots, \lambda_d)'$ are the nonzero eigenvalues of an unknown population matrix. Under optimal conditions, in which the estimator is correctly specified for the data at hand, or under conditions of so-called asymptotic robustness (e.g., Browne & Shapiro, 1988; Shapiro, 1987), the weights λ_j are all equal to one, and T_n converges to a chi-square distribution. However, in most cases the weights are not equal to one, and T_n should not be referred to a nominal chi-square distribution.

One approach to this problem is to construct a distribution that approximates the distribution of the weighted sum in Equation 1, and refer T_n to this approximating distribution. That is, using characteristics of the data and the model, a distribution is constructed that tries to emulate the distribution of $\sum \lambda_j Z_j^2$. Let X_{approx} be a random variable that follows this approximating distribution. Then the p value of the test of correct model specification is obtained as $P(X_{\text{approx}} > T_n)$, where T_n is considered fixed and the probability is with respect to X_{approx}. For instance, the scaling of Satorra and Bentler (1988) approximates the weighted sum in Equation 1 by setting all the weights equal to the average $\bar{\lambda} = \sum_{j=1}^{d} \hat{\lambda}_j / d$ of the estimated eigenvalues. That is, $X_{\text{approx}} = \sum_j \bar{\lambda} Z_j^2$, with p value $P\left(\sum_j \bar{\lambda} Z_j^2 > T_n\right)$, which can be recasted in the more familiar form $P\left(\chi_d^2 > T_n/\bar{\lambda}\right)$. Other recently proposed approximations to the distribution in Equation 1 are the scaled F distribution (Wu & Lin, 2016) and the scaled and shifted χ_d^2 (Asparouhov & Muthén, 2010). The scaled-and-shifted test statistic is closely related to the Sattertwaithe type test statistic proposed by Satorra and Bentler (1994), and these two statistics have been reported to have similar performance (Foldnes & Olsson, 2015).

If λ was known, Equation 1 motivates the "oracle" p value

Correspondence should be addressed to Njål Foldnes, Department of Economics, BI Norwegian Business School, Oslo 0484, Norway. E-mail: njal.foldnes@bi.no

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

$$p_n = P\left(\sum_{j=1}^{d} \lambda_j Z_j^2 > T_n\right), \qquad (2)$$

which would yield an asymptotically valid test of model fit. In a practical setting λ is unfortunately unknown, but consistent estimates $\hat{\lambda}$ can be obtained. This suggests the approximation $X_{approx} = \sum_j \hat{\lambda}_j Z_j^2$ and the associated p value

$$\hat{p}_n = P\left(\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2 > T_n\right). \qquad (3)$$

However, this consistency might come at a price, given the variability of the $\hat{\lambda}$. In practice, it might be better to replace the $\hat{\lambda}$ in Equation 3 with more stable weights $\tilde{\lambda}$, obtained through grouping the $\hat{\lambda}$ by magnitude in blocks and calculating block averages. We refer to this method as eigenvalue block averaging (EBA). As there are many ways to form blocks, the EBA method yields many new approximations to the limiting distribution in Equation 1.

Although the EBA idea is simple, to the best of our knowledge it has not been discussed before. However, Wu and Lin (2016) investigated the full eigenvalue approximation in Equation 3, which technically is an EBA procedure with singleton blocks. Also, at the other extreme, EBA with one single block is identical to the well-known Satorra–Bentler (SB) scaling procedure. We are not aware of any literature on EBA approximations between these two extremes. The goal of this article is to present the EBA framework, and to evaluate EBA tests both asymptotically and in finite samples, by comparing EBA to three established test statistics for structural equation models.

This article is organized as follows. First, we review the literature on test statistics for moment structural moments, followed by a section formally introducing the EBA tests. We then illustrate the established and proposed tests with a real-world example, followed by asymptotic and finite-sample evaluations of the tests for single and nested model testing. The final section contains discussion and concluding remarks.

TEST STATISTICS

A structural equation model implies a parametrization $\theta \mapsto \sigma(\theta)$, where the free parameters in the proposed model are contained in the q vector θ. The model has degrees of freedom given by $d = p^* - q$, where p^* denotes the dimension of $\sigma(\theta)$. In covariance structure models $\sigma(\theta)$ consists of second-order moments, but in more general structural equation models the means can also be included in $\sigma(\theta)$. The corresponding sample moment vector s is assumed to converge in probability to $\sigma_0 = \sigma(\theta_0)$, and be asymptotically normal; that is, $\sqrt{n}(s - \sigma_0) \underset{n\to\infty}{\overset{D}{\rightarrow}} N(0, \Gamma)$. Here Γ is the asymptotic covariance matrix of $\sqrt{n}s$. A very general class of estimators for θ_0 introduced by Browne (1982, 1984) is obtained by minimizing discrepancy functions $F = F(s, \sigma)$ that obey the following three conditions: $F(s, \sigma) \geq 0$ for all s, σ; $F(s, \sigma) = 0$ if and only if $s = \sigma$; and F is twice continuously differentiable jointly. That is, we consider estimators obtained as

$$\hat{\theta} = \arg \min_{\theta} \; F(s, \sigma(\theta)).$$

It is well known that the widely used normal-theory maximum likelihood (ML) estimator is such a minimal discrepancy estimator.

Minimum discrepancy estimation leads to the fit statistic $T_n = nF(s, \sigma(\hat{\theta}))$, which is asymptotically equivalent to several other tests for model fit (Satorra, 1989). Correct model specification and other assumptions (Shapiro, 1983) imply the convergence in Equation 1. The weights $\lambda_1, \ldots, \lambda_d$ are the nonzero eigenvalues of $U\Gamma$, where $U = V - V\Delta\{\Delta' V \Delta\}^{-1}\Delta' V$, Δ is the $p \times q$ derivative matrix $\partial\sigma(\theta)/\partial\theta'$ evaluated at θ_0, and $V = -\frac{1}{2}\frac{\partial^2 F(s,\sigma)}{\partial s \partial \sigma}$, evaluated at (σ_0, σ_0). Clearly, if all the λ are equal to one, then T_n converges to a chi-square distribution with d degrees of freedom, and we are in a so-called asymptotic robust situation. Conditions necessary for this have been characterized (e.g., Amemiya & Anderson, 1990; Browne & Shapiro, 1988; Mooijaart & Bentler, 1991; Satorra & Bentler, 1990; Shapiro, 1987). However, these conditions are hard to check in practice, and currently no practical procedure exists for verifying asymptotic robustness in a real-world setting (Yuan, 2005, p. 118).

The scaling procedure proposed by Satorra and Bentler (1988) is defined as $T_{\mathrm{SB}} = T_n/\hat{c}$, where $\hat{c} = \mathrm{trace}(\hat{U}\hat{\Gamma})/d$. Asymptotically T_{SB} converges to a distribution whose expectation equals d, the expectation of the nominal chi-square distribution. In conditions where all eigenvalues are equal, $\lambda_1 = \ldots = \lambda_d$, T_{SB} will converge in distribution to a chi-square distribution. Using T_{SB} as a test statistic is a widely used structural equation modeling (SEM) practice under conditions of nonnormal data. Simulation studies report that T_{SB} outperforms the ML fit statistic T_{ML} in such conditions, but that Type I error rates under T_{SB} might become inflated under substantial excess kurtosis in the data (Bentler & Yuan, 1999; Foldnes & Olsson, 2015; Nevitt & Hancock, 2004). Also, Yuan and Bentler (2010) demonstrated that T_{SB} departs from a chi-square with increasing dispersion of the eigenvalues $\lambda_j, j = 1, \ldots, d$.

Recently Asparouhov and Muthén (2010) proposed a test statistic that agrees with the reference chi-square distribution in both asymptotic mean and variance, obtained from T_{ML} by scaling and shifting (SS). This statistic is given by $T_{\mathrm{SS}} = a \cdot T_n + d - b$, where $a = \sqrt{d/\mathrm{trace}\left((\hat{U}\hat{\Gamma})^2\right)}$ and $b = \sqrt{d\left(\mathrm{trace}(\hat{U}\hat{\Gamma})\right)^2/\mathrm{trace}\left((\hat{U}\hat{\Gamma})^2\right)}$. In a simulation study, Foldnes and Olsson (2015) found that T_{SB} and T_{SS} tended to overreject and underreject, respectively, correctly specified models.

Very recently, Wu and Lin (2016) proposed a scaled F distribution that matches the mean, variance, and skewness of $\sum_{j=1}^{d} \hat{\lambda}_j Z_j^2$, where the $\hat{\lambda}_j$ are the eigenvalues of $\hat{U}\hat{\Gamma}$. The scaling, and the two degrees of freedom of the F distribution, are functions of $\sum_j \hat{\lambda}_j, \sum \hat{\lambda}_j^2$, and $\sum_j \hat{\lambda}_j^3$. In a simulation study, Wu and Lin (2016) found the scaled F test to perform similarly to the Sattertwaithe type test statistic proposed by Satorra and Bentler (1994).

EBA TEST STATISTICS

In this section we introduce new tests for model fit, based on the asymptotic result in Equation 1. The proposed methodology applies as long as the null distribution of a test statistic is a weighted sum of independent chi-squares and the weights can be estimated consistently. This means that the method could be used both for conventional goodness-of-fit testing of a single proposed model and for nested model comparison tests. Also, the tests could be applied in a context more general than the prototypical case of T_{ML}, for instance with diagonally weighted least squares (DWLS) estimation and model testing.

Note that the p value $\hat{p}_n$ in Equation 3 is theoretically optimal when the sample size goes to infinity. That is, because $\hat{\lambda}$ converges to λ in probability, the difference between $\hat{p}_n$ and the oracle p value p_n in Equation 2 goes to zero in probability, meaning that it has zero asymptotic bias. However, in situations with small sample sizes and highly nonnormal data, the estimates $\hat{\lambda}_j$ become unstable and highly variable. Because $\hat{p}_n$ directly employs each individual estimate $\hat{\lambda}_j$ it could inherit this instability, leading to poor finite-sample performance.

One established way of overcoming this instability is offered by the SB test. As previously discussed, this test estimates each λ_j by the grand average of all estimated eigenvalues. Clearly, unless all the population eigenvalues are identical, this method is inconsistent. However, the averaging process might result in less variability at the cost of some bias.

In this study our perspective is that of a bias–variance trade-off, in which the SB test and the full use of estimated eigenvalues in $\hat{p}_n$ are viewed as extreme end points on a spectrum. At one end of the spectrum, importance is given to stabilizing the eigenvalues, as done in the SB test. At the other end, importance is given to asymptotic bias. We propose intermediate solutions, referred to as EBA tests, between these two extremes. EBA testing involves grouping the $\hat{\lambda}_j$ in blocks by magnitude, and replacing them by group averages, as we formalize mathematically later. The resulting EBA tests might be viewed as middle grounds between the one-block EBA (the SB test) and the d-block EBA in Equation 3.

Consider first the following split-half approximation, where the lower half of the eigenvalues constitute one block, and are replaced by their mean value, and likewise for the block containing the upper half of the eigenvalues:

$$\hat{p}_{n,2} = P\left(\sum_{j=1}^{d} \tilde{\lambda}_j Z_j^2 > T_n\right),$$

where

$$\tilde{\lambda}_1 = \cdots = \tilde{\lambda}_{\lceil d/2 \rceil} = \frac{1}{\lceil d/2 \rceil} \sum_{j=1}^{\lceil d/2 \rceil} \hat{\lambda}_j, \quad \text{and}$$

$$\tilde{\lambda}_{\lceil d/2 \rceil + 1} = \cdots = \tilde{\lambda}_d = \frac{1}{d - \lceil d/2 \rceil} \sum_{j=\lceil d/2 \rceil + 1}^{d} \hat{\lambda}_j.$$

This procedure allows the p value approximation an additional degree of freedom compared to the SB statistic, where all eigenvalues are estimated to be equal to each other. In general, a class of middle grounds between one-block and d-block EBA can be defined as follows. Choose cut-off integers $1<\tau_1<\tau_2<\cdots<\tau_k<d$ with $1 \leq k<d$. Also let $\tau_0 = 1$. Then, for $\tau_{l-1} \leq r < \tau_l$ let

$$\tilde{\lambda}_r = \frac{1}{\tau_l - \tau_{l-1}} \sum_{\tau_{l-1} \leq j < \tau_l} \hat{\lambda}_j, \tag{4}$$

and for $\tau_k \leq r \leq d$,

$$\tilde{\lambda}_r = \frac{1}{d - \tau_k} \sum_{\tau_k \leq j \leq d} \hat{\lambda}_j.$$

Let us denote this choice by $\tilde{\lambda}(\tau) = (\tilde{\lambda}_1(\tau), \ldots, \tilde{\lambda}_d(\tau))'$. The proposed p-value estimator is then

$$\hat{p}_n(\tau) = P\left(\sum_{j=1}^{d} \tilde{\lambda}_j(\tau) Z_j^2 > T_n\right).$$

The cutoffs τ defining the blocks might appear with (approximately) equal distance, such that $\hat{p}_{n,3}$ is obtained from three (approximately) equally sized blocks, and $\hat{p}_{n,4}$ from four (approximately) equally sized blocks. For instance, with d = 35 and four blocks, the block sizes are 9, 9, 9, and 8. In this study, we investigated four EBA tests obtained from equally sized blocks: At one extreme is the one-block SB test, and at the other extreme is singleton blocks; that is, the full use of all estimated eigenvalues. We refer to this latter test as EBAF, whose p value is given by Equation 3. In between these two extremes we considered two midde-ground tests, namely the split-half, denoted by EBA2, and the use of four blocks, denoted by EBA4.

Instead of insisting that the blocks should have equal sizes, another strategy is to use a clustering algorithm. Such algorithms iteratively form blocks of eigenvalues of possibly unequal sizes to minimize the variability within each block while maximizing the between-block variance. They start with some set of blocks and then adjust these blocks iteratively to reduce the sum of squared deviations in each class. In this study we employed both the natural breaks classification of Jenks (1967), where the number of blocks is prespecified by the user, and a clustering method proposed by Wang and Song (2011) where the number of blocks is chosen by an optimization algorithm. The output of the Jenks algorithm is then the grouping of eigenvalues into the prespecified number of blocks. In this study we investigated the Jenks method with two or four blocks. Replacing the eigenvalues in each block by the block average yields the tests EBA2J and EBA4J, respectively. The output of the Wang and Song (2011) method is the grouping of eigenvalues into the optimal number of blocks. We denote by EBAA the test obtained by replacing eigenvalues in each block by the block average in each of the automatically chosen blocks.

An extension of this framework is tests that assess nested hypotheses in SEM. Due to its great practical importance, we include here a short discussion on this special case. Following Satorra (1989), let $H : \sigma = \sigma(\theta), \theta \in \Theta$ and $H_0 : \sigma = \sigma(\theta), \theta \in \Theta_0$ where $\Theta_0 = \{\theta \in \Theta : a(\theta) = 0\}$ for some continuously differentiable function a. We assume that the matrix $\frac{\partial a(\theta)}{\partial \theta}$ has full row rank, say m. We let

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, F(s, \sigma(\theta)), \qquad \tilde{\theta} = \underset{\theta \in \Theta_0}{\operatorname{argmin}}\, F(s, \sigma(\theta))$$

and $T_n = nF(s, \sigma(\hat{\theta}))$ and $\tilde{T}_n = nF(s, \sigma(\tilde{\theta}))$. Under H_0 and the conditions of Lemma 1 (iv) in Satorra (1989) the difference statistic converges as

$$\begin{gathered} \tilde{T}_n - T_n \xrightarrow[n \to \infty]{D} \sum_{j=1}^{m} \alpha_j Z_j^2, \\ Z_1, \ldots, Z_m \sim N(0,1)\mathrm{IID}, \end{gathered} \tag{5}$$

where $\alpha_1, \ldots, \alpha_m$ are the m nonzero eigenvalues of $U_d \Gamma$, where $U_d = \tilde{U} - U$ has rank m. Distribution-free consistent estimators $\hat{U}_d$ and $\hat{\Gamma}$ for U_d and Γ are found and discussed in Satorra and Bentler (2001).

In the next section we illustrate the EBA procedures with a real-world data sample, followed by a section where we evaluate, both asympotically and in finite samples, the performance of EBA. Probabilities of the type in Equation 3 were calculated using the R package CompQuadForm (Duchesne & De Micheaux, 2010), and model estimation and eigenvalue extraction were done with lavaan (Rosseel, 2012).

Block formation by clustering methods was done using the R packages BAMMtools (Rabosky et al., 2014) for the Jenks method and Ckmeans.1d.dp for the method of Wang and Song (2011). R code demonstrating the use of these packages can be found in the Appendix.

EXAMPLE

We consider data from a study (Foldnes, 2017) conducted among $n = 98$ students at a business school, where items from the shortened version of the Attitudes Toward Mathematics Inventory (Lim & Chapman, 2013) were used to model the correlation between enjoyment of mathematics (ENJ) and self-confidence (SC) in mathematics. The model is depicted in Figure 1, which has 13 degrees of freedom.

Two estimation methods, DWLS and ML, were considered, with test statistics $T_{\mathrm{DWLS}} = 7.90$ and $T_{\mathrm{ML}} = 25.26$. In each case we extracted the 13 estimated eigenvalues from $\hat{U}\hat{\Gamma}$. These are the weights used in EBAF, and are given in rows 1 and 9 of Table 1, which also contains the weights used by T_n, ML, SB, EBA2, EBA2J, EBA4, EBA4J, and EBAA. For both estimation methods, SB p values are smaller than the p values for the other robust tests, which is unsurprising, given the reported tendency of SB to overreject correct models under nonnormality (e.g., Foldnes & Olsson, 2015). Also, under DWLS, the automatic EBAA test yields only one cluster, so that EBAA in that condition coincided with SB, whereas under ML, EBAA has two clusters and is equivalent to EBA2. Overall, the p values vary moderately among the tests.

For the ML case, we also plotted the probability density function of X_{approx} for SB, SS (scaled-and-shifted), CF (scaled F), and three EBA tests in Figure 2. The p values associated with SS and CF were 0.223 and 0.195, respectively. We see that in this real-world situation, the distributions of CF and the three EBA tests are quite similar to each other. The SB and SS tests are seen to be based on distributions that differ quite a lot from those of the CF and the EBA tests.

In summary, Table 1 and Figure 2 indicate that there is some variability among the established and newly proposed tests. For a practitioner, the question remains about which of these tests should be used for evaluating the model. As shown in the next sections, there is unfortu-

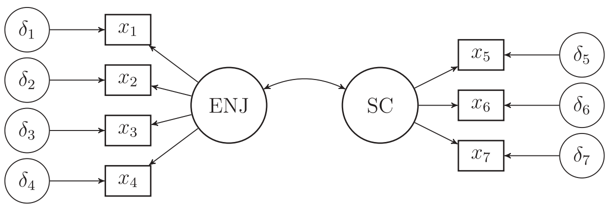

FIGURE 1 Modeling enjoyment of mathematics and self-confidence in mathematics.

TABLE 1
Estimated $\lambda_j, j = 1, \ldots, 13$, in First Row (EBAF), Together With $\tilde{\lambda}_j$ for Other Methods

	Method	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*p Value*
DWLS	EBAF	0.81	0.56	0.49	0.40	0.32	0.23	0.21	0.16	0.12	0.11	0.09	0.08	0.05	0.029
	T	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.850
	SB	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.009
	EBA2	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.10	0.10	0.10	0.10	0.10	0.10	0.019
	EBA4	0.56	0.56	0.56	0.56	0.26	0.26	0.26	0.13	0.13	0.13	0.08	0.08	0.08	0.025
	EBA2J	0.56	0.56	0.56	0.56	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.024
	EBA4J	0.81	0.48	0.48	0.48	0.26	0.26	0.26	0.10	0.10	0.10	0.10	0.10	0.10	0.028
	EBAA	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.009
ML	EBAF	5.46	2.38	2.01	1.52	1.40	1.12	1.08	0.95	0.67	0.61	0.53	0.42	0.36	0.193
	T	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.021
	SB	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	0.167
	EBA2	2.14	2.14	2.14	2.14	2.14	2.14	2.14	0.59	0.59	0.59	0.59	0.59	0.59	0.186
	EBA4	2.84	2.84	2.84	2.84	1.20	1.20	1.20	0.74	0.74	0.74	0.43	0.43	0.43	0.192
	EBA2J	5.46	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	0.181
	EBA4J	5.46	2.20	2.20	1.21	1.21	1.21	1.21	1.21	0.52	0.52	0.52	0.52	0.52	0.192
	EBAA	5.46	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	0.181

Note. DWLS = diagonally weighted least squares estimator; EBAF = full eigenvalue estimation; T = χ^2 test; SB = Satorra–Bentler; EBA*i* = *i*-block equal-size eigenvalue blocks; EBA*i*J = *i*-block Jenks eigenvalue blocks; EBAA = automatic eigenvalue clustering; ML = maximum likelihood estimator.

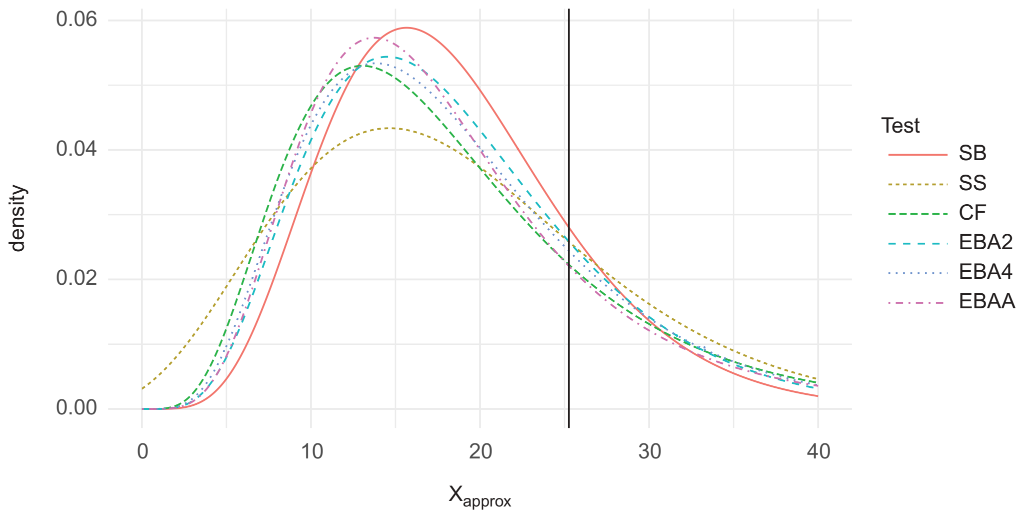

FIGURE 2 Probability density curves of X_{approx} for the case of testing a two-factor model based on n = 98 observations with the maximum likelihood (ML) estimator. Vertical line represents $T_{\text{ML}} = 25.26$. The areas below curves to the right of this line correspond to p values. SB = Satorra–Bentler; SS = scaled and shifted; CF = scaled F; EBA2 and EBA4 = eigenvalue block approximation with two and four equally sized blocks; EBAA = automatic eigenvalue clustering.

nately no single robust test that performs best under all possible conditions of sample size and underlying distribution. A possible way to select a test in a given situation is to simulate data with a distribution that is close to that of the observed data. The flexible data generating method recently proposed by Grønneberg and Foldnes (2017) could be used to emulate the characteristics of the observed sample. One can then observe which of the test candidates performs best on average on the simulated data. However, we consider this idea outside the scope of this study.

In the next section we proceed by evaluating the performance of the EBA tests and the established robust tests by Monte Carlo to gain some insight into the systematic differences with respect to empirical Type I error control.

METHOD

The performance of six EBA procedures and three established test statistics was assessed, both theoretically and empirically. The EBA procedures investigated are EBAF, EBA2, EBA4, EBA2J, EBA4J, and EBAA, and the established test statistics are SB, SS, and CF. In addition we included the oracle test in Equation 2, here denoted by OR. These test procedures are not specifically linked to ML estimation and its associated test statistic T_{ML}. In each evaluation case we therefore included a second estimator, namely DWLS, with its associated test statistic T_{DWLS}.

Theoretically, asymptotic rejection rates were computed based on eigenvalues extracted from the population matrix $U\Gamma$. This is possible due to a recently proposed method (Foldnes & Grønneberg, 2017) that allows the exact calculation of Γ, and consequently, of λ_j. For the EBA tests we solved the equation $P\left(\sum \lambda_j Z_j^2 > c\right) = 0.05$ numerically for c, and then the asymptotic rejection rate was calculated as $P\left(\sum \tilde{\lambda}_j Z_j^2 > c\right)$, where the $\tilde{\lambda}_j$ depends on the block-formation strategy.

Empirically, we conducted two simulation studies. Study 1 involves the testing of a single correctly specified model, and Study 2 involves the testing of two correctly specified nested models. The asymptotic and empirical rejection rates reported in this article were computed at the $\alpha = 0.05$ level of significance.

Models

Our model is the political democracy model discussed by Bollen (1989) in his textbook (see Figure 3), where the residual errors are not depicted for ease of presentation. There are four measures of political democracy measured twice (in 1960 and 1965), and three measures of industrialization measured once (in 1960). The model, referred to as $\mathcal{M}_1$, has $d = 35$ degrees of freedom. Study 1 involves tests of correct model specification based on $\mathcal{M}_1$. In Study 2, we considered testing a constrained model $\mathcal{M}_0$ against $\mathcal{M}_1$. $\mathcal{M}_0(d = 45)$ is nested within $\mathcal{M}_1$, and imposes 10 correctly specified equalities on unique and residual covariances.

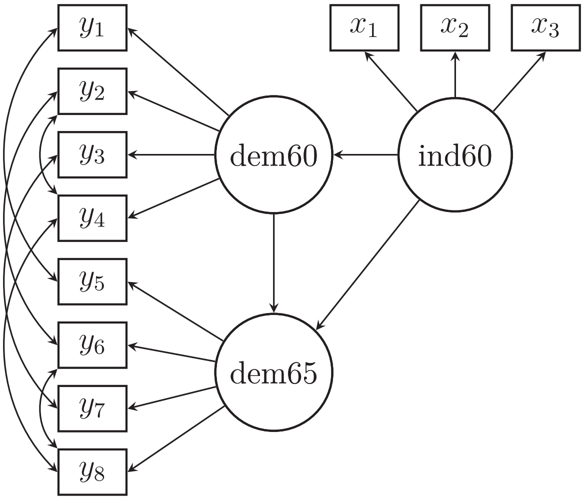

FIGURE 3 Bollen's political democracy model.

Data Generation

To theoretically evaluate the performance of the test statistics, and to evaluate the finite-sample performance of the oracle OR, the population values λ in Equation 1 must be exactly calculated. Recently, Foldnes and Grønneberg (2017) presented an algorithm for obtaining Γ under distributions produced by the Vale–Maurelli (VM) transform (Vale & Maurelli, 1983). We therefore used the VM transform in this study. We calculated Γ and U (for both ML and DWLS) and obtained population eigenvalues λ under each distributional condition. Data generation was achieved by fixing the parameters in the model, and using the model-implied covariance matrix as the target covariance matrix for the VM transform. Two nonnormal distributional conditions, denoted by D_1 and D_2, were specified by vectors containing heterogeneous skewness s and kurtosis k for the 11 univariate marginals as follows. For D_1, $s = (1,1,1,1,1,2,2,2,2,2,2)$ and $k = (5,5,5,5,5,10,10,10,10,10,10)$. For D_2, $s = (2,2,2,2,2,3,3,3,3,3,3)$ and $k = (7,7,7,7,7,21,21,21,21,21,21)$. With the terminology used by Curran, West, and Finch (1996), distributions D_1 and D_2 might be said to represent moderate and severe nonnormality, respectively. For the simulation studies, replications leading to nonconvergence or improper solutions were removed from further analysis. In each cell we simulated 10^4 replications with proper solutions, resulting in a standard error of 0.0022 for the empirical rejection rate, given that the true Type I error rate was 0.05.

RESULTS

Asymptotic Performance

Study 1

In each of six conditions (two estimators × three distributions), the 35 nonzero population eigenvalues were calculated. In Figure 4, violin plots give the distribution of these eigenvalues in each condition. With estimator ML, eigenvalues tend to get larger and span a larger range when moving from normality (N) via moderate nonnormality (D_1) to severe nonnormality (D_2). With estimator DWLS, the eigenvalues are not much affected by the underlying distribution, indicating that the large-sample distribution of T_{DWLS} is not sensitive to the underlying distribution.

Population eigenvalues were then used to compute asymptotic rejection rates for each test statistic (see Table 2). Because T_{DWLS} is not distributed as a chi-square under any distribution, T_{DWLS} rejection rates are far off the nominal level. T_{ML} is correctly specified for normal data, and hence has the optimal rejection rate of 0.05 under N, but

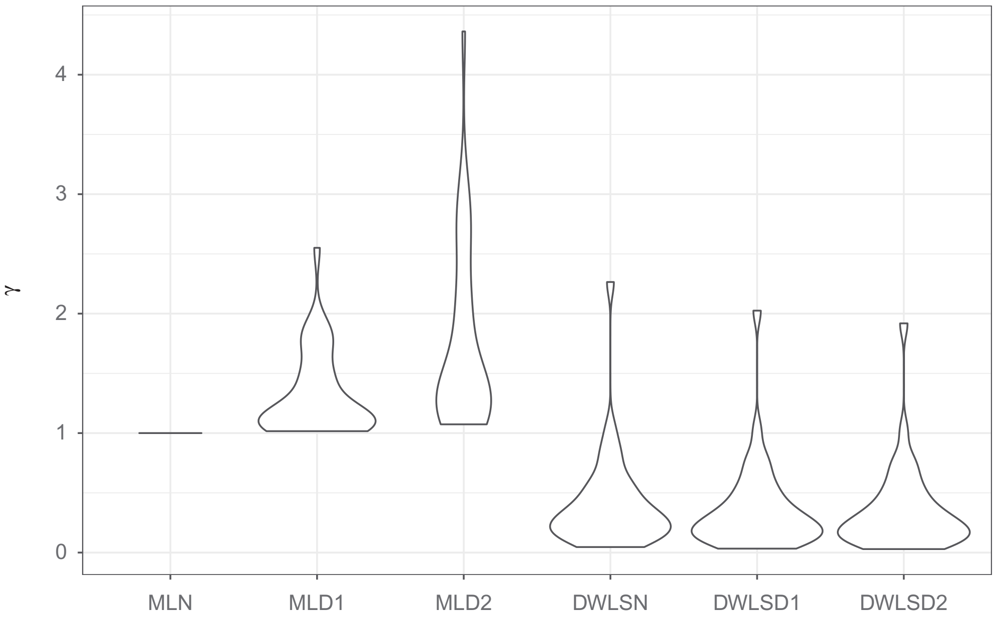

FIGURE 4 Study 1: Violin plots for the distribution of 35 population eigenvalues. MLN, MLD1, and MLD2 refer to maximum likelihood (ML) estimation under multivariate normality, and moderate and severe nonnormality, respectively. DWLSN, DWLSD1, and DWLSD2 refer to diagonally weighted least squares (DWLS) estimation under multivariate normality, and moderate and severe nonnormality, respectively.

TABLE 2
Study 1: Asymptotic Rejection Rates

	Dist	*T*	*SB*	*SS*	*CF*	*EBAF*	*EBA2*	*EBA4*	*EBA2J*	*EBA4J*	*EBAA*
DWLS	N	0.000	0.103	0.055	0.049	0.050	0.025	0.035	0.039	0.049	0.047
	D1	0.000	0.106	0.055	0.050	0.050	0.025	0.035	0.037	0.048	0.037
	D2	0.000	0.106	0.055	0.050	0.050	0.026	0.036	0.037	0.048	0.037
ML	N	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	D1	0.347	0.056	0.051	0.050	0.050	0.047	0.049	0.049	0.050	0.050
	D2	0.748	0.065	0.052	0.050	0.050	0.043	0.047	0.047	0.050	0.048

Note. T =χ^2 test; SB = Satorra–Bentler; SS = scaled and shifted; CF = scaled *F* test; EBAF = full eigenvalue estimation; EBA*i* = *i*-block equal-size eigenvalue blocks; EBA*i*J = *i*-block Jenks eigenvalue blocks; EBAA = automatic eigenvalue clustering; DWLS = diagonally weighted least squares estimator; N = normal; D_1 = moderate nonnormality; D_2= severe nonnormality; ML = maximum likelihood.

has highly inflated rejection rates under nonnormality. The SB scaling yields high rejection rates under DWLS, but close to nominal rates under ML, even with highly nonnormal data. The SS test performs much better than SB with DWLS, and is also preferrable to SB under ML. CF reaches almost perfect rejection rates under both DWLS and ML. The increasing asymptotic performance in the sequence SB, SS, and CF reflects that SB only matches the first, SS the two first, and CF the three first moments of the weighted sum in Equation 1. The full eigenvalue approximation EBAF yields perfect asymptotic Type I error control in all conditions, which is in line with the theory, as EBAF is a consistent test. The other EBA methods generally lead to underrejection, especially under DWLS, with EBA2 performing the worst, whereas EBA4J attains almost perfect Type I error control.

Study 2

The chi-square difference test has 10 degrees of freedom, and the corresponding oracle eigenvalues are presented in Table 3. Similar to the pattern in Figure 4, the eigenvalues are much more sensitive to the underlying distribution under the ML estimator compared to the DWLS estimator. With ML, the eigenvalues become larger and more varied with increasing nonnormality.

Table 4 contains asymptotic rejection rates for nested model testing. SB overrejects in all conditons except for ML under normality. SS overrejects very slightly, and the *F* test achieves perfect rejection rates. The EBA approximations tend to underreject the null, but less so compared to Study 1, with EBA4J achieving almost perfect Type I error control in all conditions.

TABLE 3
Study 2: The Population Eigenvalues of $U_d\Gamma$, Rounded to Three Decimal Places

DWLS	N	0.620	0.364	0.315	0.276	0.253	0.234	0.191	0.183	0.128	0.073
	D_1	0.503	0.326	0.288	0.244	0.184	0.174	0.157	0.133	0.102	0.058
	D_2	0.511	0.340	0.294	0.246	0.181	0.175	0.151	0.125	0.099	0.055
ML	N	1	1	1	1	1	1	1	1	1	1
	D_1	5.854	4.090	2.875	2.679	2.436	2.275	2.133	1.865	1.755	1.490
	D_2	11.280	7.607	4.724	3.990	3.702	3.564	3.276	2.784	2.629	2.093

Note. DWLS =diagonally weighted least squares estimator; N = normal; D_1= moderate nonnormality; D_2= severe nonnormality; ML = normal-theory maximum likelihood estimator.

TABLE 4
Study 2: Asymptotic Rejection Rates

	Dist	*T*	*SB*	*SS*	*CF*	*EBAF*	*EBA2*	*EBA4*	*EBA2J*	*EBA4J*	*EBAA*
DWLS	N	0.000	0.068	0.052	0.050	0.050	0.043	0.046	0.045	0.049	0.033
	D_1	0.000	0.070	0.052	0.050	0.050	0.043	0.047	0.045	0.049	0.032
	D_2	0.000	0.071	0.052	0.050	0.050	0.043	0.047	0.045	0.049	0.031
ML	N	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	D_1	0.732	0.063	0.051	0.050	0.050	0.044	0.047	0.048	0.050	0.048
	D_2	0.928	0.070	0.052	0.050	0.050	0.040	0.045	0.048	0.050	0.048

Note. T = χ^2 test; SB = Satorra–Bentler; SS = scaled and shifted; CF = scaled *F* test; EBAF = full eigenvalue approximation; EBA*i* = *i*-block eigenvalue approximation; EBA*i*J = *i*-block Jenks eigenvalue approximation; EBAA = automatic eigenvalue clustering; DWLS = diagonally weighted least squares estimator; N = normal; D_1= moderate nonnormality; D_2= severe nonnormality. ML = normal-theory maximum likelihood estimator.

Finite-Sample Performance

Study 1

Finite-sample rejection rates for testing $\mathcal{M}_1$ are given in Table 5. We discuss the DWLS case first, where, generally, all test statistics are quite robust to the underlying distribution. SB produces consistently high error rates, at about 0.1. SS error rates are consistently below the nominal level $\alpha = 0.05$, but approaches α with increasing sample size. Under conditions of small sample size and nonnormality, SS has rejections rates below 0.03. CF rejection rates are close to those of SS, but are consistently lower. EBAF

TABLE 5
Study 1: Rejection Rates

	n	*Distr*	*T*	*SB*	*SS*	*CF*	*EBAF*	*EBA2*	*EBA4*	*EBA2J*	*EBA4J*	*EBAA*	*OR*
DWLS	100	N	.000	.104	.039	.035	.034	.064	.047	.044	.036	.042	.059
		D_1	.000	.109	.030	.025	.024	.064	.041	.034	.026	.030	.099
		D_2	.000	.110	.023	.020	.019	.057	.032	.025	.020	.021	.133
	300	N	.000	.107	.049	.044	.044	.072	.057	.054	.046	.052	.057
		D_1	.000	.113	.045	.039	.039	.073	.055	.051	.041	.046	.082
		D_2	.000	.118	.039	.033	.032	.071	.049	.043	.034	.039	.106
	1,000	N	.000	.099	.050	.045	.045	.071	.060	.058	.046	.054	.049
		D_1	.000	.107	.050	.045	.045	.074	.061	.058	.047	.055	.064
		D_2	.000	.117	.048	.042	.042	.076	.057	.055	.044	.052	.077
ML	100	N	.078	.085	.044	.039	.039	.053	.044	.048	.040	.062	.078
		D_1	.277	.105	.021	.017	.016	.053	.029	.024	.018	.022	.041
		D_2	.516	.125	.014	.010	.009	.055	.026	.016	.010	.013	.018
	300	N	.059	.062	.046	.044	.044	.048	.045	.048	.045	.062	.059
		D_1	.317	.072	.024	.020	.020	.044	.031	.026	.022	.025	.051
		D_2	.617	.081	.015	.011	.011	.042	.024	.017	.012	.015	.039
	1,000	N	.049	.050	.046	.045	.045	.047	.045	.047	.045	.050	.049
		D_1	.324	.059	.031	.028	.028	.043	.036	.034	.029	.034	.050
		D_2	.689	.066	.023	.019	.019	.043	.031	.025	.019	.024	.046

Note. T = χ^2 test; SB = Satorra–Bentler; SS = scaled and shifted; CF = scaled *F*; EBAF = full eigenvalue approximation; EBA*i* = *i*-block clustering; EBA*i*J = *i*-block Jenks clustering; EBAA = automatic clustering; OR = oracle; DWLS = diagonally weighted least squares; N = normality; D_1= moderate nonnormality; D_2= severe nonnormality; ML = maximum likelihood.

and CF have almost identical rejection rates across all conditions. In contrast to SS, CF, and EBAF, the two-block EBA2 consistently has rejection rates well above α, and has poor Type I error control. EBA4 performs better than EBA2, with rejection rates lying generally between those of SS, CF, and EBAF on one hand, and EBA2 on the other hand. EBA2J lies slightly below EBA4 in terms of rejection rates, and EBA4J performs quite poorly with rejection rates below those of SS. EBAA has higher rejection rates than SS, and lower than those of EBA2J. To sum up, the procedures with best Type I error control across all DWLS conditions are SS, EBA4, EBA2J, and EBAA.

Next we consider ML estimation, where the test statistics are more sensitive to the underlying distribution than was the case for DWLS. Note that T yields exactly the same rejection rates as the oracle OR, under multivariate normal data N. Under nonnormality, however, rejection rates of T become very large. SB again has inflated rejection rates, especially under nonnormality and small sample size. SS has too low rejection rates, with especially poor performance under nonnormality. Again, CF and EBAF have near identical rejection rates across all conditions, slightly below those of SS. EBA2 has very good Type I error control in all conditions. EBA4 outperforms SS, CF, and EBAF, but still has poorer error control than EBA2. Under nonnormality, the clustering procedures EBA2J, EBA4J, and EBAA have rejection rates lower than those of EBA4. To sum up, across all ML conditions, EBA2 by far had the best Type I error control, with EBA4 as a runner-up.

Study 2

Finite-sample rejection rates for nested model testing are given in Table 6. We discuss the DWLS case first. The tests are sensitive to the underlying distribution. All tests produce rejection rates above the nominal level, especially under nonnormality. SB has the highest rejection rates. EBA2 and EBAA have lower rejection rates. However, the group of tests SS, F, EBAF, EBA4, EBA2J, and EBA4J has equal performance across all conditions, and attains better Type I error than SB, EBA2, and EBAA.

Under ML estimation, the situation is similar to the DWLS case, with a pattern of high rejection rates, decreasing toward the nominal level with increasing sample size. SB has the highest rejection rates. EBAA and EBA2 have lower rejection rates, but these are higher than the rather similar rejection rates in the group of SS, F, EBAF, EBA4, EBA2J, and EBA4J. This group achieves the lowest rejection rates, and so represents the best performing tests in terms of Type I error control.

DISCUSSION

The performance of the established and proposed new statistics has been studied, both asymptotically and in finite samples. The most important case for a practitioner is, of course, finite-sample performance. Consistent patterns among the test procedures were found across sample sizes and underlying distribution in both Study 1 and Study 2. For Study 1, the results in Table 5 suggest the following

TABLE 6
Study 2: Rejection Rates

	n	*Distr*	*T*	*SB*	*SS*	*CF*	*EBAF*	*EBA2*	*EBA4*	*EBA2J*	*EBA4J*	*EBAA*	*OR*
DWLS	100	N	.000	.096	.067	.065	.063	.075	.068	.069	.064	.091	.093
		D_1	.004	.236	.176	.171	.169	.194	.179	.179	.170	.209	.293
		D_2	.022	.313	.236	.231	.229	.260	.240	.237	.230	.270	.382
	300	N	.000	.076	.054	.052	.052	.062	.057	.058	.052	.074	.061
		D_1	.000	.151	.107	.104	.102	.121	.110	.111	.104	.136	.170
		D_2	.002	.191	.139	.136	.134	.155	.144	.141	.135	.166	.221
	1,000	N	.000	.066	.051	.049	.049	.055	.052	.053	.049	.066	.051
		D_1	.000	.106	.075	.074	.073	.083	.077	.078	.074	.100	.094
		D_2	.000	.126	.092	.089	.088	.100	.093	.094	.089	.117	.125
ML	100	N	.071	.082	.068	.066	.066	.070	.068	.069	.067	.082	.071
		D_1	.627	.198	.131	.128	.126	.154	.139	.135	.127	.164	.018
		D_2	.855	.276	.180	.176	.172	.212	.188	.184	.173	.220	.008
	300	N	.054	.058	.054	.054	.054	.054	.054	.054	.054	.058	.054
		D_1	.682	.124	.084	.082	.080	.098	.088	.086	.082	.102	.035
		D_2	.886	.165	.110	.107	.105	.129	.115	.112	.105	.133	.027
	1,000	N	.050	.052	.051	.051	.051	.051	.051	.051	.051	.052	.050
		D_1	.711	.084	.062	.060	.059	.069	.064	.063	.060	.073	.048
		D_2	.911	.111	.073	.071	.070	.087	.076	.075	.071	.087	.043

Note. T = χ^2 test; SB = Satorra–Bentler; SS = scaled and shifted; CF = scaled *F*; EBAF = full eigenvalue approximation; EBA*i* = *i*-block clustering; EBA*i*J = *i*-block Jenks clustering; EBAA = automatic clustering; OR = oracle; DWLS = diagonally weighted least squares; N = normality; D_1= moderate nonnormality; D_2= severe nonnormality; ML = maximum likelihood.

grouping of tests that perform similarly, ranked according to increasing rejection rates:

$$\text{Study1}: \text{CF/EBAF/EBA4J} < \text{SS/EBAA/EBA2J} < \text{EBA4} < \text{EBA2} < \text{SB},$$

whereas the results in Table 6 suggest the following grouping, ranked according to increasing rejection rates:

$$\text{Study2}: \text{SS/CF/EBAF/EBA4/EBA2J/EBA4J} < \text{EBA2} < \text{EBAA} < \text{SB}.$$

Also, some general observations holding across sample size, distributions, estimators, and models might be made: SB has the highest rejection rates. CF consistently has slightly lower rejection rates than SS. Remarkably, CF and EBAF have almost identical rejection rates in both models, for all sample sizes, distributions, and estimators. This echoes the findings of Wu and Lin (2016). In general, the EBA procedures perform similarly to SS and CF, with the exception of EBAA and EBA2, which tend to have somewhat higher rejection rates than SS and CF, but lower than SB.

Comparing the performance of EBA2 and EBAF across the two studies, it is noticeable that EBAF performed best in Study 2 (10 eigenvalues), whereas EBA2 performed best in Study 1 (35 eigenvalues). A possible explanation for this pattern is that in Study 2 there are more sample observations for each estimated eigenvalue. Intuitively, the eigenvalues are therefore estimated with higher precision in Study 2 compared to Study 1. The full use of the individual eigenvalues in Study 2 is more warranted than under conditions such as in Study 1, where there are far fewer observations per eigenvalue. In this latter condition it is therefore not surprising that the two-block method is found to be superior to EBAF.

We now turn to the question of evaluation. It is important to notice that there are two, sometimes conflicting, ways of evaluating test statistics. From a practical point of view, the important question is this: How well does the test control Type I error rates? This is the evaluation criterion in most simulation studies. However, the tests under consideration in this study were designed to emulate the oracle distribution in Equation 2. So theoretically, the important question is this: How well does the test approximate the oracle? Of course, it is hoped that these two evaluation criteria merge, and they certainly will for very large sample sizes. However, Tables 5 and 6 demonstrate that under realistic sample sizes, the oracle does not always achieve acceptable Type I error control. In some conditions it might therefore happen that a test statistic does a poor job approximating the oracle OR, but by some coincidence achieves good Type I error control. Consider, for instance, the condition in Study 1 of ML estimation under severe nonnormality and the smallest sample size. Here EBA2 outperforms all the other tests by a large margin, with a rejection rate of 0.055. However, the oracle has not yet reached its asymptotic limit of $\alpha = 0.05$, having a Type I rejection rate of only 0.018. Hence EBA2 does a very good job of controlling Type I error rates, although failing to achieve its theoretical aim of emulating the oracle. On the other hand, EBA2J matches the oracle rejection rate closely with a rejection rate of 0.016, but in terms of Type I error control this is unacceptably low.

Broadly speaking, evaluation in terms of Type I error control gave the following results. In Study 1, with $d = 35$ single model testing, the group SS, EBA4, EBA2J, and EBAA performed similarly, and attained the best Type I error control under DWLS, whereas EBA2 clearly outperformed all other tests under ML. In Study 2, with $d = 10$ nested model testing, the tests SS, CF, EBAF, EBA4, EBA2J, and EBA4J performed equally well, and better than SB, EBA2, and EBA4.

The second evaluation criterion considers how well the tests emulate the oracle. In Study 1, EBA2 performed the best, with the exception of DWLS under normality, where EBA4, EBA2J, and EBAA more closely matched the oracle rejection rates. In Study 2, the oracle was best approached by EBAA under DWLS, and by CF, EBAF, and EBA4J under ML.

CONCLUSION

Recently two test procedures, the scaled and shifted test SS (Asparouhov & Muthén, 2010), and the scaled *F* test CF (Wu & Lin, 2016) have been proposed based on approximating the asymptotic distribution of a weighted sum of chi-square variates in Equation 1. The SS and CF procedures match, respectively, the first two and the first three moments of the asympotic distribution, and are hence theoretically superior to the original scaling procedure of Satorra and Bentler (1988). In this article we have theoretically and empirically demonstrated, in the context of a specific model, that SS and CF outperform the SB procedure both for single and nested model testing. In accordance with earlier simulation studies, we therefore recommend SS and CF over the SB procedure, although SS and CF both tend to underreject correct models. Note that this recommendation still holds under normally distributed data.

We have also proposed new approximations to the weighted sum of i.i.d. chi-square variates, based on arranging eigenvalues in blocks and replacing them by average values. This introduces a whole new class of approximations to the asymptotic distribution of test statistics in SEM. We have compared six members of this class to the existing procedures CF and SB, both theoretically and empirically. In terms of correct Type I error control, the new procedures perform as well as SS and CF, and in some cases better. For instance, in the important case of ML testing a single model under nonnormality, a two-block eigenvalue approximation was found to outperform all other statistics.

Given the established SS and CF tests, and several well-performing EBA tests, the question then remains how one might perform model testing in a practical situation. In most cases, tests like CF, SS, and the EBA variants seem to result in similar model fit evaluations, as was the case for the illustrative example in this study. However, in some situations the tests might assess model fit differently. In such cases it would be recommended to report several test statistics. We suggest reporting one fixed-block and one dynamic-block EBA procedure. For instance SS, EBA2, and EBA4J could be reported.

REFERENCES

Amemiya, Y., & Anderson, T. W. (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

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

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181–197. doi:10.1207/S15327906Mb340203

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

Browne, M. W. (1982). Covariance structures. In D. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge: Cambridge University Press .doi:10.1017/CBO9780511897375.003

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83. doi:10.1111/bmsp.1984.37.issue-1

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, *41*, 193–208. doi:10.1111/bmsp.1988.41.issue-2

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. doi:10.1037/1082-989X.1.1.16

Duchesne, P., & De Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, *54*, 858–862. doi:10.1016/j.csda.2009.11.025

Foldnes, N. (2017). The impact of class attendance on student learning in a flipped classroom. *Nordic Journal of Digital Literacy*, *12*(1–2), 8–18. doi:10.18261/issn.1891-943x-2017-01-02-02

Foldnes, N., & Grønneberg, S. (2017). The asymptotic covariance matrix and its use in simulation studies. *Structural Equation Modeling*. Advance online publication. doi:10.1080/10705511.2017.1341320

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

Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*. Advance online pubication. doi:10.1007/s11336-017-9569-6

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, *7*(1), 186–190.

Lim, S. Y., & Chapman, E. (2013). Development of a short form of the Attitudes Toward Mathematics Inventory. *Educational Studies in Mathematics*, *82*, 145–164. doi:10.1007/s10649-012-9414-x

Mooijaart, A., & Bentler, P. M. (1991). Robustness of normal theory statistics in structural equation models. *Statistica Neerlandica*, *45*, 159–171. doi:10.1111/j.1467-9574.1991.tb01301.x

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

Rabosky, D., Grundler, M., Anderson, C., Title, P., Shi, J., Brown, J., … Larson, J. (2014). Bammtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, *5*, 701–707. doi:10.1111/2041-210X.12199

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151. doi:10.1007/BF02294453

Satorra, A., & Bentler, P. (1988). *Scaling corrections for statistics in covariance structure analysis* (UCLA Statistics Series 2). Los Angeles, CA: University of California at Los Angeles, Department of Psychology.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, *10*, 235–249. doi:10.1016/0167-9473(90)90004-2

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

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514. doi:10.1007/BF02296192

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures. *South African Statistical Journal*, *17*(1), 33–81.

Shapiro, A. (1987). Robustness properties of the MDF analysis of moment structures. *South African Statistical Journal*, *21*(1), 39–62.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465–471. doi:10.1007/BF02293687

Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal *k*-means clustering in one dimension by dynamic programming. *The R Journal*, *3*(2), 29–33. Retrieved from https://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf

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

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*, 115–148. doi:10.1207/s15327906mbr4001_5

Yuan, K.-H., & Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology*, *63*, 273–291. doi:10.1348/000711009X449771

APPENDIX
R CODE

```
# R version 3.3.1
library(lavaan) # Version0.5–22
library(CompQuadForm) # Version1.4.2
library(BAMMtools) # Version 2.1.6
library(Ckmeans.1d.dp) # Version 4.0.1

#sample size, skewness and kurtosis
n=300L
skewness=2L
kurtosis=10L
seed = 1

#Illustration based on a two-factor model
# specify population model
population.model <-

    "f1 = x1 +0.x2 +1.2*x3 +0.2*x4;
     f2 = ~ x5 +0.8*x6 +1.2*x7 +0.2*x8; f1~ ~0.5*f2"

#model to be estimated, has equality constraints on three residual variances
my.model <- "f1 = ~ x1+x2+x3+x4; f2 = ~ x5+x6+x7+x8; "

#simulate nonnormal data set
set.seed(seed)
my.dat = simulateData(population.model, sample.nobs = n,
                     skewness = rep(skewness,8), kurtosis = rep(kurtosis, 8))

#ML estimation
f = sem(my.model, data = my.dat)

#pvalues for default NTML and SB tests:
sem(my.model, data = my.dat, test = "SB")

#extract test statistic T
T = fitmeasures(f, "chisq")
#Extract U*Gamma
UG <- inspect(f, "UGamma")
#The estimated eigenvalues
df = fitmeasures(f,"DF")
eig. hat <− Re (eigen (UG) $ values [1: df])

########
## p values for various tests, based on eigenvalues
########

#NTML
pNTML <− imhof (T, rep (1, d f)) $Qq

#SB
pSB <− imhof (T, rep (mean (eig. hat), d f) ) $Qq

#EBAF
pEBAF <− imhof (T, eig. hat) $Qq

#EBA2
eigs <− c (rep (mean (eig. hat [1 : ceiling (d f/2) ] ), ceiling (d f/2) ),
       rep (mean (eig. hat [(ceiling (d f/2) +1) : df]), df − ceiling (d f/2))) pEBA2 <− imhof (T, eigs) $Qq

#Jenks EBA2
breaks <− get Jenks Breaks (eig. hat, k = 3)
block1 <− eig. hat [eig. hat ≤ breaks [2] ] ; block2 = eig. hat [eig. hat > breaks [2] ]
eigs <− c (rep (mean (block1), length (block1) ), rep (mean (block2), length (block2)) )
pEBA2J <− imhof (T, eigs) $Qq
```

```
#EBAA
t = Ckmeans.1d.dp (eig. hat)
means <− t $ centers
clusters <− t $ cluster
eigs <− sapply (clusters, function (x) means [x])
pEBAA <− imhof (T, eigs) $Qq

cat (" Simple model testing : \n " )
print (round (data. frame (pNTML, pSB, pEBAF, pEBA2, pEBA2J, pEBAA), 4) )

######
## Nested Model Testing
######

#help function. From lavaansource code.

eigen values _ dif<− function (m1, m0, A. method = " exact") { #or delta. Note that shell command lavTestLRT
     has exact, while lav_test_diff_Satorra 2000 has default delta.

     # extractinformation from m1 and m2
     T1 <− m1@test [[1] ] $ stat
     r1 <− m1@test [[1] ] $ df

     T0 <− m0@test [[1] ] $ stat
     r0 <− m0@test [[1] ] $ df

     # m = diference between the df ’ s ’
     m <− r0 − r1
     Gamma <− lav Tech (m1, "Gamma") # the same for m1 and m0
     WLS. V <− lav Tech (m1, "WLS. V" )
     PI <− lavaan : : : compute Delta (m1@Model)
     P <− lav Tech (m1, " information " )
     # needed ? (yes, if H1 already has eq constraints)
     P. inv <− lavaan : : : lav_ model_ information_ augment_ invert (m1@Model,
                                                                         information = P,
                                                                         inverted = TRUE)

     if (inherits (P. inv, " try −error ") ) {
           cat (" Error! in P. inv \n ")
           return (NA)
     }

     A <− lavaan:::lav_test_diff_A (m1, m0, method = A. method, reference = " H1 " )

     APA <− A %*% P. inv %*% t (A)
     cSums <− col Sums (APA)
     rSums <− rowSums (APA)
     empty. idx <− which (abs (cSums) < . Machine$ double. eps ^0. 5 &
                                 abs (rSums) <. Machine$ double. eps ^0. 5)
     if (length (empty. idx) > 0) {
           A <− A[− empty. idx,, drop = FALSE]
}

     # PAAPAAP
     PAAPAAP <− P. inv %*% t (A) %*% solve (A %*% P. inv %*% t (A)) %*% A %*% P. inv

     g = 1
     UG. group <− WLS. V [[g] ] %*% Gamma [[g] ] %*% WLS. V [[g] ] %*%
           PI [[g] ] %*% PAAPAAP %*% t (PI [[g] ] )

     return (Re (eigen (UG. group) $ values) [1 : m] )
}

my. model. restricted <− "f 1 = ~ x1 + b* x2 + c * x3 + d* x4 ;
                         f 2 = ~ x5 + b* x6 + c * x7 + d* x8 ;
                         x1 ~~ a * x1 ; x2 ~~ a * x2 ; x3 ~~ a * x3 ; x4 ~~ a * x4 ;
                        x5 ~~ a * x5 ; x6 ~~ a * x6 ; x7 ~~ a * x7 ; x8 ~~ a * x8 ; "
```

```
f. restricted = sem (my. model. restricted, my. dat)

#the estimated eigenvalues for diference test. 10 df.
eig. hat = eigen values _ dif(f, f. restricted)

#chisquare diference
T <− fitmeasures (f . restricted, " chisq " )− fitmeasures (f, " chisq ")
df <− fitmeasures (f . restricted, "DF" )− fitmeasures (f, "DF" )
#NTML
pNTML <− imhof (T, rep (1, df)) $Qq

#SB
pSB <− imhof (T, rep (mean (eig. hat), df) ) $Qq

#EBAF
pEBAF <− imhof (T, eig. hat) $Qq

#EBA2
eigs <− c (rep (mean (eig. hat [1 : ceiling (df/2) ] ), ceiling (df/2) ),
          rep (mean (eig. hat [(ceiling (df/2) +1) : df] ), df − ceiling (df/2)) )
pEBA2 <− imhof (T, eigs) $Qq

#Jenks EBA2
breaks <− getJenksBreaks (eig. hat, k = 3)
block1 <− eig. hat [eig. hat ≤ breaks [2] ] ; block2 = eig. hat [eig. hat > breaks [2] ]
eigs <− c (rep (mean (block1), length (block1) ), rep (mean (block2), length (block2)) )
pEBA2J <− imhof (T, eigs) $Qq

#EBAA
t = Ckmeans . 1 d . dp (eig. hat) means <− t $ centers
clusters <− t $ cluster
eigs <− sapply (clusters, function (x) means [x]) pEBAA <− imhof (T, eigs) $Qq

cat("\n Nested model testing : \n ")
print (round (data . frame (pNTML, pSB, pEBAF, pEBA2, pEBA2J, pEBAA), 4))
```