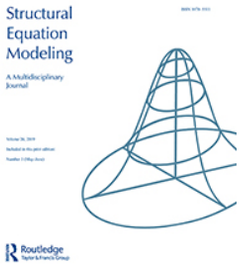
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

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

New Effect Size Measures for Structural Equation Modeling

Brenna Gomer, Ge Jiang & Ke-Hai Yuan

To cite this article: Brenna Gomer, Ge Jiang & Ke-Hai Yuan (2019) New Effect Size Measures for Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal, 26:3, 371-389, DOI: 10.1080/10705511.2018.1545231

To link to this article: https://doi.org/10.1080/10705511.2018.1545231

Published online: 18 Dec 2018.

Structural Equation Modeling: A Multidisciplinary Journal, 26: 371–389, 2019

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2018.1545231

New Effect Size Measures for Structural Equation Modeling

Brenna Gomer,[1] Ge Jiang,[2] and Ke-Hai Yuan[1]
[1]*University of Notre Dame*
[2]*University of Illinois Urbana Champaign*

Effect size is crucial for quantifying differences and a key concept behind Type I errors and power, but measures of effect size are seldom studied in structural equation modeling (SEM). While fit indices such as the root mean square error of approximation may address the severity of model misspecification, they are not a direct generalization of commonly used effect size measures such as Cohen's *d*. Moreover, with violations of normality and when a test statistic does not follow a noncentral chi-square distribution, measures of misfit that are defined through the assumed distribution of the test statistic are no longer valid.

In this study, two new classes of effect size measures for SEM are developed by generalizing Cohen's *d*. The first class consists of definitions that are theoretically equivalent to $F_{ML}(\Sigma_0, \Sigma(\boldsymbol{\theta}))$, the population counterpart of the normal-distribution-based discrepancy function. The second class of effect size measures bears a stricter resemblance to Cohen's *d* in its original form. Several versions of these generalizations are investigated to identify the one that is least affected by sample size and population distribution but most sensitive to model misspecification. Their performances under violated distributional assumptions, severity of model misspecification, and various sample sizes are examined using both normal maximum likelihood estimation and robust M-estimation. Monte Carlo results indicate that one measure in the first class of effect size measures is little affected by sample size and distribution while preserving sensitivity to model misspecification and thus is recommended for researchers to report in publications.

Keywords: Cohen's *d*, effect size, model fit, model misspecification

INTRODUCTION

There have been many developments related to fit indices in the structural equation modeling (SEM) literature, especially regarding their use in null hypothesis significance testing (NHST) (Cheng & Wu, 2017; Hu & Bentler, 1999; Zhang & Savalei, 2016). However, the topic of effect size measures is equally important and has been neglected with few exceptions (Maydeu-Olivares, 2017; Yuan & Marshall, 2004). In this article, we propose two new classes of effect size measures to fill this gap. These measures are inspired by Cohen's *d* and target overall model misspecification. Effect size measures that target model misspecification should be developed for two reasons.

First, researchers rely on effect size measures to supplement information gained from NHST. The level of significance of a test statistic does not indicate how much the model deviates from the population, which is often of greater substantive interest. Due to this limitation, some authors have advocated discontinuing significance tests altogether in favor of point estimates and confidence intervals (Schmidt & Hunter, 1997; Steiger & Fouladi, 1997).

Second, hypothesis tests in the SEM framework can reject good models simply due to large sample size or non-normally distributed data (Bentler & Yuan, 1999; Fouladi, 2000; Hu, Bentler, & Kano, 1992; Nevitt &

Correspondence should be addressed to Brenna Gomer, University of Notre Dame, E414 Corbett Hall, Notre Dame, IN 46556. E-mail: bgomer@nd.edu

Hancock, 2004; Savalei, 2008; Tanaka, 1987). Treating NHST as the major authority on a model's usefulness is worrisome. Effect size measures that are relatively unaffected by these factors can reduce this dependency and give a more complete picture.

Aren't fit indices effect size measures?

Current practice often treats fit indices as effect sizes, and while they do assess the severity of model misspecification, they cannot be used as effect sizes in their current form for several reasons.

They are used in hypothesis testing

First, and most importantly, fit indices currently operate in the context of null hypothesis testing and their cutoff values have been refined to control rejection rates, not to describe the misspecification size of the model (Hu & Bentler, 1999). While their original purpose was to describe the goodness of fit of a model (Bentler, 1990; Bentler & Bonett, 1980; Tucker & Lewis, 1973), they now play a role in NHST. This is not surprising since T_{ML} has severe limitations making it an unreliable test statistic (Bentler & Bonett, 1980). However, an effect size measure should not be simultaneously used for the purpose of hypothesis testing and as an effect size. Not only is this statistically inappropriate, it negates the usefulness of an effect size measure as a supplement to a hypothesis test.

Assumptions are often violated

Second, real data often violate the distributional assumptions underlying many fit indices. It has been said that normally distributed data are as rare as unicorns (Micceri, 1989). In the presence of non-normal data, fit indices such as the root mean square error of approximation (RMSEA) that are defined through the χ^2 distribution of the test statistic T_{ML} are no longer valid (Yuan, 2005).

Their performance is unreliable

Third, the values of fit indices are influenced by factors other than model misfit even when their underlying assumptions are satisfied. Some of these factors include model complexity, sample size, degrees of freedom, ratio of sample size to the number of variables, and even the magnitude of factor loadings (Moshagen & Auerswald, 2017; Shi, Lee, & Terry, 2018; Yuan, Jiang, & Yang, 2018). Because the influence of these factors varies from model to model, a one-size-fits-all approach to cutoff values will result in inconsistent conclusions when applied to different types of models (Beauducel & Wittmann, 2005; Fan & Sivo, 2005; Marsh, Hau, & Wen, 2004; West, Taylor, & Wu, 2012; Yuan, 2005).

While we have stated that fit indices cannot be used as effect size measures in their current form, we do not claim that fit indices can *never* be used as effect size measures. Indeed, their original goal to capture goodness of fit is an essential feature of an effect size measure. However, their use in null hypothesis testing is now somewhat entrenched and their limitations are concerning. With some modification and willingness to change old habits, it may be possible to recover the original use of fit indices as goodness-of-fit measures and apply them to real data.

However, in this article, we propose two classes of novel effect size measures that can be applied without changing current practice. Our contribution is to introduce direct generalizations of Cohen's d to SEM, a classic measure of effect size that currently has no counterpart in this area. The effect size measures we propose are global measures of model misspecification.

The remainder of this article is organized as follows. First, we define two new classes of effect size measures. Second, we describe the procedure of a Monte Carlo study investigating the performance of these new measures. Third, we present the results of the simulation study. Fourth, we develop preliminary practical guidelines for the best effect size measure and give two empirical examples. Last, we discuss our overall findings and recommendations.

NEW EFFECT SIZE MEASURES

In this section, we present two new classes of effect size measures for SEM. We will denote matrices as bold-face capital letters and vectors as bold-face lowercase letters. We begin with a brief discussion of our evaluation criteria. Next, we show how the concept of Cohen's d in combination with the evaluation criteria motivates the general form of our new measures. Then, we discuss the definitions of the new effect size measures as population parameters. We end this section with a brief note on possible substitutions of test statistics and a discussion on obtaining point estimates and confidence intervals for the effect size measures.

Evaluation criteria

In our view, sensitivity to size of model misspecification and logical justification are strict requirements for a valid effect size measure. In addition to these criteria, an effect size measure should be sample size and model size resistant so that its value reflects the severity of model misspecification and little else. A reliable effect size measure is applicable under a variety of underlying population distributions or is robust to distributional assumption violations. Ideally, it will also facilitate easy interpretation.

Effect size measures concept

A classic effect size measure is Cohen's d, a standardized mean difference between two groups (Cohen, 1988). It is a popular measure that has an intuitive meaning and forms the logic behind the two new classes of effect size measures we develop in this article. By generalizing the formula of Cohen's d to SEM, we aim to minimize the effects of sample size and population distribution while preserving sensitivity to the size of model misfit. This section describes how we translated Cohen's d to SEM. We will begin in the context of SEM and briefly introduce relevant background needed to construct our effect size measures. Next, we will introduce Cohen's d and show how we translated its formula into two new classes of effect size measures.

In Cohen's d, the numerator is simply the difference of two univariate sample means. However, we have no direct translation to this in SEM. SEM models are multivariate in nature and often the analysis is focused on the covariance structure rather than the mean structure. Instead of the sample mean, the numerical summary that is typically used for an SEM model is the test statistic T_{ML}:

$$T_{ML} = (N-1)F_{ML}(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}})) \tag{1}$$

where N is the number of observations and $F_{ML}(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}}))$ is the estimated discrepancy from the maximum likelihood discrepancy function. The discrepancy function at the population level is given by

$$F_{ML}(\Sigma_0, \Sigma(\boldsymbol{\theta})) = tr(\Sigma_0\Sigma^{-1}(\boldsymbol{\theta})) - log\left|\Sigma_0\Sigma^{-1}(\boldsymbol{\theta})\right| - p \tag{2}$$

where p is the number of manifest variables. Here, the population covariance matrix Σ_0 is replaced by the sample covariance matrix $\mathbf{S}$, and the vector of model parameters $\boldsymbol{\theta}$ in the model-implied covariance matrix $\Sigma(\boldsymbol{\theta})$ is replaced by its estimate $\hat{\boldsymbol{\theta}}$. We will refer to the discrepancy at the population level $F_{ML}(\Sigma_0, \Sigma(\boldsymbol{\theta}^*))$ as F_0, where $\boldsymbol{\theta}^*$ is the population counterpart of $\boldsymbol{\theta}$ and obtained by minimizing the function in equation (2).

The discrepancy is a single value that quantifies the difference between the hypothesized model structure and what is observed in the sample by comparing the sample covariance to the covariance under the working model. Statistics such as T_{ML} that use the discrepancy may provide a suitable single number summary of an SEM model, which is multivariate in nature. Other such statistics include the rescaled test statistic T_{RML} (Satorra & Bentler, 1988).

We will introduce T_{RML} here because we use its features to construct some effect size measures in the next section. T_{RML} is an adjustment to T_{ML} and is given by

$$T_{RML} = T_{ML}/r \tag{3}$$

where r is a correction factor that accounts for violation of normality via the sample fourth-order moments of the observed data $\mathbf{x}_i$ and the particular model structure. Specifically, r is given by $tr(\hat{\mathbf{H}}\hat{\boldsymbol{\Gamma}})/(p^* - q)$, with $p^* = p(p+1)/2$ and q equal to the number of free parameters. $\hat{\boldsymbol{\Gamma}}$ is the estimate of $\boldsymbol{\Gamma}$, the asymptotic covariance matrix of $\mathbf{y}_i = \text{vech}[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)']$. $\hat{\boldsymbol{\Gamma}}$ contains the sample fourth-order moments and thus reflects the distribution of $\mathbf{x}_i$ or distributional violations against normality. $\hat{\mathbf{H}}$ is the estimate of $\mathbf{H} = \mathbf{W} - \mathbf{W}\dot{\sigma}(\dot{\sigma}'\mathbf{W}\dot{\sigma})^{-1}\dot{\sigma}'\mathbf{W}$, where $\mathbf{W} = 2^{-1}\mathbf{D}'_p[\Sigma^{-1} \otimes \Sigma^{-1}]\mathbf{D}_p$, $\mathbf{D}_p$ is a $p^2 \times p^*$ duplication matrix (Magnus & Neudecker, 1988), $\otimes$ denotes the Kronecker product, and $\dot{\sigma}$ denotes the derivative of $\text{vech}(\Sigma(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ and evaluated at the population value. The operator $\text{vech}(\cdot)$ on a symmetric matrix yields a vector that contains the non-duplicated elements of the matrix. $\mathbf{H}$ is also known as the residual weight matrix and reflects the particular model structure (Bentler & Dudgeon, 1996).

Thus, test statistics such as T_{ML} and T_{RML} describe important features of model fit and may substitute the sample mean in Cohen's d. Note that this does not obligate us to make any distributional assumptions.

The numerator of Cohen's d is a difference of means between two samples under comparison. In a two-sample t-test,

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 are population means corresponding to the two groups. Cohen's d is given by

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \tag{4}$$

where s is the pooled standard deviation, $\bar{x}_1$ is the sample mean of the first group, and $\bar{x}_2$ is the sample mean of the second group. We denote the population version of Cohen's d as δ given by

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{5}$$

where σ is the common population standard deviation. This can be conceptualized more generally as

$$\begin{aligned}\delta &= \frac{\mu_1 - \mu_2 - 0}{\sigma} \\ &= \frac{E[(\bar{x}_1 - \bar{x}_2)|H_1] - E[(\bar{x}_1 - \bar{x}_2)|H_0]}{SD}\end{aligned} \tag{6}$$

where the notation $E[(\bar{x}_1 - \bar{x}_2)|H_1]$ denotes the expected difference between $\bar{x}_1$ and $\bar{x}_2$ under the alternative hypothesis H_1. The notation $E[(\bar{x}_1 - \bar{x}_2)|H_0]$ denotes this difference under the null hypothesis H_0. Our generalizations to SEM are inspired by this form. In the example, $\bar{x}_1 - \bar{x}_2$ is the statistic of interest, but in SEM, this idea corresponds to the discrepancy function F_{ML}. In order to describe the behavior of F_{ML} by a commonly used probability distribution, it needs to be multiplied by $N - 1$ so that the resulting T_{ML} asymptotically follows a χ^2 distribution under normality. Under non-normality, modifications of this statistic have been developed which still have the same conceptual meaning.

While in the two-sample *t*-test example the standard deviation standardizes the mean difference, T_{ML} is already standardized when distributional assumptions hold. When they do not hold, there is no universally applicable adjustment that will standardize it. Because of this, we translate the idea of a standardized mean difference in two ways which forms the basis for the two classes of effect size measures.

The first class of effect size measures resembles the form

$$\left(\frac{E(T_{ML}|H_1) - E(T_{ML}|H_0)}{N-1}\right)^{1/2} \tag{7}$$

Dividing by $N - 1$ drastically reduces the impact of sample size. With this, the square root creates some relation to the noncentrality parameter which we will discuss further in the next section. The second class of effect size measures has the form

$$\frac{E(T_{ML}|H_1) - E(T_{ML}|H_0)}{SD(T_{ML}|H_0)} \tag{8}$$

This mimics the original formula of Cohen's *d* in equation 5 more closely. We will introduce the definitions of different effect size measures within the two classes in the following section.

Definitions of effect size measures

In this section, we define two new classes of effect size measures. We begin with the definitions for the effect size measures in each class as population parameters. Then, we discuss possible substitutions of the test statistic. Finally, we provide an overview of how to obtain point estimates and confidence intervals of the effect size measures.

First class of effect size measures

The first class of effect size measures is characterized by the form given in equation 7. The first effect size measure in this class is given by

$$\mathcal{E}_1 = \frac{[E(T_{ML}|H_1) - E(T_{ML}|H_0)]}{[(N-1)\cdot(VAR(T_{ML}|H_1) - 2df)]^{1/2}/2} \tag{9}$$

where *df* denotes the model degrees of freedom. This moderately differs from the form of equation 7: The numerator is outside of the square root and the denominator can be interpreted as a difference of variances. The benefit of constructing $\mathcal{E}_1$ in this way is that it is theoretically equivalent to $\sqrt{F_0}$ if no assumptions are violated. Under the normality assumption, $T_{ML}|H_0$ follows a central χ^2 distribution with mean *df* and variance $2df$. Under this assumption, $T_{ML}|H_1$ approximately follows a noncentral χ^2 distribution with mean $df + \lambda_n$ and variance $2(df + 2\lambda_n)$, where $\lambda_n = (N-1)F_0$. Then, $E(T_{ML}|H_1) \approx df + \lambda_n$, $E(T_{ML}|H_0) \approx df$, and $VAR(T_{ML}|H_1) \approx 2df + 4\lambda_n$. This yields

$$\mathcal{E}_1 = \frac{df + \lambda_n - df}{\sqrt{N-1}\sqrt{\lambda_n}} = \frac{\sqrt{\lambda_n}}{\sqrt{N-1}} = \sqrt{F_0}$$

Defining the effect size as equation 9 also allows for greater flexibility in conditions when assumptions are violated.

The next effect size measure is

$$\mathcal{E}_2 = E\left[\left(\frac{max(T_{ML} - df, 0)}{N-1}\right)^{1/2}\right] \tag{10}$$

Unlike equation 7, this effect size measure assumes directly that $E(T_{ML}|H_0) = df$. Because $T_{ML} - df$ can be less than 0 when the distribution of the data has light tails, the numerator is adjusted to $max(T_{ML} - df, 0)$.

The third effect size measure is

$$\mathcal{E}_3 = \left(\frac{E(T_{ML}|H_1) - E(T_{ML}|H_0)}{N-1}\right)^{1/2} \tag{11}$$

Like equation 7, this effect size measure standardizes by using $N - 1$ and does not explicitly assume a distribution for T_{ML}. When applied to real data, it is possible for the numerator to be less than 0 due to sampling variability when the model is correctly specified or if the misspecification is very small. In these cases, $\mathcal{E}_3$ should be taken to be 0.

The fourth effect size measure is

$$\mathcal{E}_4 = E\left[\left(\frac{T_{ML} - tr(\mathbf{H\Gamma})}{N-1}\right)^{1/2}\right] \tag{12}$$

$\mathcal{E}_4$ assumes that $E(T_{ML}|H_0) = tr(\mathbf{H\Gamma})$. It has been shown that with sufficiently large N, $E(T_{ML}|H_0) = tr(\mathbf{H\Gamma})$ where $\mathbf{H}$ and $\mathbf{\Gamma}$ are as we previously defined (Satorra & Bentler, 1988). Under normality, $tr(\mathbf{H\Gamma}) = df$ and thus $E(T_{ML}|H_0)$ is as expected under a central χ^2 distribution. However, $tr(\mathbf{H\Gamma})$ can be greater than df when the normality assumption is violated. For example, when the underlying population follows a multivariate t-distribution, $tr(\mathbf{H\Gamma})$ is equal to df multiplied by the relative kurtosis of the t-distribution.

The next effect size measure is

$$\mathcal{E}_5 = E\left[\left(\frac{T_{RML} - df}{N-1}\right)^{1/2}\right] \tag{13}$$

This measure differs from equation 7 in the use of the rescaled statistic T_{RML} instead of T_{ML} and in assuming $E(T_{RML}|H_0) = df$. However, this assumption may not be realistic in practice (Yuan et al., 2018).

The last effect size measure in the first class is

$$\mathcal{E}_6 = \left(\frac{E(T_{RML}|H_1) - E(T_{RML}|H_0)}{N-1}\right)^{1/2} \tag{14}$$

This is identical to $\mathcal{E}_3$ except that it replaces T_{ML} with T_{RML}.

All the effect size measures in the first class have some theoretical equivalency to F_0 if distributional assumptions are satisfied. Some measures make these assumptions explicitly in their formulas ($\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_5$) while others ($\mathcal{E}_3$, $\mathcal{E}_4$, and $\mathcal{E}_6$) do not. When the assumptions do not hold, we hypothesize that the measures whose definitions do not rely on distributional assumptions will perform better. Note that the definitions of $\mathcal{E}_2$, $\mathcal{E}_4$, and $\mathcal{E}_5$ place the expectation outside of the square root. This may make them easier to execute and more suitable for meta-analysis than the other candidates. We display these effect size measures in Table 1 for reference.

TABLE 1
First Class of Effect Size Measures

Effect size	*Formula*	*Assumptions*
$\mathcal{E}_1$	$\frac{2\cdot[E(T_{ML}\|H_1)-E(T_{ML}\|H_0)]}{[(N-1)\cdot(VAR(T_{ML}\|H_1)-2df)]^{1/2}}$	$Var(T_{ML}\|H_1) = 2df + 4\lambda_n$
$\mathcal{E}_2$	$E\left[\left(\frac{max(T_{ML}-df,0)}{N-1}\right)^{1/2}\right]$	$E(T_{ML}\|H_0) = df$
$\mathcal{E}_3$	$\left(\frac{E(T_{ML}\|H_1)-E(T_{ML}\|H_0)}{N-1}\right)^{1/2}$	–
$\mathcal{E}_4$	$E\left[\left(\frac{T_{ML}-tr(\mathbf{H\Gamma})}{N-1}\right)^{1/2}\right]$	$E(T_{ML}\|H_0) = tr(\mathbf{H\Gamma})$
$\mathcal{E}_5$	$E\left[\left(\frac{T_{RML}-df}{N-1}\right)^{1/2}\right]$	$E(T_{RML}\|H_0) = df$
$\mathcal{E}_6$	$\left(\frac{E(T_{RML}\|H_1)-E(T_{RML}\|H_0)}{N-1}\right)^{1/2}$	–

Note. $\mathcal{E}_3$ and $\mathcal{E}_6$ do not make use of any assumptions in their formulas.

Second class of effect size measures

The effect size measures in the second class are formulated as a mean difference divided by the standard deviation under H_0:

$$D = \frac{E(T_{ML}|H_1) - E(T_{ML}|H_0)}{SD(T_{ML}|H_0)} \tag{15}$$

Unlike the first class, these measures have no theoretical equivalency to F_0 but are a more faithful translation of Cohen's d. We denote this class of effect size measures as D to differentiate these effect size measures from the previous class and to emphasize their close relationship to Cohen's d. We vary the measures by explicitly assuming an underlying central χ^2 null distribution in different parts of equation 15. That is, some versions assume that $E(T_{ML}|H_0) = df$ in the numerator while others assume that $SD(T_{ML}|H_0) = \sqrt{2df}$. We also test the use of T_{RML} instead of T_{ML}. These effect size measures are presented in Table 2. The columns of the table change the explicit assumptions made in the denominator of the formula while the rows change the explicit assumptions in the numerator.

Version 1 of the measures uses $\sqrt{2df}$ as its standard deviation, deviating from equation 15 by making the explicit assumption that $T_{ML}|H_0$ follows a central χ^2 distribution. The second version of the effect sizes does not rely on such an assumption. The third version of the effect sizes assumes that T_{ML} follows a central χ^2 distribution so that its variance is $2df$ but replaces $E(T_{ML}|H_0)$ with $tr(\mathbf{H\Gamma})$ since it should

TABLE 2
Second Class of Effect Size Measures

Effect size	*Version 1*	*Version 2*	*Version 3*
D_1	$E\left[\frac{T_{ML}\|H_1-df}{\sqrt{2df}}\right]$	$E\left[\frac{T_{ML}\|H_1-df}{SD(T_{ML}\|H_0)}\right]$	$E\left[\frac{T_{ML}\|H_1-df}{\sqrt{2tr(\mathbf{H\Gamma})}}\right]$
D_2	$E\left[\frac{T_{ML}\|H_1-T_{ML}\|H_0}{\sqrt{2df}}\right]$	$E\left[\frac{T_{ML}\|H_1-T_{ML}\|H_0}{SD(T_{ML}\|H_0)}\right]$	$E\left[\frac{T_{ML}\|H_1-T_{ML}\|H_0}{\sqrt{2tr(\mathbf{H\Gamma})}}\right]$
D_3	$E\left[\frac{T_{ML}\|H_1-tr(\mathbf{H\Gamma})}{\sqrt{2df}}\right]$	$E\left[\frac{T_{ML}\|H_1-tr(\mathbf{H\Gamma})}{SD(T_{ML}\|H_0)}\right]$	$E\left[\frac{T_{ML}\|H_1-tr(\mathbf{H\Gamma})}{\sqrt{2tr(\mathbf{H\Gamma})}}\right]$
D_4	–	$E\left[\frac{T_{RML}\|H_1-df}{SD(T_{RML}\|H_0)}\right]$	$E\left[\frac{T_{RML}\|H_1-df}{\sqrt{2df}}\right]$
D_5	–	$E\left[\frac{T_{RML}\|H_1-T_{RML}\|H_0}{SD(T_{RML}\|H_0)}\right]$	$E\left[\frac{T_{RML}\|H_1-T_{RML}\|H_0}{\sqrt{2df}}\right]$

Note. Each measure makes different explicit assumptions about $E(T_{ML}|H_0)$. D_1 assumes it equals df, D_2 makes no assumption, and D_3 assumes it equals $tr(\mathbf{H\Gamma})$. D_4 and D_5 use T_{RML} in place of T_{ML}. Versions 1–3 of each measure vary the assumptions made about the standard deviation. Version 1 assumes $SD(T_{ML}|H_0) = \sqrt{2df}$, while version 2 makes no assumption. Version 3 assumes $SD(T_{ML}|H_0) = \sqrt{2tr(\mathbf{H\Gamma})}$. Note that since T_{RML} is rescaled to take into account that $VAR(T_{ML}|H_0)$ does not equal $2df$, it does not make sense to create a version 1 for D_4 and D_5. D_4 assumes $E(T_{RML}|H_0) = df$ while D_5 makes no assumption.

be equivalent with large N. Under normality, this should be equal to the degrees of freedom.

Substitutions of the test statistic

Let us note that while we have only considered the test statistics T_{ML} and T_{RML} in this article, there are other test statistics (e.g., the asymptotically distribution-free statistic T_{ADF}) that are also theoretically justified for both classes of effect size measures. Of course, the performance of these substitutions should be investigated in a simulation study and their suitability verified before being applied to real data.

Point estimates and confidence intervals

So far, we have introduced definitions of the effect size measures as population parameters. We can only obtain sample estimates from real data, so here we will briefly mention how to get point estimates and confidence intervals. Interested readers can find more details in the Appendix.

Obtaining Estimates. Effect size measures that do not contain statistics under H_0 can be estimated directly from the data. However, many of the measures we have defined contain a difference of expectations. Using the procedure described in Section 3 of Yuan and Marshall (2004), estimates of $E(T_{ML}|H_1) - E(T_{ML}|H_0)$ can be obtained. This procedure can be modified to get estimates of $E(T_{RML}|H_1)$, $E(T_{RML}|H_0)$, $VAR(T_{ML}|H_1)$, and $SD(T_{ML}|H_0)$ which can be used to obtain point estimates of the effect size measures. We recommend using the point estimates together with bootstrap confidence intervals. For example, these can include percentile intervals, bias-corrected (BC) intervals, and bias-corrected and accelerated (BCa) intervals (Efron & Tibshirani, 1994). Algorithm II in Yuan and Marshall (2004) can be followed to get confidence intervals for the effect size measures we have defined. We will illustrate the performance of these confidence intervals in two applied examples later on in this article.

Statistical Properties of Estimators. Statistical properties of estimators are often of interest—particularly consistency and unbiasedness. The effect size measures we defined have estimators that enjoy one or both properties.

First, we discuss consistency. All of the effect size measures have either consistent or consistent at large[1] estimators. This is because $F_{ML}(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}}))$ is consistent for $F_{ML}(\Sigma_0, \Sigma(\boldsymbol{\theta}^*))$ and the effect size measures are functions of the test statistic $T_{ML} = (N-1)F_{ML}(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}}))$ (Kano, 1986; Shapiro, 1984). As long as this $(N-1)$ term can be cancelled out, the estimators inherit statistical consistency from $F_{ML}(\mathbf{S}, \Sigma(\hat{\boldsymbol{\theta}}))$. The bootstrap estimators of $E(T_{ML}|H_0)$ and $VAR(T_{ML}|H_1)$ are also consistent in this case (Beran & Srivastava, 1985). If the $(N-1)$ term cannot be cancelled out, then the effect size estimators are consistent at large. Thus, $\hat{\mathcal{E}}_1$, $\hat{\mathcal{E}}_3$, and $\hat{\mathcal{E}}_6$ are consistent in the first class of effect size measures. The estimators of all other effect size measures we defined are consistent at large.

Some effect size measures also have unbiased estimators. These measures are defined with expectations on the outside of their formulas, so unbiasedness of their estimators comes naturally. In the first class, $\hat{\mathcal{E}}_2$, $\hat{\mathcal{E}}_4$, and $\hat{\mathcal{E}}_5$ are unbiased. Version 1 of $\hat{D}_1$, $\hat{D}_2$, and $\hat{D}_3$ and version 3 of $\hat{D}_4$ and $\hat{D}_5$ are unbiased in the second class of effect size measures.

METHOD

In this section, we describe the procedure for a Monte Carlo simulation study to evaluate the performance of the proposed effect size measures. We begin with an overview of conditions for the simulation study. Then, we detail the data-generating process, the confirmatory factor models used in the study, and estimation methods. Finally, we discuss how the effect size measures are estimated.

Overview

To evaluate our effect size measures, we conducted a Monte Carlo simulation study using a confirmatory 3-factor model and 1,000 replications in all conditions. We tested the sensitivity of our measures to the size of model misspecification which we quantified by various values of omitted cross-factor loadings denoted as a. We also varied the sample size, model size, and the population distribution underlying the sample to determine

TABLE 3
Simulation Conditions

Size of model misspecification (a)	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
Sample size (N)	75, 100, 120, 150, 200, 300, 500, 800, 1,000
Number of manifest variables (p)	9, 15, 30
Distribution of manifest variables	Normal, t, exponential, elliptical, non-normal, normal with outliers (mix 1, mix 2, mix 3), uniform
Estimation method	ML, M-estimation

Note. Size of model misspecification is quantified by the size of omitted cross-factor loadings a.

[1] *Consistency at large* holds when the effect size measure and its estimator approach the same value, although both their limit values might be infinity.

which effect size measures are least affected by these factors. We used both ML and M-estimation methods for each of these conditions. The specific simulation conditions we tested are presented in Table 3 (5,346 conditions in total) and described in more depth in the next few pages.

Data-generating distributions

We generated normal data from a multivariate normal distribution with $p \times p$ covariance matrix Σ_0 equal to specifications we describe in the next section. Data following a t-distribution were generated from a multivariate t-distribution with 5 degrees of freedom and $p \times p$ covariance matrix Σ_0, which has substantially heavier tails than the normal distribution. We generated exponential and uniform data with the following transformation:

$$\mathbf{x} = \Sigma_0^{1/2}\mathbf{z} \tag{16}$$

where $\mathbf{z}$ has length p and each element follows a standardized exponential distribution to generate exponential data and a standardized uniform distribution to generate uniform data.

We generated elliptical data by using the following transformation:

$$\mathbf{x} = r\mathbf{z} \tag{17}$$

where $\mathbf{z}$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ of length p and covariance matrix Σ_0. Also, $r = (\chi_5^2/3)^{-1/2}$ where χ_5^2 denotes a central χ^2 distributed variable with 5 degrees of freedom, and $\mathbf{z}$ and r are independent. The purpose of this transformation is to preserve the mean and covariance while changing the multivariate relative kurtosis to 3.

We generated non-elliptical non-normal data by modifying equation (3.1) in Yuan and Bentler (1999) given below:

$$\mathbf{x} = r\Sigma^{1/2}\mathbf{z} \tag{18}$$

where the elements of $\mathbf{z}$ are independent and each follows a standardized non-normal distribution, and r is an independent nonnegative random variable. In particular, the p elements of $\mathbf{z}$ are independent and each follows a standardized exponential distribution and the same r is used as for the elliptical data.

The mixed distributions consisted of a contaminated normal distribution whose major component has mean vector $\mathbf{0}$ of length p and covariance matrix Σ_0. Mix 1 contained 5% normally distributed outliers with mean vector $\mathbf{10}$ of length p with covariance matrix Σ_0, mix 2 contained 5% outliers with mean vector $\mathbf{0}$ of length p with covariance matrix 3. Σ_0 and mix 3 contained 5% outliers with mean vector $\mathbf{10}$ of length p and covariance matrix 3, Σ_0.

Confirmatory factor model

All the population conditions in the simulation study were specified via a confirmatory three-factor model. An equation denoting this model is given by

$$\Sigma(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \Psi \tag{19}$$

where $\Sigma(\boldsymbol{\theta})$ is the working model covariance structure, $\boldsymbol{\Lambda}$ is a $p \times 3$ matrix of factor loadings, $\boldsymbol{\Phi}$ is a 3×3 matrix of factor correlations, and Ψ is a $p \times p$ diagonal matrix of measurement error variances of the manifest variables. As described in Table 3, we considered three conditions on the number of manifest variables ($p = 9$, 15, and 30). We set the error variances in Ψ to 1 and specified the factor correlations as

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & & \\ .5 & 1 & \\ .3 & .4 & 1 \end{bmatrix} \tag{20}$$

Three cross-factor loadings were in the population, but they were ignored in misspecified models. These paths are denoted a in Figure 1 and their values were set equal in the population. Larger values of a corresponded to more severe model misspecification.

The general factor loadings matrix can be expressed as

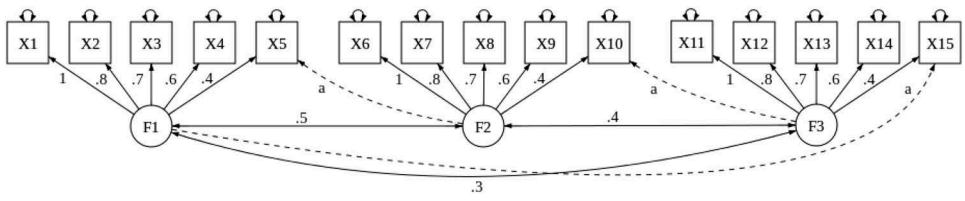


FIGURE 1 Path diagram for $p = 15$. *Note*. Dashed lines indicate paths for cross-loadings a that were omitted in misspecified models. Double-headed arrows above observed variables X1–X15 denote error variances set to 1. Double-headed arrows between factors F1 and F3 denote factor correlations. The factor loading values are noted on the left of single-headed arrows.

$$\mathbf{\Lambda} = \begin{bmatrix} c & a & \mathbf{0} \\ \mathbf{0} & c & a \\ a & \mathbf{0} & c \end{bmatrix} \quad (21)$$

where $\mathbf{0}$ denotes a vector of $p/3$ zeros and a is a vector of length $p/3$ whose last element equals a and the other $p/3 - 1$ elements are zero. For example, a population with nine manifest variables has $a = (0, 0, a)'$. The vector c has length $p/3$ with values that are adjusted according to the number of manifest variables. For a population with nine manifest variables, $c = (1, .8, .7)'$. For a population with 15 manifest variables, $c = (1, .8, .7, .6, .4)'$. For a population with 30 manifest variables,
$c = (1.00, .80, .70, .60, .40, .30, .35, .47, .58, .85)'$.

Estimation method

We employed Normal maximum-likelihood estimation as well as M-estimation using Huber-type weights, downweighting the 5% most extreme observations. M-estimation is a robust estimation method that weights observations according to how far they are from the center of the distribution of the sample. Observations that are far from the center get smaller weights than observations in the middle when estimating population means and covariances. Details of M-estimation in general can be found in Maronna, Martin, and Yohai (2006) and for SEM in particular in Yuan and Bentler (1998) and Yuan, Bentler, and Chan (2004).

Estimation of effect size measures

The purpose of the simulation study was to evaluate the definitions of the new effect size measures we proposed. To this end, we calculated the effect size measures in the following way. We generated data under H_0 according to the distribution specified in the simulation condition and setting a to 0 in the covariance matrix. For data under H_1, we modified the procedure so that the value of a varied from 0 to 1. Then, we used Fisher scoring to compute both ML and robust M-estimates and evaluated the resulting test statistics T_{ML} and T_{RML} under H_0 and H_1. To estimate $E(T_{ML}|H_0)$, $E(T_{ML}|H_1)$, $E(T_{RML}|H_0)$, and $E(T_{RML}|H_1)$, we averaged the respective test statistics across the converged replications.

RESULTS

Out of the 1,000 replications for each simulation condition, there were some nonconverged cases.[2] The worst case had only 804 converged replications. This corresponded to the simulation condition for estimating $E(T|H_1)$ under nonnormally distributed data with $p = 15$ and $N = 75$. All other conditions had at least 872 converged replications. For this article, we recorded the number of convergences and ignored the non-converging replications.

We evaluated the effect size measures according to their sensitivity to model misspecification and insensitivity to sample size, number of variables, and data distribution. First, we compared the relative performance of each effect size measure according to these features to determine which effect size measures are most promising. Then, we evaluated their performance in more depth, eliminating the other less promising candidates from further consideration.

Hierarchical regression approach

We used a hierarchical regression modeling approach to evaluate the performance of all the effect size measures. This method was used to determine how sample size, number of variables, data distribution, and estimation method influence the values of the effect size measures controlling for size of model misspecification. Specifically, our purpose was to determine how much variance in the effect size measure was accounted for by the predictor variables (i.e., N, p, etc.) by examining R^2_{adj}. The full model we considered for a single effect size measure can be written as

$$y_{full} = \beta_0 + \beta_1(\text{misspecification}) + \beta_2(N) + \beta_3(p) + \beta_4(\text{distribution}) + \beta_5(\text{estimationmethod}) + \epsilon$$

and the reduced model can be written as

$$y_{reduced} = \beta_0 + \beta_1(\text{misspecification}) + \epsilon$$

where N refers to the sample size and p is the number of variables. Population distribution and estimation method were treated as categorical variables in this analysis, while p, N, and model misspecification were treated as continuous. For the sake of simplicity, we write a single regression coefficient β_4 here for population distribution, but operationally, there are eight regression coefficients for this variable because the use of dummy coding is required for categorical predictors.

For each effect size measure, we calculated the full and reduced models and their $R^2_{adj.}$. We used $R^2_{adj.}$ instead of R^2 because $R^2_{adj.}$ contains a penalty to offset the effect of

[2] Cases were considered as nonconverged when the maximum iteration limit of 1000 was exceeded in the Fisher scoring algorithm to estimate the model parameters. This occurred when the difference between the model parameter estimates in the previous step and the current step exceeded 1×10^{-10} for all 1000 iterations.

TABLE 4
Hierarchical Regression Results

ES measure	$R^2_{adj.}(full)$	$R^2_{adj.}(reduced)$	$\Delta R^2_{adj.}$
$\mathcal{E}_1$	**0.8424**	**0.6788**	**0.1636**
$\mathcal{E}_2$	0.7678	0.2963	0.4715
$\mathcal{E}_3$	**0.9680**	**0.9254**	**0.0426**
$\mathcal{E}_4$	0.6629	0.3214	0.3414
$\mathcal{E}_5$	0.6975	0.2640	0.4335
$\mathcal{E}_6$	**0.9138**	**0.7737**	**0.1401**
$D_1(v1)$	0.6688	0.2665	0.4022
$D_1(v2)$	0.5719	0.2406	0.3313
$D_1(v3)$	0.6258	0.2568	0.3690
$D_2(v1)$	0.6721	0.2850	0.3871
$D_2(v2)$	0.5846	0.2312	0.3534
$D_2(v3)$	0.6432	0.2698	0.3734
$D_3(v1)$	0.5148	0.0826	0.4322
$D_3(v2)$	0.5276	0.0754	0.4522
$D_3(v3)$	0.6220	0.1756	0.4464
$D_4(v2)$	0.6165	0.1908	0.4257
$D_4(v3)$	0.61649	0.19080	0.42569
$D_5(v2)$	0.6062	0.2296	0.3766
$D_5(v3)$	0.5965	0.2335	0.3629

Note. $R^2_{adj.}$ for the full and reduced models of each effect size measure. Subscripts v_1, v_2, and v_3 denote version 1, version 2, and version 3 of effect size measures from the second class.

Bold rows correspond to the three effect size measures with the lowest $\Delta R^2_{adj.}$

including more variables in the model which artificially inflates the value of R^2 (Rencher & Schaalje, 2008).

As can be seen in Table 4, $\mathcal{E}_3$ performed the best out of all 19 effect size measures we considered. The full model that included sample size, number of variables, data distribution, and misspecification size accounted for 96.8% of the variance in the values of $\mathcal{E}_3$ while the reduced model that only considered misspecification size accounted for 92.5% of the variance. This is encouraging for two reasons. First, the sample size, number of variables, and data distribution did little to improve the prediction of the values of $\mathcal{E}_3$, indicating that this effect size measure may be most affected by the size of model misspecification and little affected by these other factors. These characteristics are needed in a good effect size measure. Second, since the full model accounted for 96.8% of the variance in values of $\mathcal{E}_3$, there was little influence on $\mathcal{E}_3$ from unidentified factors present in this study.

The other two bolded effect size measures in Table 4 ($\mathcal{E}_1$ and $\mathcal{E}_6$) also performed reasonably well. Between the two measures, $\mathcal{E}_6$ had a full model that explained more of its variance than $\mathcal{E}_1$, which was the main difference in their performance. The full model only explained 84.2% of the variance of $\mathcal{E}_1$; this suggests other factors not explicitly considered in our model substantially affected its value.

TABLE 5
Distance to F_0 for Three Effect Size Measures Closest to F_0

		Average distance from F_0					
		$\mathcal{E}_1$		$\mathcal{E}_3$		$\mathcal{E}_6$	
p	*Distribution*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*
	Normal	0.0349	0.0403	0.0060	0.0062	0.0080	0.0079
	t	0.2547	0.0958	0.0097	0.0076	0.1199	0.2018
	Exponential	0.0354	0.0356	0.0040	0.0047	0.0094	0.1006
	Elliptical	0.2605	0.1041	0.0100	0.0061	0.1192	0.2007
9	Non-normal	0.2436	0.0673	0.0082	0.0056	0.1089	0.2389
	Mix 1	0.0633	0.0643	0.0396	0.0334	0.0423	0.0618
	Mix 2	0.0821	0.0516	0.0060	0.0055	0.0328	0.0541
	Mix 3	0.1014	0.0575	0.0394	0.0298	0.0626	0.1268
	Uniform	0.0330	0.0364	0.0061	0.0062	0.0065	0.0073
	Normal	0.0420	0.0494	0.0064	0.0064	0.0080	0.0071
	t	0.3683	0.1239	0.0124	0.0070	0.1538	0.2700
	Exponential	0.0591	0.0267	0.0038	0.0038	0.0090	0.0963
	Elliptical	0.3696	0.1231	0.0119	0.0071	0.1539	0.2701
15	Non-normal	0.3516	0.0988	0.0077	0.0042	0.1392	0.3028
	Mix 1	0.1041	0.1040	0.0466	0.0451	0.0482	0.0549
	Mix 2	0.1052	0.0570	0.0074	0.0064	0.0408	0.0702
	Mix 3	0.1574	0.0896	0.0463	0.0415	0.0674	0.1165
	Uniform	0.0423	0.0444	0.0084	0.0079	0.0087	0.0068
	Normal	0.1173	0.1198	0.0028	0.0029	0.0034	0.0023
	t	0.4689	0.2021	0.0070	0.0036	0.1576	0.2852
	Exponential	0.1255	0.0856	0.0024	0.0026	0.0048	0.0580
	Elliptical	0.4690	0.2023	0.0094	0.0041	0.1576	0.2852
30	Non-normal	0.4641	0.2189	0.0056	0.0097	0.1480	0.3108
	Mix 1	0.3019	0.3027	0.0753	0.0764	0.0772	0.0812
	Mix 2	0.1871	0.1234	0.0032	0.0028	0.0372	0.0704
	Mix 3	0.3363	0.3006	0.0745	0.0882	0.1087	0.2150
	Uniform	0.1098	0.1099	0.0044	0.0045	0.0040	0.0043

Note. Absolute distances are averaged across model misspecification size and then sample size. $\mathcal{E}_1$, $\mathcal{E}_3$, and $\mathcal{E}_6$ are the three measures that are closest to F_0 in the most simulation conditions.

Comparison to F_0

Since the first class of effect size measures has some equivalency to F_0, we conducted a preliminary evaluation of these measures by their distance from F_0. Distance was calculated as the absolute difference between the value of an effect size measure and the value of F_0 for a specific level of model misspecification for each simulation condition. These absolute differences were then averaged across size of model misspecification and collapsed across sample size. All measures were ranked according to their distance from F_0 and their rankings recorded for each simulation condition. The measures that were most consistently ranked in the top 3 ($\mathcal{E}_1$, $\mathcal{E}_3$, and $\mathcal{E}_6$) are summarized in Table 5.

From Table 5, it is clear that $\mathcal{E}_3$ was consistently closer to F_0 than the other effect size measures in almost every simulation condition. $\mathcal{E}_6$ was only marginally better in conditions with uniform data and 15 or 30 manifest variables. Note that this may not be a fair comparison because F_0 assumes normality and its value may not be an

appropriate target for some statistics under non-normal conditions. Nevertheless, this comparison is informative as a preliminary tool to investigate the behavior of the effect size measures.

Most promising effect size measure

The pattern that emerges from Tables 4 and 5 is that $\mathcal{E}_3$ performed the best while $\mathcal{E}_1$ and $\mathcal{E}_6$ were not as satisfactory. In addition to having a larger ΔR^2_{adj}, they were not ranked in the top 3 effect size measures consistently across the simulation conditions. The other effect size measures considered in this study performed even worse. For the rest of our analysis, we considered only $\mathcal{E}_3$ and examined its properties. For the sake of convenience and clarity, we will refer to $\mathcal{E}_3$ for the remainder of this article as simply $\mathcal{E}$ (pronounced as "E").

Stability of $\mathcal{E}$

While a measure of model misspecification in the form of $[E(T|H_1) - E(T|H_0)]$ has been proposed in Yuan and Marshall (2004), $\mathcal{E}$ is new. Problematically, there is no established ideal to compare $\mathcal{E}$ against. Thus, we evaluated its stability by comparing its performance under various simulation conditions to its performance in a reference condition. We considered the reference to be the simulation condition with normally distributed data, $p = 9$, $N = 1000$, and using ML estimation. This condition was chosen because there was an adequate sample size, the number of manifest variables was small enough to avoid estimation

TABLE 6
Distance of $\mathcal{E}_i$ from $\mathcal{E}_r$

		Sample size N									
		75		*150*		*300*		*500*		*800*	
p	*Dist*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*
	Norm	0.0128	0.0138	0.0069	0.0076	0.0022	0.0025	0.0004	0.0005	0.0020	0.0024
	t	0.0242	0.0185	0.0134	0.0085	0.0113	0.0029	0.0028	0.0006	0.0025	0.0028
	Exp	0.0123	0.0106	0.0036	0.0049	0.0011	0.0014	0.0007	0.0002	0.0004	0.0005
	Ellip	0.0307	0.0079	0.0081	0.0082	0.0054	0.0029	0.0024	0.0006	0.0023	0.0028
9	Non	0.0167	0.0117	0.0115	0.0082	0.0058	0.0031	0.0027	0.0016	0.0008	0.0012
	Mix 1	0.0395	0.0347	0.0402	0.0351	0.0374	0.0308	0.0375	0.0301	0.0404	0.0335
	Mix2	0.0126	0.0098	0.0069	0.0067	0.0023	0.0020	0.0007	0.0006	0.0028	0.0028
	Mix 3	0.0395	0.0316	0.0400	0.0314	0.0373	0.0270	0.0377	0.0261	0.0406	0.0293
	Unif	0.0119	0.0123	0.0074	0.0073	0.0021	0.0022	0.0038	0.0039	0.0008	0.0008
	Norm	0.1030	0.1032	0.1178	0.1180	0.1202	0.1203	0.1173	0.1172	0.1196	0.1196
	t	0.1159	0.1040	0.1135	0.1174	0.1176	0.1205	0.1152	0.1162	0.1226	0.1198
	Exp	0.1189	0.1183	0.1215	0.1197	0.1220	0.1206	0.1214	0.1200	0.1202	0.1183
	Ellip	0.1108	0.1035	0.1076	0.1174	0.1176	0.1205	0.1154	0.1162	0.1181	0.1198
15	Non	0.1209	0.1208	0.1221	0.1175	0.1142	0.1204	0.1188	0.1218	0.1174	0.1202
	Mix 1	0.0950	0.0974	0.0979	0.1008	0.0984	0.1024	0.0929	0.0966	0.0963	0.0999
	Mix 2	0.1028	0.1040	0.1158	0.1175	0.1176	0.1190	0.1157	0.1165	0.1200	0.1199
	Mix 3	0.0938	0.1025	0.0975	0.1064	0.0975	0.1082	0.0926	0.1027	0.0968	0.1060
	Unif	0.1081	0.1027	0.1097	0.1096	0.1129	0.1130	0.1182	0.1182	0.1198	0.1198
	Norm	0.1282	0.1284	0.1312	0.1313	0.1282	0.1282	0.1281	0.1281	0.1300	0.1299
	t	0.1313	0.1259	0.1297	0.1310	0.1364	0.1272	0.1253	0.1277	0.1276	0.1297
	Exp	0.1322	0.1322	0.1344	0.1350	0.1285	0.1303	0.1332	0.1347	0.1324	0.1339
	Ellip	0.1398	0.1204	0.1282	0.1310	0.1265	0.1272	0.1254	0.1277	0.1277	0.1297
30	Non	0.1320	0.1170	0.1332	0.1313	0.1301	0.1263	0.1348	0.1319	0.1312	0.1309
	Mix 1	0.0902	0.0926	0.0945	0.0938	0.0913	0.0902	0.0913	0.0900	0.0927	0.0914
	Mix 2	0.1273	0.1292	0.1302	0.1309	0.1273	0.1277	0.1282	0.1284	0.1301	0.1299
	Mix 3	0.0911	0.0865	0.0946	0.0855	0.0910	0.0822	0.0909	0.0820	0.0926	0.0826
	Unif	0.1216	0.1210	0.1239	0.1239	0.1303	0.1303	0.1297	0.1297	0.1294	0.1294

Note. Distance of $\mathcal{E}_i$ from $\mathcal{E}_r$ for a given level of model misspecification was calculated as the absolute difference $|\mathcal{E}_i - \mathcal{E}_r|$. These absolute differences were averaged across misspecification size for each simulation condition to yield the distance values. Dist column contains the abbreviated names of distributions used in the simulation conditions, where "Norm" denotes the normal distribution condition, "Exp" denotes the exponential distribution condition, "Ellip" denotes the elliptical distribution condition, "Non" denotes the non-normal distribution condition, and "Unif" denotes the uniform distribution condition.

TABLE 7
Distance of $\mathcal{E}_i$ from $\mathcal{E}_r$ When Collapsing Across N

	Model size					
	$p = 9$		$p = 15$		$p = 30$	
Distribution	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*
Normal	0.0053	0.0057	0.1145	0.1145	0.1288	0.1288
t	0.0093	0.0071	0.1136	0.1139	0.1285	0.1280
Exponential	0.0035	0.0038	0.1191	0.1178	0.1322	0.1331
Elliptical	0.0097	0.0057	0.1110	0.1139	0.1276	0.1274
Non-normal	0.0079	0.0051	0.1168	0.1190	0.1328	0.1315
Mix 1	0.0390	0.0328	0.0953	0.0988	0.0922	0.0914
Mix 2	0.0055	0.0051	0.1131	0.1142	0.1284	0.1288
Mix 3	0.0389	0.0292	0.0946	0.1044	0.0924	0.0835
Uniform	0.0055	0.0056	0.1136	0.1129	0.1266	0.1266

problems, and ML estimation was appropriately paired with normally distributed data. We will denote the values of $\mathcal{E}$ under the reference condition as $\mathcal{E}_r$, and the values of $\mathcal{E}$ in other simulation conditions as $\mathcal{E}_i$, where i refers to a particular simulation condition.

Table 6 contains the distances of $\mathcal{E}_i$ from $\mathcal{E}_r$. Within each model size, there was only minor variation in the distance of $\mathcal{E}_i$ from $\mathcal{E}_r$. However, there was considerably more variation in $\mathcal{E}$ as the model size varied. The distances of $\mathcal{E}_i$ from $\mathcal{E}_r$ for a 9-variable model were fairly small from around 0.0008 to 0.0395. However, the distances of $\mathcal{E}_i$ from $\mathcal{E}_r$ for a 15- or 30-variable model fell roughly between 0.08 and 0.14.

Table 7 condenses the results shown in Table 6 by collapsing across sample size. Table 7 shows that there was more variation between model sizes than there was across distributions and estimation methods. When the model size increased from $p = 9$ to $p = 15$, the distances of $\mathcal{E}_i$ from $\mathcal{E}_r$ suddenly jumped from about 0.03 to about 0.11. However, when the model size increased from $p = 15$ to $p = 30$, the distances didn't increase as much.

Together, Tables 6 and 7 indicate that $\mathcal{E}$ was little affected by sample size and data distribution but was moderately affected by the number of variables. The difference between the value of $\mathcal{E}_i$ from that of $\mathcal{E}_r$ is as much as 0.1493 for the conditions studied when considering bigger model sizes. This is relatively large considering that the scale of $\mathcal{E}$ generally ran from 0 to 1 (maximum value obtained in this study was 1.01).

Figure 2 consists of several plots on the values of $\mathcal{E}$ across several simulation conditions for a sample size of 300. The values of $\mathcal{E}_i$ were compared to the values of $\mathcal{E}_r$, where the values of $\mathcal{E}_r$ are denoted by a solid line. The $\mathcal{E}_i$ values using ML and M-estimation were almost identical for the conditions studied. The $\mathcal{E}_i$ values across population distributions were also very similar. For models with 15 or 30 manifest variables, the $\mathcal{E}_i$ values were higher than the $\mathcal{E}_r$ values. This gap between the values of $\mathcal{E}_i$ and $\mathcal{E}_r$ became larger as the size of model misspecification increased.

Figure 3 shows the change in $\mathcal{E}$ as N varies for the simulation condition in which the values of $\mathcal{E}$ were furthest from the values of $\mathcal{E}_r$. In this scenario, increasing the sample size did little to improve the value. A sample size of 100 yielded comparable values to those from a sample size of 1000.

The values of the difference $\mathcal{E}_i - \mathcal{E}_r$ along with their relative frequencies are shown in Figure 4. For most simulation conditions, $\mathcal{E}_i$ was only slightly smaller than $\mathcal{E}_r$. In conditions where $\mathcal{E}_i$ values were far from $\mathcal{E}_r$, it was most common for $\mathcal{E}_i$ to be larger than $\mathcal{E}_r$. In these cases, $\mathcal{E}$ values may overrepresent rather than underrepresent the size of model misspecification.

FOLLOW-UP INVESTIGATION OF THE BEST EFFECT SIZE MEASURE

In this section, we extend our investigation of $\mathcal{E}$. First, we untangle the effect of model size on $\mathcal{E}$. Then, we provide rough guidelines for cutoff values. Finally, we demonstrate the performance of $\mathcal{E}$ and its confidence intervals on two real datasets.

The effect of model size

We have stated that $\mathcal{E}$ appears to be influenced by model size based on the results of our simulation study. However, considering three model sizes is not adequate to infer the exact nature of the relationship. To remedy this, we considered the effect of a wider range of model sizes (p = 9, 15, 30, 45, 60, and 90) on the definition analytically. For this purpose, an analytical approach may be sufficient to determine the effect of p holding other factors constant.

Building from Equation 11, we have

$$\mathcal{E} = \left(\frac{(N-1)F_{ML}(\Sigma_0, \Sigma(\boldsymbol{\theta})) - (N-1)F_{ML}(\Sigma_{H_0}, \Sigma(\boldsymbol{\theta}))}{N-1} \right)^{1/2} \tag{22}$$

Figure 5 shows the results graphically. Holding a constant, we expected horizontal lines indicating that the values of $\mathcal{E}$ remained unchanged when varying p. For $a \leq .5$, there was only minor fluctuation. For larger a and models with less than 45 variables, the fluctuation of $\mathcal{E}$ values was more pronounced. Once $p \geq 45$, the values of $\mathcal{E}$ were stable at somewhat higher magnitudes.

Based on these results, $\mathcal{E}$ seems to be moderately impacted by model size when there is substantial model misspecification. Under these conditions and when $p \geq 45$, the values of $\mathcal{E}$ were larger. The practical impact of this may be minimal with a suitable choice of cutoff values. We

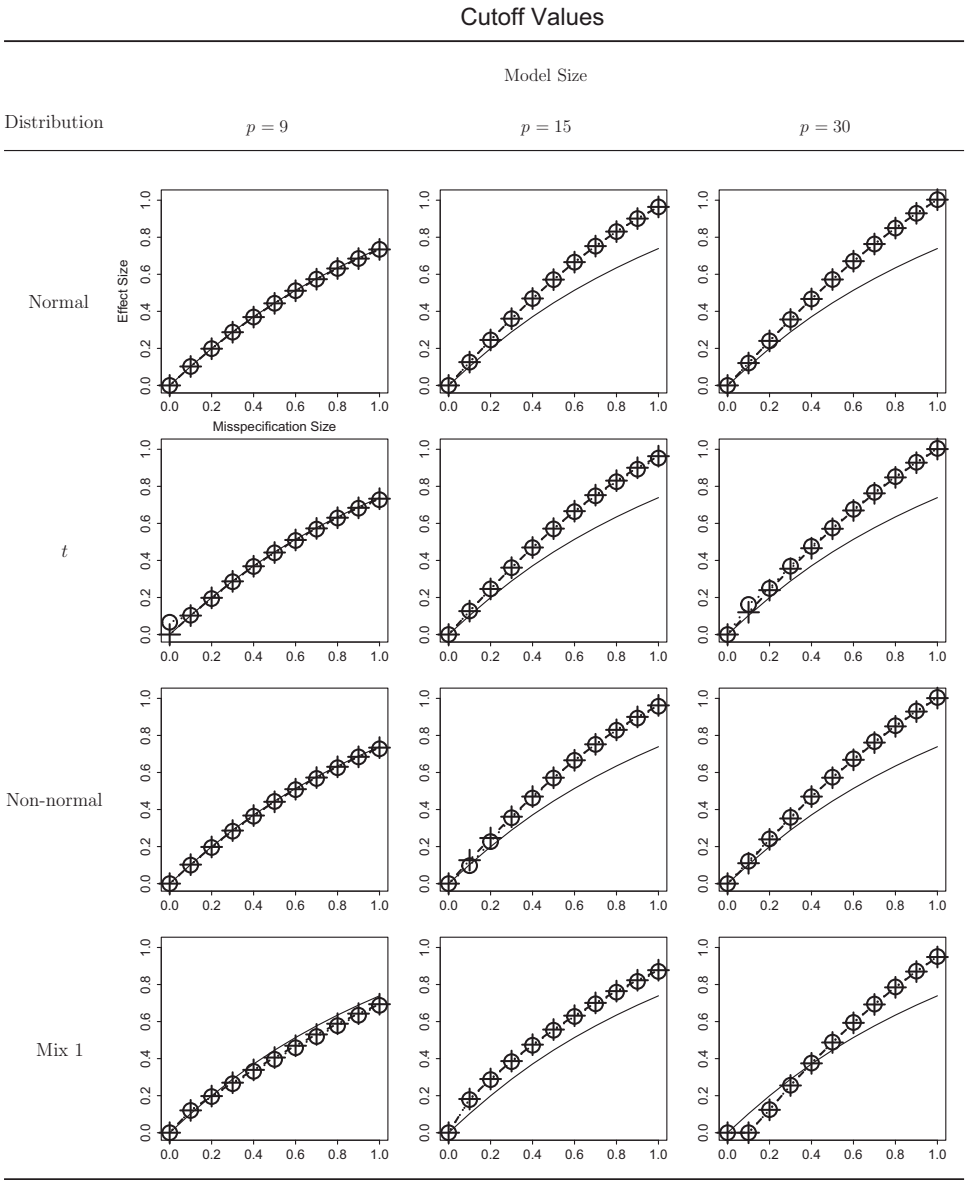


FIGURE 2 Performance of $\mathcal{E}_i$ compared to $\mathcal{E}_r$. *Note*. $N = 300$, the line with circles denotes the values of $\mathcal{E}_i$ under ML estimation while the line with crosses denotes M-estimation. The solid line corresponds to the values of $\mathcal{E}_r$.

discuss such a choice of cutoff values in the following section.

Cutoff Values

This study has focused primarily on the definitions of new effect size measures and their performance in conditions where the population values were known. However, we recognize that there is a practical need for cutoff values of effect size measures. This is more of a substantive issue than a statistical one, but we will present preliminary recommendations based on our findings and current conventions.

Since we have measured the severity of model misspecification by the size of omitted cross-loadings, a natural approach to establishing cutoff values of $\mathcal{E}$ is to translate cutoffs used to assess practical significance of factor loadings in exploratory factor analysis (EFA). Stevens (1992) suggested using a cutoff of 0.4 for interpretation purposes while Comrey and Lee (1992) suggested cutoffs ranging from 0.32 (poor), 0.45 (fair), 0.55 (good), 0.63 (very good), to 0.71 (excellent). That is, standardized loadings that are less than the chosen cutoff contribute very little to the factor and should be dropped. Analogously in CFA, an omitted factor loading with the same magnitude should correspond to minor model

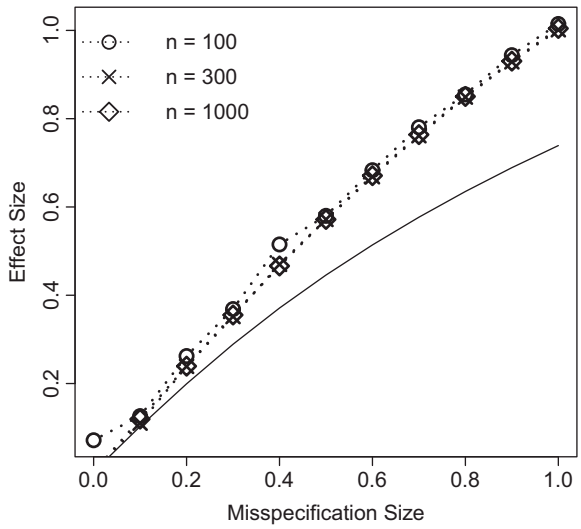


FIGURE 3 The effects of increasing N on the values of $\mathcal{E}$. *Note*. The effects of increasing N on the values of $\mathcal{E}$ for $p = 30$, M-estimation, and non-normal data. The solid line denotes the values of $\mathcal{E}_r$. Model misspecification is quantified by the size of omitted cross-factor loadings a.

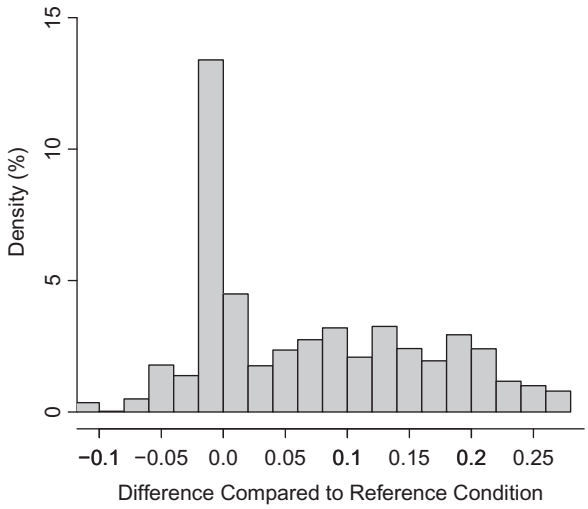


FIGURE 4 Distribution of $\mathcal{E}_i - \mathcal{E}_r$. *Note*. Distribution of the differences $\mathcal{E}_i - \mathcal{E}_r$, where $\mathcal{E}_i - \mathcal{E}_r$ denotes the average difference between $\mathcal{E}_i$ and $\mathcal{E}_r$ when holding model misspecification size constant.

misspecification. We will incorporate both suggestions in our choice of cutoff thresholds.

The average values of $\mathcal{E}$ obtained in the simulation study for each misspecification size (unstandardized a and standardized a_s) are presented in Table 8. Cutoff thresholds are displayed in Table 9 and were chosen according to the suggestions discussed in the previous paragraph and the misspecification sizes considered in the study. $\mathcal{E}<.42$ can be regarded as very small since this corresponds to a standardized omitted factor loading of approximately 0.3 or less. Small effect sizes have values between 0.42 and 0.6, based on the criteria suggested by Stevens (1992) for interpretable factor loadings. Medium effect sizes have values between 0.6 and 0.82, where 0.82 corresponds to an omitted factor loading of approximately 0.55 [“good” according to Comrey and Lee (1992)]. A large effect size exceeds 0.82 in value.

These four cutoff categories have two advantages: (1) A “very small” effect size category provides a bonus

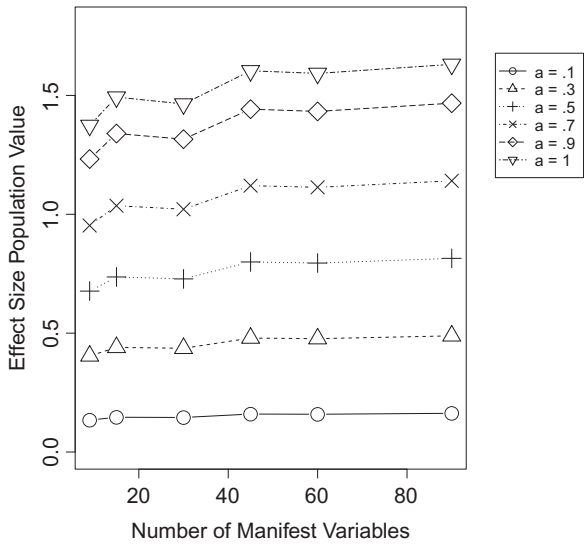


FIGURE 5 Effect of p on $\mathcal{E}$. *Note*. The effect of p on population values of $\mathcal{E}$ holding misspecification size constant.

TABLE 8
$\mathcal{E}$ Values by Misspecification Size

a	a_s	*Mean*	*SD*
0.0	0.0	0.0024	0.0111
0.1	0.0820	0.1121	0.0386
0.2	0.1597	0.2216	0.0378
0.3	0.2323	0.3265	0.0426
0.4	0.2991	0.4248	0.0529
0.5	0.3600	0.5164	0.0647
0.6	0.4150	0.6017	0.0770
0.7	0.4644	0.6807	0.0889
0.8	0.5087	0.7536	0.0997
0.9	0.5482	0.8205	0.1093
1.0	0.5835	0.8815	0.1182

Note. a_s denotes the standardized loadings that correspond to a. The mean and standard deviation columns refer to $\mathcal{E}$ values.

TABLE 9
$\mathcal{E}$ Cutoff Values

Very small	*Small*	*Medium*	*Large*
<.42	0.42–0.6	0.6–0.82	>.82

reward for excellent models, since models with minimal misspecification are of great interest. (2) Most of the fluctuations in $\mathcal{E}$ values due to model size occur in the "medium" and "large" effect size categories, where the stakes are lower.

Let us caution that these cutoff guidelines are preliminary and should be treated as rules of thumb. There is necessarily some level of arbitrariness in any approach to cutoff values. However, a comprehensive investigation is warranted to refine the thresholds so they reflect the distribution of effect sizes in empirical studies using SEM. Field-specific norms should also be developed, since measurement error can be unavoidably larger for some research areas. In addition, more types of model structures should be considered. The thresholds in Table 9 are intended to serve in the interim.

Applied examples

We will demonstrate the practical performance of $\mathcal{E}$ on two real datasets in this section. The first example is a dataset with distributional assumption violations and shows the performance of $\mathcal{E}$ under moderate model misspecification. The second example demonstrates the performance of $\mathcal{E}$ under very little model misspecification. In both examples, we provide confidence intervals to illustrate the actual performance of the bootstrap confidence intervals discussed previously. We will implement percentile intervals, BC intervals, and BCa intervals with 90% and 95% confidence levels.

Example 1: KIMS: Baer, Smith, and Allen (2004) developed the Kentucky Inventory of Mindfulness Skills (KIMS) containing four subscales: observing, describing, acting with awareness, and accepting without judgment. The inventory consists of 39 items rated on a 5-point Likert scale. Publicly available data from openpsychometrics.org was obtained from 601 participants who agreed to share their responses for research purposes. Histograms

of the item responses suggested moderately skewed distributions for many items.

While a scree plot indicated a preference for three factors, EFA showed that three factors had cumulative variance of 0.366 (sum of the first three eigenvalues of the sample correlation matrix divided by 39), while four factors had 0.437. We conducted an EFA using the psych package in R, extracting four factors using promax rotation. Our results mostly agreed with the original study with the exception of item 11. In our analysis, item 11 was sorted into the observing subscale while it was originally part of the acting with awareness subscale. Like the original study, our analysis indicated several potential cross-loadings (Baer et al., 2004).

We fit two models for our demonstration. Model 1 was a four-factor model based on the original subscales with no cross-loadings and Model 2 was a four-factor model based on our EFA results (four factors, with item 11 on the observing subscale and cross-loadings with magnitude greater than 0.23). The test statistics and fit indices in Table 10 provided a conflicted view of model fit. The Comparative Fit Index (CFI), Tucker Lewis Index (TLI), and RMSEA slightly preferred Model 2 (although CFI and TLI indicate a poor model fit in both cases) while the Standardized Root Mean Square Residual (SRMR) preferred Model 1. As is typical with real data, it is unclear which fit indices to trust and which model should be preferred.

$\hat{\mathcal{E}}$ values in Table 11 suggested substantial model misspecification for both models. Referencing Table 9 from the previous section, the 90% and 95% confidence intervals for both models fell entirely within the "large" effect size category. The widths of the percentile intervals, BC intervals, and BCa intervals were fairly narrow and did not differ much from each other. Increasing the confidence level from 90% to 95% did not widen the confidence intervals substantially for either model. Overall, $\hat{\mathcal{E}}$ indicated that there was less misspecification in Model 2 which matched our expectations based on the EFA results. However, neither model was very good.

TABLE 10
Example 1 Test Statistics and Fit Indices

Model	T_{ML}	*df*	*CFI*	*TLI*	*RMSEA*	*RMSEA CI*	*SRMR*
Model 1	2,451.289	696	0.830	0.819	0.065	(0.062, 0.068)	0.074
Model 2	2,380.451	694	0.837	0.826	0.064	(0.061, 0.066)	0.085

Note. RMSEA CI denotes 90% confidence intervals for RMSEA.

TABLE 12
Example 2 Test Statistic and Fit Indices

T_{ML}	*df*	*CFI*	*TLI*	*RMSEA*	*RMSEA CI*	*SRMR*
2.097	4	1	1.025	0	(0, 0.118)	0.019

Note. RMSEA CI denotes 90% confidence intervals for RMSEA.

Example 2: Open–closed book data: This classic dataset from Table 1.2.1 in Mardia, Kent, and Bibby (1979) contains test scores from 88 students on 5 subjects: mechanics, vectors, algebra, analysis, and statistics. The first two subjects were closed-book exams while the last three were open-book exams. We fit the tried-and-true two-factor model with open book and closed book factors originally proposed by Tanaka, Watadani, and Moon (1991). The test statistic and fit indices in Table 12 did not reject the model.

We obtained $\hat{\mathcal{E}} = .155$ which indicated a "very small" effect size according to Table 9. The 90% and 95% confidence intervals in Table 13 also indicated a "very small" effect size. The average width of the 90% confidence intervals was 0.295 while the average width of the 95% confidence intervals was 0.336. This is not a very large difference. Overall, the model appears to perform extremely well and had little misspecification.

A NOTE ON FIT INDICES

In the introduction, we noted that fit indices cannot be used as effect size measures while they are used in hypothesis tests. In

TABLE 11
Example 1 Results

	Model 1, $\hat{\mathcal{E}} = 2.021$		*Model 2,* $\hat{\mathcal{E}} = 1.607$	
Type	90%	95%	90%	95%
Percentile interval	(1.918, 2.138)	(1.905, 2.160)	(1.491, 1.740)	(1.473, 1.767)
BC interval	(1.925, 2.145)	(1.909, 2.165)	(1.493, 1.747)	(1.477, 1.771)
BC α (α_1, α_2)	(0.0625, 0.960)	(0.032, 0.981)	(0.058, 0.957)	(0.030, 0.979)
BCa interval	(1.929, 2.152)	(1.914, 2.181)	(1.498, 1.755)	(1.483, 1.781)
BCa α (α_1, α_2)	(0.072, 0.968)	(0.040, 0.986)	(0.068, 0.965)	(0.038, 0.985)

Note. These confidence intervals were based on a nested bootstrap with $B_1 = 800$ and $B_0 = 200$.

TABLE 13
Example 2 Results

Level	*Percentile*	*BC*	*BC* α (α_1, α_2)	*BCa*	*BCa* α (α_1, α_2)
90%	(0.021, 0.313)	(0.027, 0.322)	(0.063, 0.960)	(0.029, 0.328)	(0.066, 0.963)
95%	(0.009, 0.343)	(0.016, 0.351)	(0.032, 0.981)	(0.017, 0.356)	(0.035, 0.983)

Note. "Level" denotes the confidence level of the confidence intervals. The intervals were based on a nested bootstrap with $B_1 = 800$ and $B_0 = 200$.

TABLE 14
Performance of Popular Fit Indices for $a = .4$

			CFI		*TLI*		*RMSEA*		*SRMR*	
p	*N*	*Dist*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*	*ML*	*M-est*
	75	Norm	0.9334	0.9315	0.9001	0.8973	0.0678	0.0691	0.0761	0.0781
	75	*t*	0.8609	0.9055	0.7914	0.8583	0.1142	0.0854	0.0929	0.1689
	75	Ellip	0.8609	0.9055	0.7914	0.8583	0.1142	0.0854	0.0929	0.1689
	75	Non	0.8708	0.9341	0.8062	0.9011	0.1090	0.0695	0.0904	0.2152
	300	Norm	0.9314	0.9310	0.8971	0.8964	0.0761	0.0763	0.0556	0.0563
9	300	*t*	0.8972	0.9246	0.8458	0.8869	0.0957	0.0801	0.0644	0.1621
	300	Ellip	0.8972	0.9246	0.8458	0.8869	0.0957	0.0801	0.0644	0.1621
	300	Non	0.9056	0.9387	0.8584	0.9080	0.0920	0.0731	0.0632	0.2201
	800	Norm	0.9344	0.9344	0.9016	0.9017	0.0745	0.0745	0.0499	0.0501
	800	*t*	0.9168	0.9324	0.8752	0.8986	0.0847	0.0759	0.0543	0.1565
	800	Ellip	0.9168	0.9324	0.8752	0.8986	0.0847	0.0759	0.0543	0.1565
	800	Non	0.9196	0.9366	0.8794	0.9048	0.0838	0.0749	0.0541	0.2223
	75	Norm	0.8984	0.8954	0.8774	0.8738	0.0569	0.0580	0.0925	0.0929
	75	*t*	0.7256	0.8358	0.6688	0.8018	0.1144	0.0784	0.1174	0.1743
	75	Ellip	0.7256	0.8358	0.6688	0.8018	0.1144	0.0784	0.1174	0.1743
	75	Non	0.7796	0.8863	0.7341	0.8628	0.1005	0.0628	0.1089	0.1844
	300	Norm	0.9226	0.9216	0.9065	0.9054	0.0515	0.0518	0.0644	0.0646
15	300	*t*	0.8430	0.9063	0.8105	0.8869	0.0785	0.0574	0.0769	0.1539
	300	Ellip	0.8430	0.9063	0.8105	0.8869	0.0785	0.0574	0.0769	0.1539
	300	Non	0.8530	0.9230	0.8226	0.9071	0.0761	0.0525	0.0759	0.1855
	800	Norm	0.9264	0.9262	0.9112	0.9109	0.0502	0.0503	0.0558	0.0559
	800	*t*	0.8829	0.9209	0.8587	0.9045	0.0654	0.0522	0.0640	0.1474
	800	Ellip	0.8829	0.9209	0.8587	0.9045	0.0654	0.0522	0.0640	0.1474
	800	Non	0.8911	0.9290	0.8686	0.9143	0.0629	0.0502	0.0616	0.1813
	75	Norm	0.8152	0.8137	0.8000	0.7985	0.0558	0.0561	0.0952	0.0953
	75	*t*	0.5685	0.7024	0.5331	0.6780	0.1080	0.0775	0.1248	0.1583
	75	Ellip	0.5685	0.7024	0.5331	0.6780	0.1080	0.0775	0.1248	0.1583
	75	Non	0.5874	0.7562	0.5535	0.7362	0.1053	0.0686	0.1222	0.1643
	300	Norm	0.9536	0.9527	0.9497	0.9488	0.0254	0.0257	0.0528	0.0528
30	300	*t*	0.7847	0.9215	0.7670	0.9151	0.0622	0.0341	0.0727	0.1310
	300	Ellip	0.7847	0.9215	0.7670	0.9151	0.0622	0.0341	0.0727	0.1310
	300	Non	0.7931	0.9464	0.7761	0.9420	0.0613	0.0279	0.0719	0.1474
	800	Norm	0.9606	0.9604	0.9574	0.9571	0.0236	0.0237	0.0386	0.0387
	800	*t*	0.8624	0.9491	0.8511	0.9449	0.0470	0.0271	0.0526	0.1246
	800	Ellip	0.8624	0.9491	0.8511	0.9449	0.0470	0.0271	0.0526	0.1246
	800	Non	0.8749	0.9593	0.8646	0.9560	0.0445	0.0244	0.0500	0.1387

Note. Values of popular fit indices for a fixed level of model misspecification ($a = .4$). Dist column contains the abbreviated names of distributions used in the simulation conditions, where "Norm" denotes the normal distribution condition, "Ellip" denotes the elliptical distribution condition, and "Non" denotes the non-normal distribution condition.

addition to this point, we stated that their performance is unstable and affected by factors unrelated to model fit. In Table 14, we briefly demonstrate this by showing the performance of CFI, TLI, RMSEA, and SRMR under some of the conditions considered in our simulation study.

Although the model misspecification was held constant at $a = .4$, the values of every fit index ranged from poor to good fit according to the cutoff guidelines established in Hu and Bentler (1999). This was true even within model sizes. Based on these results, we do not advocate using these fit indices as effect size measures even if their use in hypothesis testing is discontinued.

DISCUSSION

The results of the Monte Carlo simulation study indicate that $\mathcal{E}$ ($\mathcal{E}_3$ in the first class of effect size measures) is a useful effect size measure. Its performance on two real datasets echoes the results of the simulation study. The analyses we performed imply that (1) $\mathcal{E}$ is little influenced by sample size, distribution, or estimation method; (2) the range of values of $\mathcal{E}$ reflect the severity of model misspecification; and (3) while $\mathcal{E}$ is somewhat influenced by model size, the practical impact of this relationship may be reduced by choosing suitable cutoff values.

Thus, generalizing Cohen's d to SEM has yielded an effect size measure that has several desirable properties. However, this approach has also produced effect size measures that do not have desirable properties. Overall, the first class of effect size measures performed better than the second class. While the second class of effect size measures was a stricter translation of Cohen's d, the first class of effect size measures had some equivalency to F_0. These may be important distinguishing traits and may inform future researchers in this area.

Another feature of this study was that it compared effect size measures that did not make any distributional assumptions in their definitions to effect size measures that did. Results of the simulation study showed that the effect size measures that relied on distributional assumptions in their definitions performed badly. Avoiding these assumptions when developing new statistical measures may be beneficial.

Limitations

This study has only evaluated the performance of effect size measures with complete data and the performance of $\mathcal{E}$ with incomplete data is unknown. The simulation study was limited to three-factor models with the number of manifest variables uniformly increasing among the factors. The best effect size measure $\mathcal{E}$ is also somewhat impacted by model size.

Future directions

Future directions include further investigation of $\mathcal{E}$ under more types of model complexity and different numbers of manifest variables. More rigorous cutoff thresholds should also be developed based on an extensive literature review of empirical studies that use SEM. The performance of $\mathcal{E}$ with incomplete data should also be investigated.

CONCLUSION

While the purpose of $\mathcal{E}$ is not to replace test statistics or fit indices, it does complement statistical analysis in applications. It provides additional information that cannot be gained from existing measures in SEM. Supplemental information to hypothesis tests that capture the size of the misspecification without being subject to the need of controlling Type I errors is extremely valuable. A model with small misspecification can still be rejected in a hypothesis test but can be rewarded by the supplemental effect size $\mathcal{E}$. For these reasons, we hope that the topic of effect size in SEM will become a more active area of research. For now, the development of $\mathcal{E}$ is a promising start.

FUNDING

This work was supported by the National Science Foundation: [Grant Number SES-1461355].

REFERENCES

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206. doi:10.1177/1073191104268029

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*(1), 41–75. doi:10.1207/s15328007sem1201_3

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88* (3), 588–606. doi:10.1037/0033-2909.88.3.588

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, *47*(1), 563–592. doi:10.1146/annurev.psych.47.1.563

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197. doi:10.1207/S15327906Mb340203

Beran, R., & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, *13* (1), 95–115. doi:10.1214/aos/1176346579

Cheng, C., & Wu, H. (2017). Confidence intervals of fit indexes by inverting a bootstrap test. *Structural Equation Modeling*, *24*(6), 870–880. doi:10.1080/10705511.2017.1333432

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd* ed.). Malwah, NJ: Erlbaum.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Eribaum.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(3), 343–367. doi:10.1207/s15328007sem1203_1

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*(3), 356–410. doi:10.1207/S15328007SEM0703_2

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362. doi:10.1037/0033-2909.112.2.351

Kano, Y. (1986). Conditions on consistency of estimators in covariance structure model. *Journal of the Japan Statistical Society*, *16*(1), 75–80.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York, NY: Wiley.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York, NY: Academic Press.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York, NY: Wiley.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. doi:10.1207/s15328007sem1103_2

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. doi:10.1007/s11336-016-9552-7

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi:10.1037/0033-2909.105.1.156

Moshagen, M., & Auerswald, M. (2017). On congruence and incongruence of measures of fit in structural equation modeling. *Psychological methods*. *23*(2), 318–336.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478. doi:10.1207/S15327906MBR3903_3

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. New York, NY: Wiley.

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. doi:10.3168/jds.S0022-0302(88)79586-7

Savalei, V. (2008). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling*, *15* (1), 1–22. doi:10.1080/10705510701758091

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Shapiro, A. (1984). A note on consistency of estimators in the analysis of moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 84–88. doi:10.1111/j.2044-8317.1984.tb00790.x

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling*, *25*(1), 21–40. doi:10.1080/10705511.2017.1369088

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*(1), 134–146. doi:10.2307/1130296

Tanaka, Y., Watadani, S., & Moon, S. H. (1991). Influence in covariance structure analysis: With an application to confirmatory factor analysis. *Communications in Statistics-Theory and Methods*, *20*(12), 3805–3821. doi:10.1080/03610929108830742

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. doi:10.1007/BF02291170

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.

Yuan, K. H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika*, *31*(1), 67–90. doi:10.2333/bhmk.31.67

Yuan, K.-H., & Bentler, P. M. (1998). Structural equation modeling with robust covariances. In A. E. Raftery (Ed.), *Sociological methodology 1998* (pp. 363–396). Boston, MA: Blackwell Publishers.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115–148. doi:10.1207/s15327906mbr4001_5

Yuan, K.-H., & Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica*, *9*, 831–853.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. doi:10.1007/BF02295644

Yuan, K.-H., Jiang, G., & Yang, M. (2018). Mean and mean-and-variance corrections with big data. *Structural Equation Modeling*, *25*(2), 214–229. doi:10.1080/10705511.2017.1379012

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling*, *23*(3), 392–408. doi:10.1080/10705511.2015.1118692

APPENDIX

This appendix provides the details on obtaining point estimates and confidence intervals for the effect size measures. We focus on $\mathcal{E}_3$ since the procedure is parallel for the other effect size measures. As in the main body of the article, we will denote $\mathcal{E}_3$ as $\mathcal{E}$ here. To obtain a point estimate of $\mathcal{E}$, the procedure in Section 3 of Yuan and Marshall (2004) is slightly modified and outlined below:

- Estimate $E(T_{ML}|H_1)$ by T_{ML}
- Estimate $E(T_{ML}|H_0)$:
  - Choose an admissible $\tilde{\boldsymbol{\theta}}$, such as the ML estimate $\hat{\boldsymbol{\theta}}$.
  - Get the model-implied covariance $\tilde{\Sigma} = \Sigma(\tilde{\boldsymbol{\theta}})$.
  - Transform the sample data x_i to $y_i = \Sigma^{1/2}(\boldsymbol{\theta})\mathbf{S}^{-1/2}x_i$ for $i = 1, 2, ..., n$.
  - Take B_0 bootstrap samples from $\mathbf{y} = (y_1, y_2, ..., y_n)$.
  - Calculate T_{ML} for each of the B_0 samples, denoted as T_b^*.
  - Estimate $E(T_{ML}|H_0)$ by $\bar{T}_{ML}^* = \sum_{b=1}^{B_0} T_b^*/B_0$.
- $\hat{\mathcal{E}} = \left(\frac{T_{ML} - \bar{T}_{ML}^*}{N-1}\right)^{1/2}$

To create a confidence interval, another layer of bootstrap is added to the process we described above. This corresponds to Algorithm II from Yuan and Marshall (2004):

- Generate B_1 bootstrap samples from $\mathbf{x} = (x_1, x_2, ..., x_n)$ denoted $x_b^* = (x_1^*, x_2^*, ..., x_n^*)$.
- Calculate T_{ML} for each bootstrap sample denoted T_b^*.
- Choose any admissible $\tilde{\boldsymbol{\theta}}$.
- Get the model-implied covariance $\tilde{\Sigma} = \Sigma(\tilde{\boldsymbol{\theta}})$.
- Transform each bootstrap sample x_b^* into $y_b^* = (y_1^*, ..., y_n^*)$ where $y_i^* = \Sigma^{1/2}(\tilde{\boldsymbol{\theta}})\mathbf{S}_b^{*-1/2}x_i^*$ for $i = 1, 2, ..., n$ and where $\mathbf{S}_b^*$ is the covariance matrix of x_b^*.
- Take B_0 bootstrap samples from $y_b^* = (y_1^*, ..., y_n^*)$ denoted as $y_{bj}^{**} = (y_1^{**}, y_2^{**}, ..., y_n^{**})$.
- Calculate T_{ML} for each of the y_{bj}^{**} bootstrap samples denoted as T_{bj}^{**}.
- Take the average of the second-layer test statistics $\bar{T}_b^{**} = \sum_{j=1}^{B_0} T_{bj}^{**}/B_0$.
- Obtain B_1 estimates of $\hat{\mathcal{E}}_b^* = (\frac{T_b^* - \bar{T}_b^{**}}{N-1})^{1/2}$.
- Rank the $\hat{\mathcal{E}}_b^*$ estimates and use them to construct percentile, BC, and BCa confidence intervals as usual (see Efron & Tibshirani, 1994).