

ASSESSING THE SIZE OF MODEL MISFIT IN STRUCTURAL EQUATION MODELS

ALBERTO MAYDEU-OLIVARES 

UNIVERSITY OF BARCELONA

UNIVERSITY OF SOUTH CAROLINA

When a statistically significant mean difference is found, the magnitude of the difference is judged qualitatively using an effect size such as Cohen's d . In contrast, in a structural equation model (SEM), the result of the statistical test of model fit is often disregarded if significant, and inferences are drawn using "close" models retained based on point estimates of sample statistics (goodness-of-fit indices). However, when a SEM cannot be retained using a test of exact fit, all substantive inferences drawn from it are suspect. It is therefore important to determine the size of the model misfit. Standardized residual covariances and residual correlations provide standardized effect sizes of the misfit of SEM models. They can be summarized using the Standardized Root Mean Squared Residual (SRMSR) and the Correlation Root Mean Squared Residual (CRMSR) which can be used as overall effect sizes of the misfit. Statistical theory is provided that allows the construction of confidence intervals and tests of close fit based on the SRMSR and CRMSR. It is hoped that the use of standardized effect sizes of misfit will help reconcile current practices in SEM and elsewhere in statistics.

Key words: goodness-of-fit, RMSEA, effect size.

Goodness-of-fit assessment refers to how well our model reproduces the data-generating process. When fitting a regression model, researchers assess the goodness-of-fit of their model by inspecting a plot of the standardized residuals versus standardized fitted values in order to check the linearity and homoscedasticity assumption of the model. Failure to meet the model assumptions will likely result in incorrect parameter estimates. Simply put, when the fitted model is incorrectly specified, all inferences based on it are suspect.

In structural equation modeling (SEM), we fit a system of regression-like equations to our data (possibly involving latent variables), and procedures similar to those used in regression can be used to assess each of the equations in the model (Bollen & Arminger, 1991; Coffman & Millsap, 2006; Hildreth, 2013; Yuan & Hayashi, 2010). However, they hardly seem to be used in applications. Rather, for historical reasons, the assessment of model fit in structural equation models has relied on the use of grouped data (residual means, covariances, and correlations) as opposed to individual residual observations. There is evidence that residual covariances (i.e., the differences between observed and fitted covariances) are sensitive to linear mis-specifications (Raykov, 2000) but I am not aware of any research that has examined whether they are sensitive to the presence of heteroscedasticity. Residual summary statistics such as residual covariances can be summarized into a single test statistic to assess the overall goodness-of-fit of a structural equations model. Although overall test statistics are invariably reported in applications, they are most often ignored as they usually suggest that the model is incorrectly specified. Rather,

Presidential Address to the Psychometric Society, delivered at the annual meeting in Madison (WI), July 2014. This research was supported by an ICREA-Academia Award and Grant SGR 2014 1500 from the Catalan Government and Grant PSI2012-33601 from the Spanish Ministry of Education. I am indebted to Peter Bentler, Ke-Hai Yuan, Albert Satorra, Jim Steiger, Haruhiko Ogasawara, and Yves Rosseel for their helpful comments. I am also most thankful to Yves Rosseel for implementing these methods in the Lavaan package in R.

Correspondence should be made to Alberto Maydeu-Olivares, Department of Psychology, University of South Carolina, Barnwell College, 1512 Pendleton St., Columbia, SC 29208, USA. Email: amaydeu@sc.edu

current practice in SEM relies on the use of goodness-of-fit indices, summary statistics that are provided in a descriptive fashion, along with some cutoff values that suggest a “good” fit. In so doing, I feel that these current practices in SEM contradict the most basic statistical principles, as taught in introductory statistics courses. In fact, I believe that assessing the fit of a structural equation model is not fundamentally different from assessing a mean difference using a z test.

The aim of this presentation is to provide specific links between our practices in introductory statistics and in SEM in the hope that it helps to standardize current practices. To simplify the presentation, I discuss solely the frequentist approach. Also, I focus on “classical” structural equation models, in other words, on covariance structure models.

1. Why Should We Assess the Goodness-of-Fit of a SEM Model?

What do we gain by properly examining the fit of a SEM model? And what do we lose by not doing so? To answer these questions, it is best to use a simple model. Consider the mediation model involving variables Y , X and M sketched in Fig. 1 as Model 0. Edwards and Lambert (2007) discussed how to use regression analysis to estimate the moderating effects of a fourth variable (variable z in Fig. 1). Consider Model 0, in Fig. 1. Moderation may be present only in path a , leading to Model a , or only in path b , leading to Model b , etc. Assuming that Model 0 holds, there are seven possible moderated mediation models with these four variables (in addition to Model 0, no moderation). These models can be estimated by including the interactions ZX and ZM in the model leading to the eight models and their degrees of freedom sketched in Fig. 1.

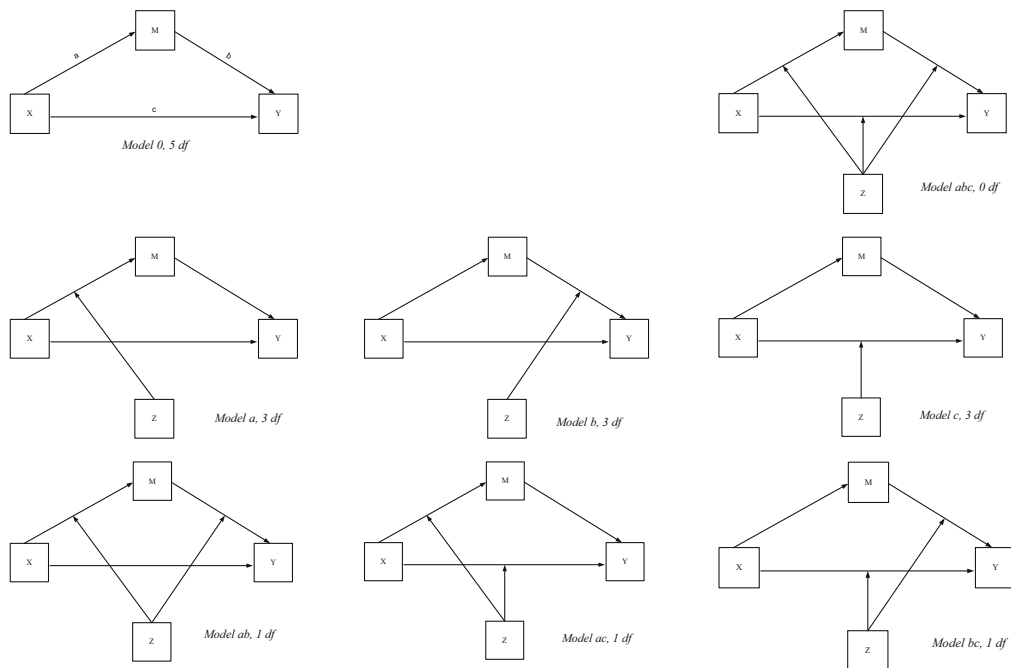


FIGURE 1.
Sketch of moderated mediation models.

The purpose of the analysis is to estimate the mediation parameter (indirect effect) $a \times b$. Because of the nature of the models involved, both regression and SEM provide the same parameter estimates for any of the models depicted. However, SEM methods provide an overall goodness of fit of the model being fitted, something not available using the equation by equation regression methods. What is the use of these goodness-of-fit tests? One could argue that they are not needed and that all that a researcher needs to do when confronted with this problem is to fit the saturated model, Model abc . This argument is incorrect. If any of the other models (Models a to bc) is the data-generating process and Model abc is fitted, one loses power for detecting the parameter of interest, $a \times b$, because one uses up degrees of freedom in estimating unneeded parameters. More generally, if the data-generating process is nested within (but not equal to) the fitted model, the power to detect the mediational effect is reduced. For instance, if Model a is the data-generating model and Models ab , ac , or abc are fitted, the mediational effect will be consistently estimated, but power will be less (confidence intervals for the parameter $a \times b$ will be wider) than the power resulting from using Model a . This is clearly undesirable if the population effect size is small, and/or if sample size is small. In contrast, if the data-generating model is not nested within the fitted model, the mediational effect will not be consistently estimated. For instance, if Model a is the data-generating model, and Models b , c , or bc are fitted, the estimate of the parameter of interest $a \times b$ will be biased, regardless of sample size. Fitting an overparameterized SEM model will lead to a loss of power for detecting effects of interest, whereas fitting an underparameterized SEM model (or more generally, a mis-specified model) will lead to inconsistent estimates in one or more model parameters.

Correct model specification is critical for sound substantive inferences—not just in a SEM but for any model, even the simplest possible one. For instance, consider investigating the extent to which individuals in a high-stakes job assessment situation distort their responses to a personality questionnaire. Suppose we have baseline (B) measures of their personality as well as their scores obtained when they apply (A) for a job. To what extent did individuals distort their responses? Provided the sample size is large enough, this question can be answered using a paired samples z test.

What assumptions are needed to make an inference in this case given a large enough sample? Apparently, none. Although the data (difference scores) may appear remarkably non-normal, unless the sample size is small, the central limit theorem ensures the validity of the statistical inference. However, there are “hidden” model assumptions (Cohen, Cohen, West, & Aiken, 2003). The easiest way to see this is by rephrasing the problem as a regression model, $y_A - y_B = \beta_0 + \varepsilon$. This regression model can be rewritten as $y_A = \beta_0 + y_B + \varepsilon$, and we now realize a model assumption made when using a paired samples z test, namely, that the slope between the two assessments is 1. The slope between two psychological measurements taken a couple of weeks apart is never 1. Therefore, the model assumption made when performing our paired sample z test is likely to be incorrect, thus biasing the results. To investigate whether the assumption used in our paired sample z test is tenable, we could fit $y_A = \beta_0 + \beta_1 y_B + \varepsilon$. However, this model in turn implies the standard assumption that the predictor (baseline score) is uncorrelated with the error (which contains unmodeled predictors of the assessment score). This assumption is likely to be incorrect as well, as we should expect individuals with a high level on a desirable trait to be less likely to distort their responses than those with a low level on the trait, leading to a heteroscedastic relationship between baseline and assessment scores. Thus, fitting a regression model to this problem is also likely to result in incorrect substantive conclusions. A model that accounts for the heteroscedasticity is likely to be needed. The morale of this digression is clear: even in extremely simple models such as a paired sample z test there are model assumptions that, when not met, lead to incorrect substantive conclusions.

2. Current Practices in Elementary Statistics

2.1. Classical Approach

Consider the paired sample z test for the score distortion example introduced above. A null hypothesis of no effect can be stated as $H_0 : \mu = \mu_A - \mu_B = 0$, where μ denotes the population mean of the difference scores. More generally, the null hypothesis can specify a non-null value, $H_0 : \mu = \mu_0$ versus say $H_1 : \mu \neq \mu_0$ and H_0 can be tested in large samples using a z statistic. Should a statistically significant difference be found, we examine the magnitude of the difference using an effect size and judge it qualitatively.

To me, an effect size is a population parameter that conveys the magnitude of the effect. We can convey this information for a paired sample z test in different ways. One way is by providing an estimate of the mean difference. This is reasonable if the units are meaningful. For the score distortion example, such an effect size (e.g., 10 points) is not useful to convey the magnitude of the effect. In this case, we could provide an estimate of the percentage increase (e.g., assessment scores are on average 20% higher than baseline scores). Often, these relative effect sizes are more informative than raw (or unstandardized) effect sizes. However, relative effect sizes are most useful when (a) they are coupled with information on the baseline, and (b) baseline units can be substantively interpreted. When the baseline units cannot be substantively interpreted, preference has evolved for unitless (standardized) effect sizes. Regardless of whether unstandardized, relative, or standardized effect sizes are used, it is necessary to provide a confidence interval for them instead of simply a point estimate to convey the precision with which we estimate the effect size, a population parameter.

2.2. Interval Estimation Approach

Some authors (Schmidt & Hunter, 1997; Steiger & Fouladi, 1997; and references therein) have strongly argued for bypassing the significance test altogether and focusing the analysis on the estimation of a confidence interval for an effect size of interest. One basic argument put forth by these authors is that, from a substantive point of view, it is far more interesting (and fruitful) to ask the question of how large an effect is (in a probabilistic sense, i.e., with a confidence interval) than asking whether an effect is exactly zero (which is very unlikely). For a paired sample z test, the unstandardized effect size is simply $\varepsilon_u = \mu - \mu_0$ and the standardized effect size is $\varepsilon_s = \frac{\mu - \mu_0}{\sigma}$, where σ denotes the population standard deviation of the difference scores and note that ε_s is Cohen's delta (e.g., Cohen, 1988). In large samples, a $(100 - \alpha)\%$ confidence interval for ε (that is, ε_s or ε_u), can be obtained using

$$\Pr(\hat{\varepsilon} - z_{\alpha/2}\text{ASE}(\hat{\varepsilon}) \leq \varepsilon \leq \hat{\varepsilon} + z_{\alpha/2}\text{ASE}(\hat{\varepsilon})) = 1 - \alpha, \quad (1)$$

where $\hat{\varepsilon}$ denotes the estimate of the effect size and ASE denotes its asymptotic standard error.

3. Current Practices in SEM

Before describing current practices in SEM, I shall provide some background. Covariance structure modeling most often begins by specifying a set of equations for mean-centered data, possibly involving latent variables. The model implies a covariance structure $\boldsymbol{\sigma}_0 = \boldsymbol{\sigma}(\boldsymbol{\theta}_0)$. Let p denote the number of observed variables being modeled; then $\boldsymbol{\sigma}$ is a $t = p(p + 1)/2$ vector of population covariances, and $\boldsymbol{\theta}_0$ is a $q \leq t$ vector of parameters to be estimated from the data.

The two best known procedures for estimating the model parameters involve minimizing the discrepancy functions

$$F = (\mathbf{s} - \boldsymbol{\sigma}_0)' \hat{\mathbf{W}} (\mathbf{s} - \boldsymbol{\sigma}_0), \quad (2)$$

and

$$F_{\text{ML}} = \ln |\boldsymbol{\Sigma}_0| - \ln |\mathbf{S}| + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0^{-1}) - p, \quad (3)$$

with respect to $\boldsymbol{\theta}$. In (3), \mathbf{S} and $\boldsymbol{\Sigma}_0$ denote the sample and population covariance matrices, respectively, whereas in (2), \mathbf{s} is a t -vector of sample covariances and $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, a fixed matrix. In covariance structure analysis, maximum likelihood (ML) parameter estimates under normality assumptions are frequently obtained by minimizing (3)—see Browne and Arminger (1995) for details. In contrast, (2) defines a class of functions. Different choices of the weight matrix $\hat{\mathbf{W}}$ lead to different estimators such as unweighted least squares, diagonally weighted least squares, or the asymptotically distribution free (ADF) weighted least squares proposed by Browne (1982, 1984).

One choice of weight matrix in (2) is

$$\hat{\mathbf{W}} = 2^{-1} \mathbf{D}' \left(\hat{\boldsymbol{\Sigma}}_0^{-1} \otimes \hat{\boldsymbol{\Sigma}}_0^{-1} \right) \mathbf{D}, \quad (4)$$

updated at each iteration, where \mathbf{D} is a duplication matrix (see Browne & Arminger, 1995). I shall denote (2) with (4) as F_{IRLS} (for Iteratively Reweighted Least Squares). Minimizing F_{IRLS} has the same solution as (3) but a different minimum (for details see Lee & Jennrich, 1979). Thus, F_{IRLS} can also be used to obtain normal theory (NT) ML estimates. NT ML is by far the most widely used estimator in covariance structure modeling both when normality assumptions are invoked but also under the ADF assumptions set forth by Browne (1982). In the latter case, asymptotic standard errors that are robust to non-normality are used (Satorra & Bentler, 1994).

3.1. Goodness-of-Fit Assessment: Classical Approach

Within a classical approach, the null hypothesis to be tested is $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}_0$ versus $H_1 : \boldsymbol{\sigma} \neq \boldsymbol{\sigma}_0$. There exist a number of test statistics with known asymptotic distribution under the null that can be used to test this hypothesis (Browne, 1984; Satorra & Bentler, 1994; Yuan & Bentler, 1997, 1998, 1999), some of which rely on normality assumptions, and others on ADF assumptions. These test statistics are simply a summary (often a weighted average) of the residual covariances, $\mathbf{s} - \hat{\boldsymbol{\sigma}}$, where $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$. For instance, for ML estimates obtained under normality assumptions, $\hat{T}_{\text{IRLS}} = n\hat{F}_{\text{IRLS}}$ is one suitable test statistic, where $n = N - 1$ and N denotes sample size. Under the null hypothesis, this statistic follows an asymptotic Chi-square distribution with $df = t - q$ degrees of freedom. However, in applications involving ML estimation under normality assumptions, the most widely used test statistic is the likelihood ratio test of the model against a saturated model, $LR = n\hat{F}_{\text{ML}} \xrightarrow{d} \chi_{df}^2$.

Yet, when (as is most often the case) the null hypothesis is rejected, the result of the goodness-of-fit test is usually disregarded on a number of grounds, such as (a) the model will be rejected for substantively trivial misfits when the sample size is very large, (b) the null hypothesis is a substantively uninteresting hypothesis, since by definition all models (that is, approximations) are wrong, or (c) we should focus on comparative model fit (i.e., model selection) rather than on model fit.

As a result, within a classical approach SEM practice usually relies on the use of goodness-of-fit indices. These are summary sample statistics used in a descriptive fashion such as

(a) the Comparative Fit Index (Bentler, 1990)

$$\widehat{\text{CFI}} = 1 - \frac{\max(\hat{T}_0 - df_0, 0)}{\max(\hat{T}_{\text{baseline}} - df_{\text{baseline}}, 0)}, \quad (5)$$

where \hat{T}_0 denotes the estimated value of the statistic used to test the fitted model (with df_0), and $\hat{T}_{\text{baseline}}$ denotes the value of the statistic when testing a baseline model (with df_{baseline}), usually the independence model,

(b) the Goodness-of-Fit Index (GFI: Jöreskog & Sörbom, 1981) which as presented in Tanaka and Huba (1989) is

$$\widehat{\text{GFI}} = 1 - \frac{(\mathbf{s} - \hat{\boldsymbol{\sigma}})' \hat{\mathbf{W}} (\mathbf{s} - \hat{\boldsymbol{\sigma}})}{\mathbf{s}' \hat{\mathbf{W}} \mathbf{s}}, \quad (6)$$

(c) the Standardized Root Mean Squared Residual (SRMR: Bentler, 1995; Jöreskog & Sörbom, 1988)

$$\widehat{\text{SRMR}} = \sqrt{\frac{(\mathbf{s} - \hat{\boldsymbol{\sigma}})' \hat{\mathbf{G}}^{-1} (\mathbf{s} - \hat{\boldsymbol{\sigma}})}{t}} = \sqrt{\frac{1}{t} \sum_{i < j} \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{\sqrt{s_{ii} s_{jj}}}}, \quad (7)$$

where $\hat{\mathbf{G}}$ is a $t \times t$ diagonal matrix with elements $s_{ii} s_{jj}$, that is, letting \mathbf{s}_{ii} be a p vector of sample variances, $\hat{\mathbf{G}} = \text{diag}(\text{vecs}(\mathbf{s}_{ii} \mathbf{s}'_{ii}))$, and $\text{vecs}()$ denotes an operator that stacks onto a column vector the elements at or below the diagonal of a symmetric matrix,

(d) the Correlation Root Mean Squared Residual (CRMRR: Bollen, 1989; Ogasawara, 2001)

$$\widehat{\text{CRMRR}} = \sqrt{\frac{(\hat{\mathbf{G}}^{-1/2} \mathbf{s} - \hat{\mathbf{G}}_0^{-1/2} \hat{\boldsymbol{\sigma}})' (\hat{\mathbf{G}}^{-1/2} \mathbf{s} - \hat{\mathbf{G}}_0^{-1/2} \hat{\boldsymbol{\sigma}})}{t - p}} = \sqrt{\frac{1}{t - p} \sum_{i < j} (r_{ij} - \hat{\rho}_{ij})^2}, \quad (8)$$

where $\hat{\mathbf{G}}_0 = \text{diag}(\text{vecs}(\hat{\boldsymbol{\sigma}}_{ii} \hat{\boldsymbol{\sigma}}'_{ii}))$.

Many more goodness-of-fit indices have been proposed in the literature. For an overview, see Marsh, Hau, and Grayson (2005) and Bollen and Long (1993). In practice, goodness-of-fit indices such as the GFI, CFI or SRMR are used as if they were test statistics in a classical null hypothesis testing framework using cutoff values that have been proposed in the literature (e.g., Hu & Bentler, 1999). Thus, if a goodness-of-fit index meets the recommended cutoff values, the model is retained and used as if it were correctly specified. In other words, fit indices coupled with fixed cutoff values are used as if they were test statistics of a vaguely specified hypothesis of “close” fit: the model fit is close if it meets the fixed cutoff criterion for that index, and not close if it exceeds the cutoff (Barrett, 2007; Marsh, Hau, & Wen, 2004; Saris, Satorra, & van der Veld, 2009). The problem is compounded by the existence of different fit indices, each one with one or more proposed cutoff values. As a result, unscrupulous researchers may choose the most favorable index and the most favorable cutoff value (McDonald & Ho, 2002).

There are several problems with current practices. I will just highlight three. First, they are pre-statistical. The parameter being estimated by each index is often not described, nor is care taken to ensure that the index is an asymptotically unbiased (or at least consistent) estimate of the parameter of interest. Furthermore, the sampling variability of the statistic is blatantly ignored. In

many ways, retaining a model if, say, the sample CFI is larger than some cutoff value (say .99) is analogous to concluding in a z test that there is a mean difference worth speaking of if the observed mean percentage increase is larger than some fixed cutoff value (say 10%). Second, partly as a result of the approach being pre-statistical, no agreement is possible regarding which index to use, or what cutoff value. Third, and most importantly, inferences on the model parameters are made as if the model were correct, when in fact, it is mis-specified.

3.2. Goodness-of-Fit Assessment: Interval Estimation Approach

Within this approach, a confidence interval for an effect size of model misfit is estimated. The effect size of model misfit reported almost invariably in current SEM practice is the Root Mean Squared Error of Approximation (RMSEA). The RMSEA is often referred to as a goodness-of-fit index. I prefer to reserve the term “goodness-of-fit index” for sample statistics whose distribution is unknown, or if known, is not used in practice. The asymptotic distribution of the sample RMSEA is known under general conditions (Browne & Cudeck, 1993, Steiger, 1989), and is widely used in practice for statistical inference.

The RMSEA was introduced by Steiger (Steiger, 1989, 1990; Steiger & Lind, 1980) with contributions by Browne and Cudeck (1993). Consider a discrepancy function of the sample and model-implied population covariances such as (2) and (3). As described by Steiger (1989) and Browne and Cudeck (1993), the population RMSEA is defined as

$$\text{RMSEA} = \sqrt{\frac{F(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)}{df}} \quad (9)$$

where $F(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$ denotes the discrepancy between the true and unknown population covariance matrix, $\boldsymbol{\Sigma}$, and the covariance matrix specified under the null hypothesis, $\boldsymbol{\Sigma}_0$.

The sample RMSEA is defined as (Browne & Cudeck, 1993; Steiger, 1989)

$$\widehat{\text{RMSEA}} = \sqrt{\max\left(\frac{n\hat{F} - df}{n \times df}, 0\right)}. \quad (10)$$

Provided $n\hat{F}$ follows, asymptotically, a non-central Chi-square distribution under a sequence of local alternatives, confidence intervals for the population RMSEA can be obtained. Alternatively, the RMSEA may be used to test an hypothesis of approximate (or “close”) fit of the model, $H_0^* : \text{RMSEA} \leq c$ versus $H_1^* : \text{RMSEA} > c$, where $c > 0$ is some arbitrary cutoff value. When $c = 0$, H_0^* is equivalent to the null hypothesis of exact fit $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta})$. Browne and Cudeck (1993, p. 144) famously stated that, based on their practical experience, they felt that models with $\text{RMSEA} \leq 0.05$ provided a close fit to the unknown population covariance matrix (in relation to the degrees of freedom), that models with an RMSEA around 0.08 provided a reasonable approximation, and that models with an RMSEA greater than 0.1 provided a poor fit. As a result, the hypothesis of close fit being tested in applications is $H_0^* : \text{RMSEA} \leq 0.05$, which corresponds to Steiger’s (1989) cutoff for a very good fit.

4. Reconciling Current Practices in Introductory Statistics and in SEM

The use of the RMSEA effect size of the misfit (9) and accompanying statistical theory to obtain a confidence interval for it (within an interval estimation approach) or a p value for a test

of close fit $H_0^* : \text{RMSEA} \leq 0.05$ (within a classical approach) provides a marked improvement over practices involving the comparison of sample goodness-of-fit indices to fixed cutoff values. However, upon realizing that the population RMSEA is an effect size of the misfit of a SEM model, I saw that we need to examine more closely than we have done so far how to assess the size of the misfit of a SEM model. For it is the notion of effect size of the misfit that will enable us to reconcile current practices in introductory statistics (e.g., when reporting z tests results) and in SEM.

Indeed, I believe that current practices in SEM will substantially improve by mimicking current practices elsewhere in statistics. Thus, within a classical approach, if the null hypothesis is tested and rejected, one should report the effect size of the misfit (with a confidence interval), and judge it qualitatively, just as in a z test. On the other hand, within an interval estimation approach, one should first judiciously choose the effect size of the misfit and report a confidence interval for it.

However, there are three key differences between the null hypothesis considered in covariance structure modeling [$H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0 = \boldsymbol{\sigma}(\boldsymbol{\theta}_0)$] and a paired sample z test ($H_0 : \mu = \mu_0$). The first is that in the z test, the parameter involved in the null hypothesis is the parameter of interest, whereas in SEM, the parameters involved in the null hypothesis of model fit, $\boldsymbol{\sigma}$, are most often simply a means to an end: the interest lies in the fundamental parameters $\boldsymbol{\theta}$. A second difference is that H_0 in a z test is univariate and only one parameter is involved, while in SEM we deal with a multivariate null hypothesis. A third difference is that in a z test H_0 is a simple null hypothesis (μ_0 is fixed), whereas in SEM we deal with a composite null hypothesis ($\boldsymbol{\theta}_0$ is to be estimated from the data).

The first difference implies that one needs to relate the model misfit results to the consistency with which parameters of interest are estimated. There is not much research on this topic, though see for instance Yuan, Marshall, and Bentler (2003) and Reise, Scheines, Widaman, and Haviland (2012). The second difference implies that in a z test there is only one effect size to consider, although we are free to choose the most suitable one for the analysis (e.g., unstandardized, standardized, or relative), whereas in SEM one may consider t effect sizes in which the overall effect size of the misfit will simply be a summary measure, however imperfect, of these t effect sizes. Also, because only one parameter is considered in a z test, assessing power involves a single dimension; in contrast, in SEM there are multiple possible directions of misfit to consider when assessing power. The third difference simply implies that the statistical theory for obtaining confidence intervals will necessarily be more involved in SEM than in a z test.

5. Effect Sizes of SEM Misfit

5.1. A Multivariate Problem

The size of the misfit in a covariance structure model cannot be captured by a single effect size parameter because of the multivariate nature of the data. Rather, for each variance and covariance we can define an unstandardized effect size $\varepsilon_{ij}^{(u)}$ as well as a standardized effect size parameter $\varepsilon_{ij}^{(s)}$ analogous to the ones used in a paired sample z test where $\varepsilon_{ij}^{(u)} = \sigma_{ij} - \sigma_{ij}^0$ and

$$\varepsilon_{ij}^{(s)} = \frac{\sigma_{ij} - \sigma_{ij}^0}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \quad (11)$$

Here, σ_{ij} denotes the true and unknown population covariance between variables i and j (or variance if $i = j$) and σ_{ij}^0 denotes the population covariance (or variance) under the fitted model

obtained as the solution to (13)—see below. Almost invariably, we will not be able to interpret the magnitude of the unstandardized effect size $\varepsilon_{ij}^{(u)}$. Therefore, as an effect size it is most often useless. To obtain an effect size whose magnitude can be interpreted, in (11) the unstandardized effect size is standardized by dividing it by the product of the population standard deviations.

Another possible effect size is

$$\varepsilon_{ij}^{(r)} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} - \frac{\sigma_{ij}^0}{\sqrt{\sigma_{ii}^0\sigma_{jj}^0}} = \rho_{ij} - \rho_{ij}^0. \quad (12)$$

Here, the population covariance is standardized by dividing by the population standard deviations, and the covariance under the null hypothesis is standardized by dividing by the null standard deviations. $\varepsilon_{ij}^{(r)}$ is simply the population residual correlation between the population and the fitted model.

Letting $\hat{\sigma}_{ij} \equiv \sigma_{ij}(\hat{\boldsymbol{\theta}})$, $\varepsilon_{ij}^{(u)}$ and $\varepsilon_{ij}^{(s)}$ can be consistently estimated using the sample residual covariance and standardized residual covariance respectively, $e_{ij}^{(u)} = s_{ij} - \hat{\sigma}_{ij}$, $e_{ij}^{(s)} = \frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii}s_{jj}}}$ and note that the standardized residual covariance need not be in the interval $(-1, 1)$, but one expects it to be within this metric in most cases. In contrast, the sample residual correlation, $e_{ij}^{(r)} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}} = r_{ij} - \hat{\rho}_{ij}$, is most often within the interval $(-1, 1)$.

Now, to obtain confidence intervals for these population effect sizes, I assume that the population covariances $\boldsymbol{\sigma}$ are related to the fitted model, $\boldsymbol{\sigma}_0$, by the standard assumption of a sequence of local alternatives (aka parameter drift assumption), $\boldsymbol{\sigma} = \boldsymbol{\sigma}_0 + \boldsymbol{\delta}/\sqrt{N}$, where $\boldsymbol{\sigma}_0 = \boldsymbol{\sigma}(\boldsymbol{\theta}_0)$. I also assume that $\boldsymbol{\sigma}_0$ corresponds to the value of $\boldsymbol{\theta}_0$ that minimizes $(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)' \mathbf{W} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)$ so that $\boldsymbol{\theta}_0$ satisfies

$$\boldsymbol{\Delta}'_0 \mathbf{W} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0) = \mathbf{0}, \quad (13)$$

where $\boldsymbol{\Delta}_0 = \left. \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. Finally, with $\boldsymbol{\delta} = \sqrt{N}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)$ I assume that

$$\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma}_0) \xrightarrow{d} N(\boldsymbol{\delta}, \boldsymbol{\Gamma}). \quad (14)$$

Consider now the asymptotic distribution of the unstandardized, $\boldsymbol{\varepsilon}_u = \mathbf{s} - \hat{\boldsymbol{\sigma}}$, standardized, $\boldsymbol{\varepsilon}_s = \hat{\mathbf{G}}^{-1/2}(\mathbf{s} - \hat{\boldsymbol{\sigma}})$, and correlation residuals, $\boldsymbol{\varepsilon}_r = \hat{\mathbf{G}}^{-1/2}\mathbf{s} - \hat{\mathbf{G}}_0^{-1/2}\hat{\boldsymbol{\sigma}} = \mathbf{r} - \hat{\boldsymbol{\rho}}$ respectively, with population parameters

$$\boldsymbol{\varepsilon}_u = \boldsymbol{\sigma} - \boldsymbol{\sigma}_0 = \boldsymbol{\delta}/\sqrt{N}, \quad \boldsymbol{\varepsilon}_s = \mathbf{G}^{-1/2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0), \quad \text{and} \quad \boldsymbol{\varepsilon}_r = \mathbf{G}^{-1/2}\boldsymbol{\sigma} - \mathbf{G}_0^{-1/2}\boldsymbol{\sigma}_0, \quad (15)$$

where $\mathbf{G} = \text{diag}(\text{vecs}(\sigma_{ii}\sigma'_{ii}))$. With $\mathbf{H} = \boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{W}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{W}$, and $\mathbf{F} = \left. \frac{\partial(\mathbf{r}-\hat{\boldsymbol{\rho}})}{\partial \hat{\boldsymbol{\sigma}}'} \right|_{\hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}}$, we obtain $\boldsymbol{\varepsilon}_u \xrightarrow{d} N(\boldsymbol{\varepsilon}_u, \boldsymbol{\Xi}_u)$, $\boldsymbol{\Xi}_u = N^{-1}(\mathbf{I} - \mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \mathbf{H})'$,

$$\boldsymbol{\varepsilon}_s \xrightarrow{d} N(\boldsymbol{\varepsilon}_s, \boldsymbol{\Xi}_s), \quad \boldsymbol{\Xi}_s = \mathbf{G}^{-1/2}\boldsymbol{\Xi}_u\mathbf{G}^{-1/2}, \quad (16)$$

$$\boldsymbol{\varepsilon}_r \xrightarrow{d} N(\boldsymbol{\varepsilon}_r, \boldsymbol{\Xi}_r), \quad \boldsymbol{\Xi}_r = \mathbf{F}\boldsymbol{\Xi}_u\mathbf{F}', \quad (17)$$

where $\boldsymbol{\Gamma}$ denotes the asymptotic covariance matrix of $\sqrt{N}\mathbf{s}$, which can be computed under normality assumptions, $\boldsymbol{\Gamma}_{NT}$, or under ADF assumptions, $\boldsymbol{\Gamma}_{ADF}$ (see for instance Satorra & Bentler, 1994).

Therefore, the sample effect sizes are asymptotically unbiased estimators of their respective population effect sizes. Furthermore, using (1), for any estimator within the family of estimators (2)—which includes the MLE—one can obtain confidence intervals (under normality, or “robust” to non-normality) for these effect sizes of the misfit. Thus, for instance, the asymptotic standard error of a standardized residual covariance is $\text{ASE}(e_{ij}^{(s)}) = \sqrt{\Xi_s^{(ij)}} = \sqrt{\Xi_u^{(ij)}} / \sqrt{\sigma_{ii}\sigma_{jj}}$ (Ogasawara, 2001), where $\Xi_s^{(ij)}$ is the corresponding diagonal element of Ξ_s .

5.2. Overall Effect Size of the Misfit

In practice, applied researchers will be interested in a single parameter that summarizes the effect of the misfit. A “natural” overall unstandardized effect size is the population Root Mean Squared Residual (RMSR), $\iota_u := \text{RMR} = \sqrt{\frac{\mathbf{e}'_u \mathbf{e}_u}{t}} = \sqrt{\frac{(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)'(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)}{t}}$. It is clear that this overall effect size cannot be substantively interpreted. In contrast, the population Standardized Root Mean Residual

$$\iota_s := \text{SRMR} = \sqrt{\frac{\mathbf{e}'_s \mathbf{e}_s}{t}} = \sqrt{\frac{(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)' \mathbf{G}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)}{t}}, \quad (18)$$

can be approximately interpreted as an average of the standardized effect sizes. Finally, we can define a population Correlation Root Mean Squared Residual

$$\iota_r := \text{CRMR} = \sqrt{\frac{\mathbf{e}'_r \mathbf{e}_r}{t-p}} = \sqrt{\frac{(\mathbf{G}^{-1/2} \boldsymbol{\sigma} - \mathbf{G}_0^{-1/2} \boldsymbol{\sigma}_0)' (\mathbf{G}^{-1/2} \boldsymbol{\sigma} - \mathbf{G}_0^{-1/2} \boldsymbol{\sigma}_0)}{t-p}}, \quad (19)$$

which can be approximately interpreted as the average of the population residual correlations.

To obtain suitable estimators for these population parameters, consider the sum of squared residuals, $T = T_s$ or T_r , where $T_s = \mathbf{e}'_s \mathbf{e}_s$ and $T_r = \mathbf{e}'_r \mathbf{e}_r$ denote the sum of squared standardized and correlation residuals, respectively. The asymptotic mean and variance of T under parameter drift assumptions are (see the “Appendix”)

$$\text{E}(T) = \mu_T = \text{tr}(\Xi) + \boldsymbol{\epsilon}' \boldsymbol{\epsilon}, \quad (20)$$

$$\text{var}(T) = \sigma_T^2 = 2\text{tr}(\Xi^2) + 4\boldsymbol{\epsilon}' \Xi \boldsymbol{\epsilon}, \quad (21)$$

where $\Xi = \Xi_s$ or Ξ_r . Therefore, using the method of moments, an asymptotically unbiased estimate of the population sum of squared residuals $\boldsymbol{\epsilon}' \boldsymbol{\epsilon}$ is $\hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} = T - \text{tr}(\Xi)$ leading to the estimator for the overall effect sizes $\iota = \iota_s$ or ι_r $\tilde{\iota} = \sqrt{\frac{T - \text{tr}(\Xi)}{t}}$, where for ι_r , we use $t - p$ instead of t . In turn, the asymptotic mean and variance of this estimator are

$$\text{E}(\tilde{\iota}) = \left(\frac{\mu_T - \text{tr}(\Xi)}{t} \right)^{-\frac{3}{2}} \frac{8(\mu_T - \text{tr}(\Xi))^2 - \sigma_T^2}{8t^2} = \iota \left(1 - \frac{\sigma_T^2}{8(\boldsymbol{\epsilon}' \boldsymbol{\epsilon})^2} \right), \quad (22)$$

$$\text{var}(\tilde{\iota}) = \frac{\sigma_T^2}{4t(\mu_T - \text{tr}(\Xi))}. \quad (23)$$

Equation (22) reveals that $\tilde{\iota}$ is an asymptotically biased estimator of the overall population effect sizes and that an asymptotically unbiased estimator is $\hat{\iota} = k^{-1} \tilde{\iota}$, where $k = 1 - \frac{\sigma_T^2}{8(\boldsymbol{\epsilon}' \boldsymbol{\epsilon})^2}$. As a result, asymptotically unbiased estimates of the population SRMR and CRMR can be obtained using

$$\hat{l}_s = \hat{k}_s^{-1} \sqrt{\frac{\max(T_s - \text{tr}(\hat{\Xi}_s), 0)}{t}}, \quad \hat{l}_r = \hat{k}_r^{-1} \sqrt{\frac{\max(T_r - \text{tr}(\hat{\Xi}_r), 0)}{t - p}}, \quad (24)$$

$$\hat{k}_s = 1 - \frac{\text{tr}(\hat{\Xi}_s^2) + 2\mathbf{e}'_s \hat{\Xi}_s \mathbf{e}_s}{4T_s^2}, \quad \hat{k}_r = 1 - \frac{\text{tr}(\hat{\Xi}_r^2) + 2\mathbf{e}'_r \hat{\Xi}_r \mathbf{e}_r}{4T_r^2}. \quad (25)$$

Equation (22) also reveals that the sample RMR (Jöreskog & Sörbom, 1988), $\widehat{RMSR} = \sqrt{(\mathbf{s} - \hat{\boldsymbol{\sigma}})'(\mathbf{s} - \hat{\boldsymbol{\sigma}})/t}$, and the samples SRMR and CRMR given in Eqs. (7) and (8) overestimate their population parameters.

6. Confidence Intervals for the SRMR and CRMR and Associated Tests of Close Fit

If the sampling distribution of $\hat{l} = \hat{l}_s, \hat{l}_r$ is well approximated in large samples by a normal distribution, then a $(100 - \alpha)\%$ confidence interval for $\iota = \iota_s, \iota_r$ can be obtained using an expression analogous to (1) where the asymptotic standard errors are (see ‘‘Appendix’’)

$$\text{ASE}(\hat{l}_s) = \sqrt{k_s^{-2} \frac{\text{tr}(\Xi_s^2) + 2\mathbf{e}'_s \Xi_s \mathbf{e}_s}{2t \times T_s}}, \quad + \text{ASE}(\hat{l}_r) = \sqrt{k_r^{-2} \frac{\text{tr}(\Xi_r^2) + 2\mathbf{e}'_r \Xi_r \mathbf{e}_r}{2(t - p)T_r}}. \quad (26)$$

If a test of close fit based on the SRMR or CRMR is desired, of the type

$$H_0^* : \iota \leq \iota_0 \text{ versus } H_1^* : \iota > \iota_0, \quad (27)$$

where $\iota_0 > 0$ is an arbitrary value of the SRMR or CRMR, p values are obtained using $p = 1 - \Phi(z)$, where $\Phi(\cdot)$ denotes a standard normal distribution function and $z = \frac{\hat{l} - \iota_0}{\text{ASE}(\hat{l})}$.

7. Tests of Exact Fit Based on the SRMR and CRMR

For tests of exact fit of the type $H_0 : \iota = 0$ versus $H_1 : \iota > 0$, where ι denotes either the population SRMR or CRMR, I consider the statistics $T^* = \widehat{SRMSR}$ or \widehat{CRMSR} given in Eqs. (7) and (8). Under this null hypothesis, the asymptotic mean and variance of the sample SRMR and CRMR are (see ‘‘Appendix’’)

$$\text{E}(T^*) = \mu_{T^*} = \sqrt{\frac{\text{tr}(\Xi) 4\text{tr}(\Xi)^2 - \text{tr}(\Xi^2)}{t 4\text{tr}(\Xi)^2}}, \quad (28)$$

$$\text{var}(T^*) = \sigma_{T^*}^2 = \frac{\text{tr}(\Xi^2)}{2t \text{tr}(\Xi)}, \quad (29)$$

where for the CRMR $t - p$ is used instead of t (see also Ogasawara, 2001). Again, assuming that the sampling distribution of $T^* = \widehat{SRMR}$ or \widehat{CRMR} is well approximated in large samples by a normal distribution, p values for the null hypothesis of exact fit are obtained using

$$p = 1 - \Phi(z), \quad z = \frac{T^* - \mu_{T^*}}{\sqrt{\sigma_{T^*}^2}}. \quad (30)$$

8. Small Sample Behavior of SRMR and CRMR Confidence Intervals and Goodness-of-Fit Tests

To investigate the performance of the methods described above, I performed a small simulation study involving a two-factor model. More specifically, multivariate normal data for $p = 8$ variables were generated with mean zero and a covariance structure $\Sigma = \Lambda\Psi\Lambda' + \Theta$ with $\Lambda' = \begin{pmatrix} 0.8 & 0.7 & 0.7 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.7 & 0.7 & 0.8 \end{pmatrix}$, $\Psi = \begin{pmatrix} 1 & \\ \rho & 1 \end{pmatrix}$, and $\Theta = \mathbf{I} - \text{diag}(\Lambda\Psi\Lambda')$. Sample sizes were 100, 200, 500, 1000, and 3000; interfactor correlations were $\rho = 0.7, 0.8$ and 0.9 . A one-factor model was estimated in all cases by maximum likelihood under normality assumptions. 1000 replications per condition were used. In Table 1 I provide the population values corresponding to each condition, the average sample SRMR (7) and CRMR (8) and the unbiased estimates of the population parameters (24). We see in this table that the simulation setup yields an array of population values: the SRMR values investigated are 0.074, 0.049, and 0.025, and the CRMR values investigated are 0.084, 0.056 and 0.028.

We also see in Table 1 that the samples SRMR and CRMR overestimate their population parameters substantially as sample size and degree of misfit decreases. For instance, when the correlation between the two factors is 0.9 and sample size is 100, the population CRMR is 0.028, but the average sample CRMR is a whopping 0.048. In contrast, the unbiased estimates are right on target: for the above example, the average of the unbiased estimates is 0.027. I also provide in Table 1 the standard deviation of the unbiased estimates, the average of the standard errors (SEs) of the unbiased estimates, and coverage rates for 90% confidence intervals computed using a reference normal distribution. We see that the SEs slightly overestimate the true variability of the estimates when the sample size is 200 or less, particularly as the degree of misfit decreases. This is because of the use of a truncated estimate. As a result, coverage rates are unacceptably small when the sample size is 200 or less and the fitted model is rather close to the true model (population parameters of SRMR = 0.025 and CRMR = 0.028). In all other conditions, coverage is at or below nominal rates, but it is at least 0.86 even at the smallest sample size considered ($N = 100$). I also examined coverage rates for individual standardized residuals (11) and individual residual correlations (12). For the worst case ($\rho = 0.9$, or SRMR = 0.025 and CRMR = 0.028) and smallest sample size considered ($N = 100$), coverage rates for 90% confidence intervals are at worst 0.86.

Table 2 shows the empirical rejection rates (Type I errors) for selected values of nominal rejection rates of the test of close fit (27) using a reference normal distribution. In all cases, the value being tested is whether the SRMR or CRMR parameter equals its true value. We see in this table that the proposed test provides accurate Type I errors for all levels of SRMR (0.025–0.074) and CRMR (0.028 to 0.084) even at the smallest sample size considered ($N = 100$) and for all significance levels except when $\alpha > 0.60$, $N < 500$ and at the smallest model discrepancy considered ($\rho = 0.9$, or SRMR = 0.025 and CRMR = 0.028).

To investigate the performance of the methods described above when the model holds, I used the same setup as before, except that now data were generated under a one-factor model. Table 3 shows the empirical rejection rates (Type I errors) for selected values of nominal rejection rates of the test of exact fit $H_0 : \iota = 0$ versus $H_1 : \iota > 0$ using a reference normal distribution. In this table, we see that the proposed test provides accurate Type I errors even at the smallest sample size considered ($N = 100$) and for all significance levels.

Additional Monte Carlo results on the accuracy of confidence intervals and tests for single standardized residuals and residual correlations are given in Maydeu-Olivares and Shi (2017)—see also Ogasawara (2001).

TABLE 1.

Standardized Root Mean Squared Residual (SRMR) and Correlation Root Mean Squared Residual (CRM): population values, average and standard deviation of sample estimates, average and standard error of unbiased estimates, average standard errors for unbiased estimates, and coverage rates for 90% confidence intervals.

ρ	N	Stat.	Pop.	Sample mean	Sample SD	Estimate mean	Estimate SD	Estimate mean SE	Coverage	
0.7	100	SRMR	0.074	0.084	0.017	0.074	0.019	0.018	0.88	
		CRM	0.084	0.095	0.019	0.084	0.021	0.019	0.86	
	200	SRMR	0.074	0.079	0.012	0.075	0.013	0.012	0.89	
		CRM	0.084	0.090	0.014	0.085	0.015	0.014	0.87	
	500	SRMR	0.074	0.076	0.008	0.074	0.008	0.008	0.89	
		CRM	0.084	0.086	0.009	0.084	0.009	0.009	0.89	
	1000	SRMR	0.074	0.075	0.005	0.074	0.006	0.005	0.89	
		CRM	0.084	0.085	0.006	0.084	0.006	0.006	0.89	
	3000	SRMR	0.074	0.074	0.003	0.074	0.003	0.003	0.90	
		CRM	0.084	0.084	0.004	0.084	0.004	0.004	0.90	
	0.8	100	SRMR	0.049	0.061	0.013	0.048	0.016	0.014	0.86
			CRM	0.056	0.070	0.014	0.055	0.018	0.016	0.86
200		SRMR	0.049	0.056	0.010	0.049	0.011	0.010	0.87	
		CRM	0.056	0.063	0.011	0.056	0.012	0.011	0.87	
500		SRMR	0.049	0.052	0.006	0.049	0.006	0.006	0.88	
		CRM	0.056	0.059	0.007	0.056	0.007	0.007	0.88	
1000		SRMR	0.049	0.051	0.004	0.049	0.005	0.004	0.89	
		CRM	0.056	0.057	0.005	0.056	0.005	0.005	0.89	
3000		SRMR	0.049	0.050	0.003	0.049	0.003	0.003	0.89	
		CRM	0.056	0.056	0.003	0.056	0.003	0.003	0.89	
0.9		100	SRMR	0.025	0.043	0.009	0.022	0.014	0.012	0.78
			CRM	0.028	0.048	0.010	0.026	0.016	0.013	0.78
	200	SRMR	0.025	0.035	0.007	0.024	0.010	0.008	0.83	
		CRM	0.028	0.039	0.007	0.027	0.011	0.009	0.83	
	500	SRMR	0.025	0.029	0.004	0.024	0.005	0.005	0.87	
		CRM	0.028	0.033	0.005	0.028	0.006	0.005	0.87	
	1000	SRMR	0.025	0.027	0.003	0.025	0.003	0.003	0.88	
		CRM	0.028	0.030	0.004	0.028	0.004	0.004	0.88	
	3000	SRMR	0.025	0.025	0.002	0.025	0.002	0.002	0.89	
		CRM	0.028	0.029	0.002	0.028	0.002	0.002	0.89	

8.1. The RMSEA as an Overall Effect Size

The population RMSEA (9) as defined by Steiger (1989) and Browne and Cudeck (1993) is not unique; rather, it changes depending on the method used to estimate the model parameters. For instance, when ML estimates under normality are obtained by minimizing F_{IRLS} —i.e., (2) using the weight matrix (4)—the population RMSEA being estimated is

$$\text{RMSEA}_{\text{NT},1} = \sqrt{\frac{(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0) \boldsymbol{\Gamma}_{\text{NT}}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)}{df}}, \quad (31)$$

where $\boldsymbol{\Gamma}_{\text{NT}}$ denotes the asymptotic covariance matrix of $\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma})$ under normality assumptions.

In contrast, when these ML estimates are obtained by minimizing the F_{ML} function (3), the population RMSEA being estimated is

TABLE 2.

Test of close model fit based on the Standardized Root Mean Squared Residual (SRMR) and Correlation Root Mean Squared Residual (CRM): empirical rejection rates for selected levels of significance level.

ρ	N	Stat.	Pop.	α											
				0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
0.7	100	SRMR	0.074	<0.01	0.03	0.07	0.18	0.29	0.39	0.49	0.59	0.68	0.77	0.86	
		CRM	0.084	0.01	0.05	0.09	0.20	0.30	0.40	0.50	0.60	0.68	0.77	0.86	
	200	SRMR	0.074	<0.01	0.04	0.10	0.21	0.31	0.40	0.49	0.58	0.68	0.78	0.88	
		CRM	0.084	0.01	0.06	0.11	0.23	0.32	0.41	0.49	0.58	0.68	0.78	0.88	
	500	SRMR	0.074	0.01	0.04	0.08	0.19	0.30	0.41	0.50	0.60	0.69	0.79	0.89	
		CRM	0.084	0.01	0.05	0.09	0.20	0.30	0.41	0.51	0.60	0.69	0.79	0.89	
	1000	SRMR	0.074	0.01	0.04	0.10	0.19	0.29	0.41	0.50	0.61	0.70	0.79	0.90	
		CRM	0.084	0.01	0.05	0.11	0.20	0.30	0.41	0.50	0.61	0.70	0.79	0.89	
	3000	SRMR	0.074	0.01	0.04	0.10	0.21	0.32	0.41	0.50	0.61	0.71	0.81	0.91	
		CRM	0.084	0.01	0.05	0.11	0.21	0.32	0.41	0.50	0.61	0.70	0.81	0.91	
	0.8	100	SRMR	0.049	<0.01	0.05	0.10	0.22	0.30	0.39	0.47	0.57	0.67	0.76	0.86
			CRM	0.056	<0.01	0.05	0.11	0.22	0.31	0.39	0.48	0.57	0.67	0.77	0.87
200		SRMR	0.049	<0.01	0.05	0.10	0.22	0.30	0.39	0.47	0.57	0.67	0.76	0.86	
		CRM	0.056	<0.01	0.05	0.11	0.22	0.31	0.39	0.48	0.57	0.67	0.77	0.87	
500		SRMR	0.049	0.01	0.04	0.09	0.20	0.29	0.40	0.50	0.59	0.68	0.77	0.88	
		CRM	0.056	0.01	0.04	0.09	0.20	0.29	0.40	0.50	0.60	0.68	0.77	0.88	
1000		SRMR	0.049	0.01	0.05	0.10	0.19	0.29	0.40	0.50	0.60	0.71	0.78	0.88	
		CRM	0.056	0.01	0.05	0.10	0.19	0.29	0.40	0.50	0.60	0.71	0.78	0.88	
3000		SRMR	0.049	0.01	0.05	0.10	0.21	0.31	0.41	0.50	0.60	0.70	0.81	0.90	
		CRM	0.056	0.01	0.05	0.10	0.21	0.31	0.41	0.50	0.60	0.70	0.81	0.90	
0.9		100	SRMR	0.025	0.01	0.04	0.08	0.19	0.30	0.39	0.48	0.57	0.65	0.71	0.77
			CRM	0.028	0.01	0.04	0.09	0.20	0.30	0.40	0.49	0.58	0.65	0.71	0.78
	200	SRMR	0.025	0.01	0.06	0.12	0.22	0.29	0.37	0.46	0.57	0.65	0.75	0.84	
		CRM	0.028	0.01	0.06	0.12	0.22	0.29	0.37	0.47	0.57	0.65	0.75	0.84	
	500	SRMR	0.025	0.01	0.05	0.09	0.20	0.29	0.40	0.49	0.59	0.68	0.76	0.85	
		CRM	0.028	0.01	0.05	0.09	0.20	0.29	0.40	0.50	0.59	0.68	0.76	0.85	
	1000	SRMR	0.025	0.01	0.05	0.09	0.19	0.30	0.39	0.50	0.60	0.69	0.78	0.87	
		CRM	0.028	0.01	0.05	0.09	0.19	0.30	0.39	0.50	0.60	0.69	0.78	0.87	
	3000	SRMR	0.025	0.01	0.04	0.09	0.21	0.31	0.40	0.49	0.60	0.70	0.80	0.89	
		CRM	0.028	0.01	0.04	0.09	0.21	0.31	0.40	0.49	0.60	0.70	0.80	0.89	

The null hypothesis tested is that the SRMR and CRM equal their population values.

$$RMSEA_{NT,2} = \sqrt{\frac{\ln |\Sigma_0| - \ln |\Sigma| + \text{tr}(\Sigma \Sigma_0^{-1}) - p}{df}} \quad (32)$$

Furthermore, if the model parameters are estimated by minimizing (2) using the weight matrix $\hat{W} = \hat{\Gamma}_{ADF}^{-1}$, where $\hat{\Gamma}_{ADF}$ denotes the consistent estimator of the asymptotic covariance matrix of $\sqrt{N}(s - \sigma)$ under ADF assumptions, the RMSEA being estimated is

$$RMSEA_{ADF} = \sqrt{\frac{(\sigma - \sigma_0)' \Gamma_{ADF}^{-1} (\sigma - \sigma_0)}{df}} \quad (33)$$

TABLE 3.

Test of exact model fit based on the Standardized Root Mean Squared Residual (SRMR) and Correlation Root Mean Squared Residual (CRMR): empirical rejection rates for selected levels of significance level.

N	Stat.	α										
		0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
100	SRMR	0.01	0.05	0.11	0.20	0.30	0.42	0.53	0.62	0.74	0.84	0.92
	CRMR	0.01	0.05	0.11	0.20	0.30	0.43	0.53	0.63	0.74	0.84	0.92
200	SRMR	0.01	0.05	0.10	0.20	0.30	0.42	0.53	0.64	0.74	0.83	0.91
	CRMR	0.01	0.06	0.10	0.20	0.30	0.42	0.53	0.64	0.74	0.83	0.91
500	SRMR	0.02	0.06	0.11	0.22	0.30	0.40	0.49	0.60	0.70	0.81	0.92
	CRMR	0.02	0.06	0.11	0.22	0.30	0.40	0.49	0.60	0.70	0.81	0.92
1000	SRMR	0.01	0.05	0.10	0.20	0.29	0.39	0.49	0.59	0.71	0.82	0.92
	CRMR	0.01	0.05	0.10	0.20	0.29	0.39	0.49	0.59	0.71	0.82	0.92
3000	SRMR	0.01	0.05	0.10	0.20	0.30	0.40	0.51	0.58	0.70	0.79	0.91
	CRMR	0.01	0.05	0.10	0.20	0.30	0.40	0.51	0.58	0.70	0.79	0.91

The null hypothesis tested is that the population SRMR and CRMR equal zero.

I find it undesirable that the population parameter being estimated to communicate the size of the model misfit changes depending on the estimator used; I think that the same parameter should be estimated regardless of the estimator and distributional assumptions. I will return to this point later.

For now, let us assume for ease of exposition that the RMSEA parameter being estimated is (31). The $RMSEA_{NT,1}$ differs from the RMR in that (a) a weighted sum of the unstandardized effect sizes is used instead of an unweighted sum, and (b) the sum is divided by degrees of freedom instead of by the number of effect sizes. I will examine each of these features of $RMSEA_{NT,1}$ in turn.

In my opinion, the only weighting of population parameters that can be justified is one that facilitates the substantive interpretation of the parameter being estimated. The weight matrix used in the population SRMR (18) is one such choice. The weight matrix (4) does not facilitate the substantive interpretation of the parameter: quite the contrary. So why was such a weight matrix used? Because of computational convenience: under normality assumptions, an assumption of parameter drift, and if an asymptotically optimal estimator such as the MLE is used, the statistic $\hat{T}_{IRLS} = n\hat{F}_{IRLS}$ leads to a non-central Chi-square distribution which allows a straightforward computation of confidence intervals for the RMSEA (31)—see Steiger (1989) and Browne and Cudeck (1993) for details. The use of $\hat{\Gamma}_{NT}^{-1}$ in \hat{T}_{IRLS} is fully justified: it takes into account the sampling variability of the statistics used in the quadratic form. But the use of Γ_{NT}^{-1} in (31) is not. The unstandardized population effect sizes $\sigma - \sigma_0$ are parameters and therefore have no sampling variability. Recently, however, Wu and Browne (2015, see also Satorra, 2015) have provided one such justification. They show that if one is willing to assume that the data arise from a two-level setup in which the sampled population is an instance of a superpopulation, the RMSEA can be interpreted as a “distance” between the sampled population and the superpopulation. In many instances, I feel uncomfortable with this two-level setup because its main result is to widen the confidence intervals, whereas parameter estimates are unaffected. I prefer to rely on the usual setup in which inferences are made solely on the population from which the sample is drawn and, if indeed a superpopulation is of interest, samples from multiple populations from the superpopulation are drawn (as in any two-level model), thereby resulting in different parameter estimates than if only a single sample is used, as in Wu and Browne’s (2015) setup.

Leaving the superpopulation issue aside, the RMSEA parameter cannot be interpreted substantively. What does a population RMSEA of 0.07 mean? We believe it is a value that is not too large, but not small either. But we believe this because of the cutoff values given by Browne and Cudeck, or by Steiger (1989), which are slightly different; if they had not given their famous cutoff values, we would be left stranded with an RMSEA value we would be unable to interpret.

It is the uninterpretable nature of the RMSEA that makes cutoff values necessary. A close reading of Browne and Cudeck (1993) reveals that their cutoff values were largely based on their experience fitting exploratory factor models. However, any given population value of the RMSEA (say 0.05) has a different meaning depending on the structure and size of the model (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Savalei, 2012). In contrast, a standardized effect size such as the CRMR always has the same interpretation: it can be interpreted as approximately the average of the residual correlations, which are in a $(-1, 1)$ interval. As a result, cutoff values for the CRMR can be avoided altogether, as we are able to interpret what a $\text{CRMR} = 0.05$ versus $\text{CRMR} = 0.10$ means.

Steiger (2000) anticipated many of the points made in the present article regarding inconveniences with the population non-centrality parameter and related measures such as the RMSEA. In particular, he pointed out how the fact that the weighting of the residuals is stochastically motivated makes precise population cutoff points problematic.

9. Adjusting for Model Parsimony: Model Selection Versus Goodness-of-Fit

The population RMSEA divides the (weighted) sum of unstandardized effect sizes by degrees of freedom to correct for model complexity. Indeed, models with more parameters will generally lead to smaller overall effect sizes. By doing so, the RMSEA attempts to “reduce the conflict between the desire for parsimony and the desire for close fit” (Browne & Cudeck, 1993: p. 136). Indeed, the sample RMSEA can be used as an index for model selection and also as a test statistic to assess the (approximate) goodness-of-fit of the model.

Model selection involves selecting the best model for the data at hand using a particular criterion. Goodness-of-fit proper addresses the question of how close our model is to the data-generating mechanism. I prefer to separate model selection from goodness-of-fit assessment proper. In applications, often both need to be performed. An investigator who performs model selection first, after selecting the best model for the data at hand, needs to examine the extent to which the selected model mirrors the unknown data-generating mechanism. This involves substantive considerations, but also a goodness-of-fit assessment, as the data at hand is what provides us the actual information on the data-generating mechanism.

On the other hand, an investigator who first assesses the goodness-of-fit of several substantively different models, after determining that more than one model provides a close fit to the data at hand, must select a model. This can involve a model selection statistic, but in my opinion, it will involve mostly substantive considerations. The extreme case, of course, involves choosing among equivalent models (MacCallum, Wegener, Uchino, & Fabrigar, 1993).

To help understand the effect that adjusting for model parsimony has on the overall effect sizes of the misfit, I computed the population $\text{RMSEA}_{\text{NT},2}$ (32), $\text{RMSEA}_{\text{NT},2}$ (31), the SRMR (18), and the CRMR (19) when the discrepancy between the true data-generating mechanism and the fitted model substantively “stays the same” while model size (the number of variables being modeled) increases. The true data-generating mechanism is again a two-independent cluster factor model with equal number of indicators per factor where population factor loading values were in all cases 0.8 and error variances $1 - 0.8^2 = 0.36$, the number of variables ranged from $p = 8$ to $p = 48$, and ρ ranged from 0.8 to 0.9. The null model corresponds to the value of the parameter vector that minimizes the Kullback–Leibler discrepancy (i.e., the ML fitting function under normality)

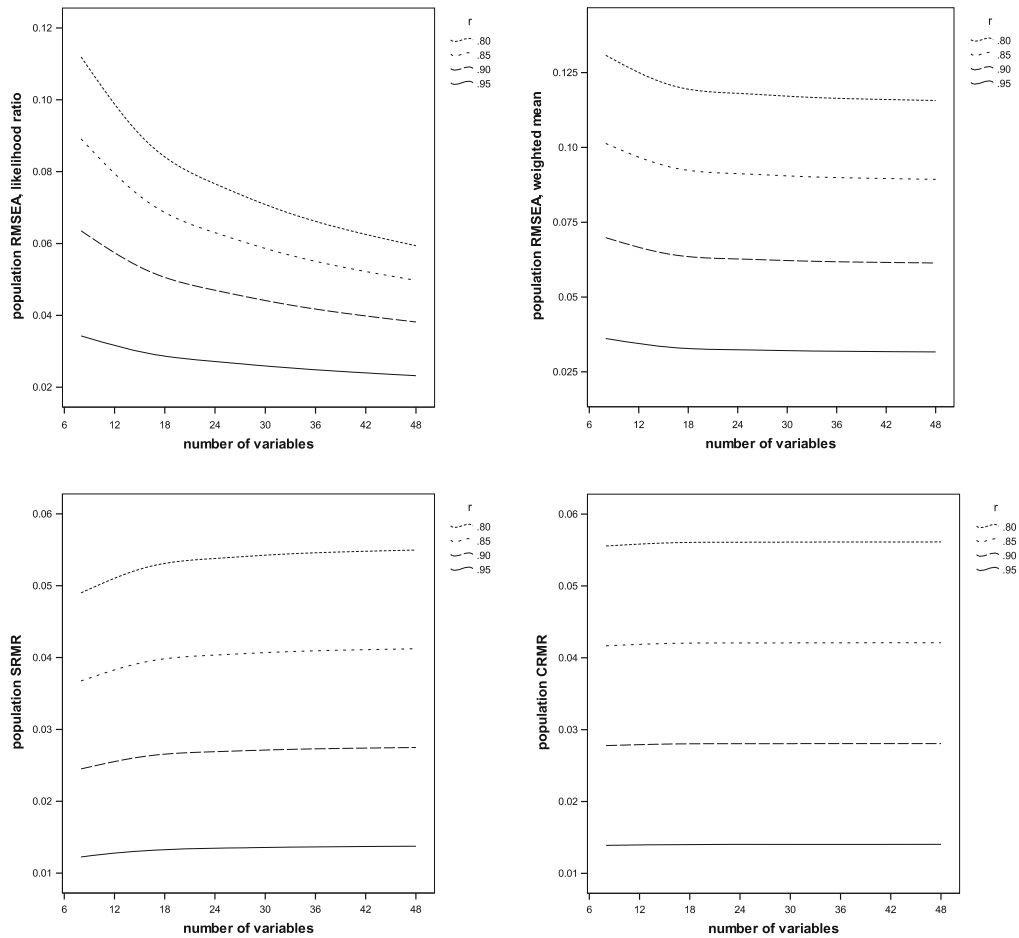


FIGURE 2.

Overall effect size of the misfit when a one-factor model is fitted to an independent clusters two-factor model as the number of variables and the correlation between the factors increase.

between the data-generating model and a one-factor model. Results are presented graphically in Fig. 2. We see in Fig. 2 that when a one-factor model is fitted to an independent cluster factor model and the correlation between the factors is held constant (say $\rho = 0.85$), as the number of indicators per factor increases the population $RMSEA_{NT,2}$ decreases. This implies that if a criterion such as a $RMSEA = 0.05$ is used to retain a model, and if the $RMSEA$ is estimated using the likelihood ratio test statistic, it is easier to retain a model because the corresponding population parameter decreases as model size increases even when the discrepancy between the data-generating mechanism and the fitted model “stays the same.” In contrast, in this setup the value of the $RMSEA_{NT,1}$ parameter also decreases from $p = 8$ to 16 and then flattens. Thus, the $RMSEA_{NT,1}$ is preferable in this setup. The $SRMR$ shows the opposite pattern to the $RMSEA_{NT,1}$: it increases from $p = 8$ to 16 and then flattens. Finally, the $CRMR$ shows the pattern that I would like to see in this setup: it reflects the fact that the overall effect of the misfit does not depend on model size, it only depends on ρ , the correlation between the factors.

This example serves to illustrate the following point: Almost invariably, a fixed cutoff for the $RMSEA$, such as $RMSEA = 0.05$, is used. Also, almost invariably the $RMSEA$ is estimated using the likelihood ratio test (possibly after a mean and variance correction to account for non-

normality). In my experience (see also Kenny & McCoach, 2003), in these situations it is easier to retain the model when a large number of variables is used (say, $p = 50$, rather than, say, $p = 10$). This runs counter to the obvious fact that finding a well-fitting model is increasingly difficult as the number of variables to be modeled increases. In contrast, an average of standardized effect sizes, such as the CRMR, is invariant to increasing model size as the discrepancy between the data-generating mechanism and the fitted model “stays the same.” As a result, one should expect larger values of the SRMR or CRMR as the number of variables being modeled increases.

10. Assessing the Goodness-of-Fit of a SEM Model in Practice

Figure 3 displays graphically a “typical” SEM model, taken from Alden, Steenkamp, and Batra (2006).¹ In this model, a number of indicators are used to measure the variables “media,” “migration,” “SNI,” “materia,” “ethno,” and “GBA” without error. In addition, the variables “GCO attitude” and “GCO intens” are measured with error, and a number of variables (e.g., age) are used as covariates. The hypotheses of interest are displayed graphically in Fig. 4.

The results obtained using robust methods are very similar to those obtained under normality assumptions, and here I just report the results obtained under normality. The R package lavaan (Rosseel, 2012) was used to obtain all the results presented here. The likelihood ratio statistic yields 1,090.99 on 267 df , $p = 0$. We conclude that the proposed model is not the data-generating mechanism. This is not a surprising outcome, given the streamlined nature of the model and the number of variables being modeled.

How far away are we from the data-generating mechanism? In other words, what is the size of the misfit? If the size of the misfit is trivial, we may decide to ignore it. The current standard for judging the size of model misfit (adjusted for model parsimony) is to use the RMSEA. A 90% confidence interval (CI) for the RMSEA based on the likelihood ratio test yields (0.061, 0.069), the point estimate is 0.065, and notice that another CI could have been constructed using T_{IRLS} , yielding a slightly different result. Does this estimate of the RMSEA reveal a large or a small degree of misfit? We do not know. All we have are Browne and Cudeck’s famous cutoffs, based on their experience. But their experience is limited to just a handful of models and misspecifications. As Savalei (2012) has shown, a RMSEA value (say 0.065) may indicate a large or small misfit depending on a number of factors such as the type of model, the number of variables being modeled, and so on. The magnitude of the RMSEA is uninterpretable because it is based on a weighted combination of unstandardized residuals. But the RMSEA also aims for model parsimony. Aiming for model parsimony makes sense in exploratory factor analysis, where we do not want to add another factor to the model to increase fit when the number of general factors may be correctly specified, but the model misfits for other reasons, such as the need of correlated residuals. But in a general model such as the one displayed in Fig. 3, do we want our model to be parsimonious? I think the answer is no. The aim of the study is to obtain accurate parameter estimates and standard errors to make inferences about the hypotheses outlined in Fig. 4.

Alternatively, the overall magnitude of the model misfit can be gauged using the CRMR (or the SRMR) without the need for any cutoff and without adjusting for model parsimony. The unbiased estimate of the CRMR, given by (24), equals 0.065, and a 90% CI for the population parameter is (0.059, 0.072). Notice that in this example the point estimates of the RMSEA and CRMR are equal (to three digits), whereas the CI for the CRMR is wider. Using the CRMR, is the magnitude of this model misfit large? In applications, I disregard statistically significant correlations smaller than $|0.1|$ as being of no substantive interest. I use the same criterion to judge the magnitude of residual correlations: I cannot disregard a statistically significant residual correlation larger than $|0.1|$. Of

¹I am indebted to JBS Steenkamp for providing the data used in this example.

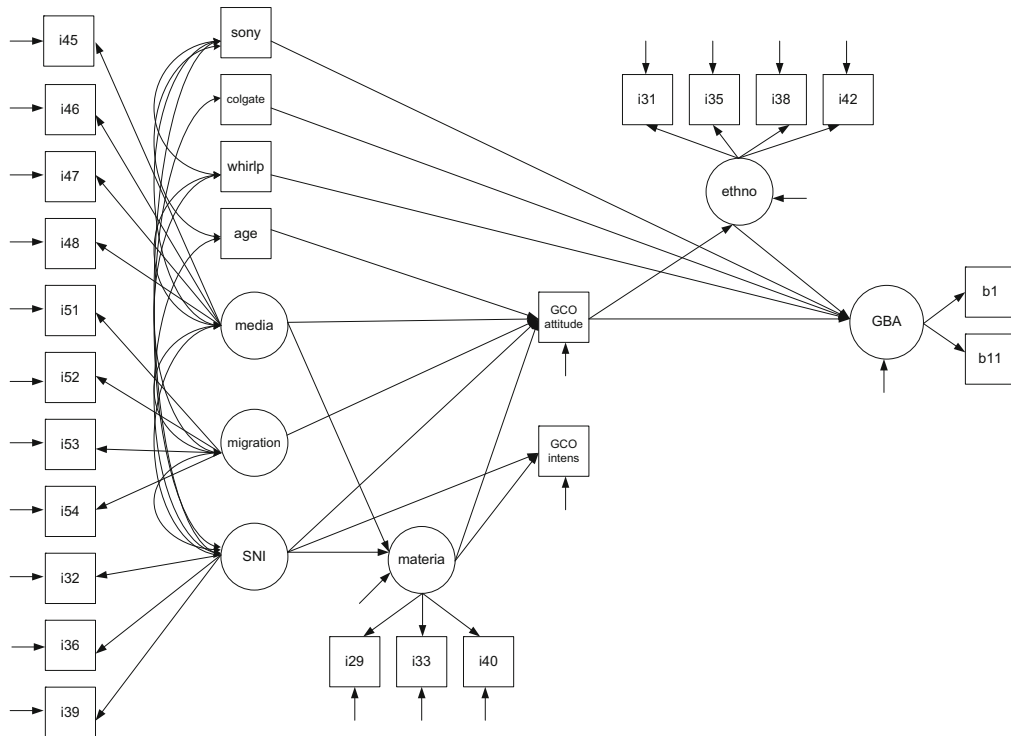


FIGURE 3. "Typical" SEM model, taken from Alden et al. (2006).

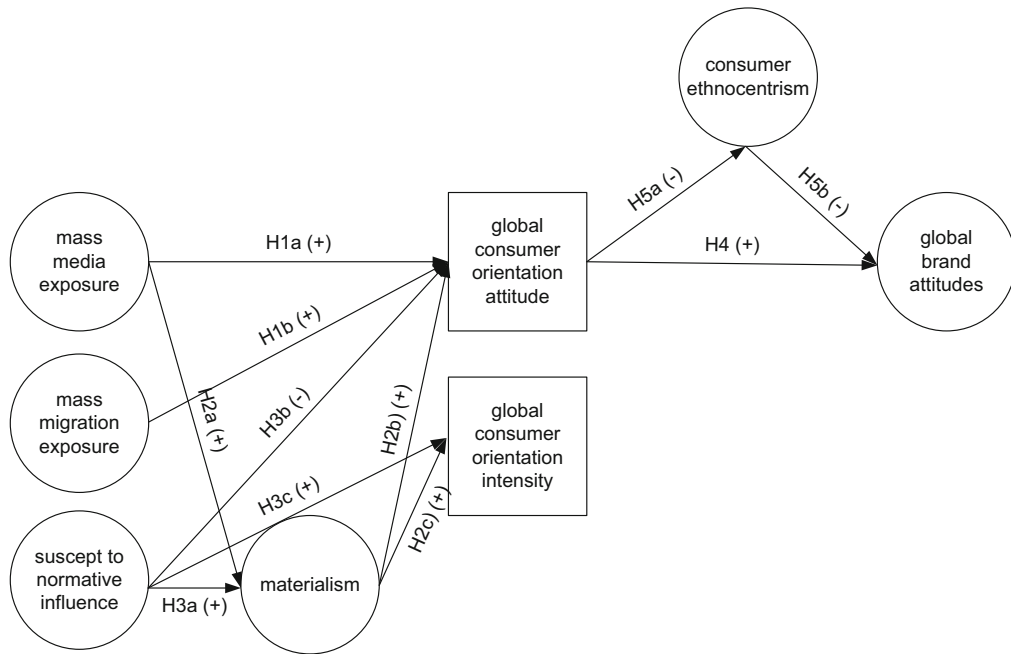


FIGURE 4. Alden et al.'s (2006) substantive hypotheses.

course, the smaller the residual correlations the better, and I consider that the degree of misfit of a model is negligible when all residual correlations are smaller than 0.05. Therefore, I consider that a model fits closely when $CRMR \leq 0.05$, provided no statistically significant residual correlation is larger $|0.1|$. Do I consider a CRMR value of 0.065 to be acceptable? I cannot tell without examining the largest residual correlations (in absolute values), as the CRMR is only (roughly) the average of the residual correlations. This model yields many large residual correlations: nine of them are larger than $|0.20|$, and they cannot be attributed to chance. I conclude that the model fits poorly.

It has repeatedly been advised (Jöreskog & Sörbom, 1988; McDonald & Ho, 2002; Raykov, 2000) that all standardized residual covariances (or residual correlations) be examined carefully. To do so, the variables need to be ordered in consonance with the fitted model. No pattern between these residuals should be apparent: rather, they should be well scattered. The matrix of residuals will often reveal structural zeroes, residual correlations that must be zero under the fitted model and for the choice of estimator used. When there is a large number of structural zeroes in the model, the SRMR and CRMR do not provide an accurate representation of the overall fit of the model.

Researchers should examine the extent to which standardized residual covariances (or residual correlations) can be attributed to chance using confidence intervals, or z statistics based on their standard errors. In so doing, it is necessary to account for multiple testing, for instance using a Bonferroni correction.

To understand how the model misfits, I often find it helpful to compute the average of the absolute value of the residual correlations for each variable (Maydeu-Olivares, 2015). In this case, the average corresponding to *i54* (one of the indicators of “migra”—mass mediation exposure), 0.138, is much larger than the averages for the other variables (which range from 0.013 to 0.087). The model needs to be modified to reduce the size of the residual correlations involving this variable. An alternative is to remove the variable from the model as it is not a key variable.

How to modify the model to obtain a better fit? I would begin by inspecting graphs of individuals’ residuals versus expected values for each equation in the model, just as in regression. These may be used to assess the linearity and homoscedasticity assumptions in each equation. In particular, the presence of heteroscedasticity may be an indication of a moderating relationship, possibly involving another variable in the dataset. Only after I am satisfied that these critical assumptions are met would I explore the modification indices (score tests) for the model in combination with their power and the expected parameter change, as described in Saris et al. (2009). Their procedure requires a fair amount of work, particularly with a model of this size, but this is as it should be: fitting a set of equations (possibly involving latent variables) involves considerably more work than fitting a single regression equation.

Hopefully, the application of Saris et al.’s (2009) method will shed some light into the validity of our inferences on the substantive hypotheses of interest outlined in Fig. 4. What do the residual correlations (or standardized residuals)—the size of the model misfit—tell us about the validity of our inferences? This matter requires further research.

11. Remarks on Some Recurring Themes in SEM Goodness-of-Fit Assessment

11.1. *The Role of Sample Size*

The role of sample size in assessing the goodness-of-fit of a SEM model is as it should be, and it is no different from a regression model or a simple z -test: increasing sample size leads to increasing power to detect discrepancies between the model and the data-generating mechanism. This is a good thing. Sample size should be as large as possible. I have yet to encounter a z -test application involving a mean difference in which researchers complained that their power was too high. And yes, the value of all test statistics (including the z test) depends on sample size,

not just the value of the Chi-square statistic in SEM. Nonetheless, one encounters numerous applications in which the dependency of the SEM's Chi-square test statistic on sample size is used to disregard its result, based on grounds of excessive power, without actually checking the power of the test against any meaningful alternative. In any case, the "antidote" against excessive power is to examine the effect size (in our case, the effect size of the misfit).

11.2. Relative Effect Sizes of the Overall Misfit

There exist population counterparts of the CFI (Bentler, 1990), and of the GFI (Steiger, 1990), as well of other popular relative goodness-of-fit indices. Therefore, we could use a parameter to convey the magnitude of the misfit of a model relative to that of another model. Should we do so? We could, but by construction, they are relative effect sizes of the misfit. Relative overall misfit effect sizes can be used to compare two substantively motivated models, or a substantively motivated model against a standard baseline model. The standard baseline model used in the CFI is the independence model, and the one used in the GFI is the saturated model. I believe that the independence model should never be used as a baseline model, because if, as our theory suggests, we believe that our variables are correlated and we model their dependencies, why do we use an independence model as a baseline to communicate how well we are modeling these dependencies? Put differently, if we believe that our variables are uncorrelated (our baseline model), why are we modeling their dependencies? The use of a saturated model as a standard baseline has some justification. For if the core features of our model are correctly specified (and this is a strong assumption) then hopefully our parameter estimates will be unbiased when a restricted model is fitted, and therefore it is of some interest to gauge our fitted model against a saturated model. In any case, if the relative effect sizes of the overall misfit are of interest, confidence intervals for the population parameters of interest should be used, not (possibly biased) point estimates.

12. Conclusions

As in any modeling enterprise, such as fitting a regression model, goodness-of-fit assessment in SEM should be performed judiciously. We wish to reproduce the features of the data-generating process, but not the idiosyncratic characteristics of the sample. The best way to distinguish them is by replication. Regrettably, in SEM not much value is assigned to replication studies, with the possible exception of factor analysis studies. In addition, it is often difficult to replicate the conditions of previous studies due to the non-experimental nature of most studies. By construction, a perfect replication study is a cross-validation. Since SEM estimation methods are asymptotic in nature, researchers use a sample as large as possible, with no data to spare in a cross-validation. However, with the increasing sample sizes available in some planned large scale studies, I expect that cross-validation methods in SEM will be more frequently used. This is good, because SEM is often used to generate new theoretical models (Jöreskog, 1993). The researcher begins with a tentative initial model. When the initial model does not fit well, the researcher repeatedly modifies the model using her substantive theory and the model-fit results, until a better fitting model is found that is a reasonable compromise between theory and data. There is nothing wrong with using SEM in this fashion. However, we need cross-validation (or replication) to ensure that we have not captured idiosyncrasies in the data. Browne (2000) and Browne and Cudeck (1989) are excellent sources on cross-validation methods.

Goodness-of-fit assessment is not an end in itself; it is simply a means to draw inferences about the world (Sarlis et al., 2009). The substantive inferences we make are the final objective of our analysis. There has been so much abuse of SEM fit indices as test statistics that one may be tempted to avoid examining the goodness-of-fit of the model altogether and simply perform

model selection. However, I believe that model selection proper does not tell us anything about the validity of the inferences we make.

Assessing the goodness-of-fit of a SEM model is a time-consuming process. The time required is a function of a number of determinants, such as the complexity of the model, and the presence of latent variables. But to a large extent, it is a function of the number of equations in the model. A researcher modeling ten equations simultaneously in a SEM should expect to devote at least a tenfold increase of the time devoted to check the model assumptions of a single regression equation.

When only a few equations are being estimated, researchers should be able to find a well-fitting model. However, when fifty equations are being modeled, as is often the case in the social sciences, it is not realistic to expect to find a well-fitting model in the amount of time a researcher can devote to the enterprise and as a result, she is obliged to settle for a model that fits closely. Thus, it is time constraints and the number of equations being modeled that justify the use of approximate fit assessment, not sample size considerations. There is nothing wrong with the notion of testing for approximate fit in large models, as opposed to testing for exact fit, but one has to realize that an approximate model is, by definition, incorrect and that all inferences based on an approximate model are suspect (Yuan et al., 2003).

Here, I have introduced statistical theory for assessing the magnitude of the misfit in SEM models using standardized residual covariances and residual correlations. For if a model is rejected and due to time constraints we cannot find a well-fitting model, we may be able to gauge how far we are from the data-generating process and decide whether to retain it as a “close fitting model” based on a confidence interval for the SRMR or the CRMR. Alternatively, in a classical fashion one can use a test of close fit for these parameters. Because the interpretation of these parameters is straightforward, any arbitrary value, such as 0.05, can be used in such a test. However, as SEM is a multivariate technique, it is only meaningful to report the overall degree of misfit of the model if none of the standardized residual covariances (or residual correlations) is too large ($< |0.1|$ is my personal criterion). Standardized residual covariances (or residual correlations) are seldom reported in applications. I would urge researchers to report the six largest standardized residual covariances (or residual correlations)—in absolute values—in applications, along with their statistical significance or CIs, for the SRMR and CRMR are simply summary measures of the model misfit.

The use of the sample SRMR and CRMR should be avoided, as they are biased upwards estimates of their corresponding parameters. For instance, the sample CRMR estimate in the application described previously is 0.072, $N = 725$, but this equals the upper bound of the 90% CI for the population parameter; the asymptotically unbiased estimate of the population parameter is 0.065.

Hopefully, this presentation will spur further research on this topic. An immediate topic that requires attention is whether we should use standardized residual covariances or residual correlations. Personally, I have a mild preference for residual correlations, as their interpretation is more straightforward than for standardized residual covariances. However, their use in models that are not scale invariant needs to be investigated. Another topic that requires research is how to assess the degree of misfit in models with a mean structure. It seems to me that it is perhaps best to separate the model misfit in the mean structure from the model misfit in the covariance structure, rather than combining them. For the mean structure, a suitable standardized parameter

is $\iota_\mu = \sqrt{\frac{1}{p} \sum_{i=1}^p \frac{(\mu_i - \mu_i^0)^2}{\sigma_i^2}}$ where μ_i and σ_i denote the unknown population mean and standard deviation and μ_i^0 denote the population mean under the fitted model. In models for ordinal data (Maydeu-Olivares, 2006; Muthén, 1984) and more generally in item response theory (IRT) models for ordinal data, the CRMR can certainly be used to gauge the size of model misfit at the bivariate level (Maydeu-Olivares, 2015).

But, above all, when due to time constraints we are obliged to settle for a model that fits closely, what do the effect sizes of the misfit tell us about the validity of the inferences we make? This is the area where I would like to see much more research.

Appendix

Asymptotic Distribution of Standardized Residual Covariances and Residual Correlations

From standard theory (e.g., Satorra, 1989)

$$\sqrt{N}(\mathbf{s} - \hat{\boldsymbol{\sigma}}) \stackrel{a}{=} \mathbf{H}\sqrt{N}(\mathbf{s} - \boldsymbol{\sigma}_0) \quad (34)$$

and the asymptotic distribution of the unstandardized (raw) residual covariances follows from (14) under parameter drift assumptions. The asymptotic distribution of the standardized residuals is obtained from a first-order Taylor expansion of $\hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})$ at $\mathbf{s} = \boldsymbol{\sigma}$ and $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$

$$\begin{aligned} \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}}) &\approx \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})\Big|_{\mathbf{s}=\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} + \frac{\partial \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})}{\partial \mathbf{s}'}\Big|_{\mathbf{s}=\boldsymbol{\sigma}} (\mathbf{s} - \boldsymbol{\sigma}) + \frac{\partial \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})}{\partial \hat{\boldsymbol{\sigma}}'}\Big|_{\hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} (\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \\ &= \mathbf{G}(\mathbf{s} - \boldsymbol{\sigma}) - \mathbf{G}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) = \mathbf{G}(\mathbf{s} - \hat{\boldsymbol{\sigma}}), \end{aligned} \quad (35)$$

since $\hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})\Big|_{\mathbf{s}=\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} = \mathbf{0}$, $\frac{\partial \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})}{\partial \mathbf{s}'}\Big|_{\mathbf{s}=\boldsymbol{\sigma}} = \mathbf{G}$ and $\frac{\partial \hat{\mathbf{G}}(\mathbf{s} - \hat{\boldsymbol{\sigma}})}{\partial \hat{\boldsymbol{\sigma}}'}\Big|_{\hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} = -\mathbf{G}$. Equation (16) follows from (14), (34), and (35).

The asymptotic distribution of the residual correlations is obtained in a similar fashion. Using a first-order Taylor expansion of $\mathbf{r} - \hat{\boldsymbol{\rho}}$ at $\mathbf{s} = \boldsymbol{\sigma}$ and $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$

$$\begin{aligned} (\mathbf{r} - \hat{\boldsymbol{\rho}}) &\approx (\mathbf{r} - \hat{\boldsymbol{\rho}})\Big|_{\mathbf{s}=\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} + \frac{\partial (\mathbf{r} - \hat{\boldsymbol{\rho}})}{\partial \mathbf{s}'}\Big|_{\mathbf{s}=\boldsymbol{\sigma}} (\mathbf{s} - \boldsymbol{\sigma}) + \frac{\partial (\mathbf{r} - \hat{\boldsymbol{\rho}})}{\partial \hat{\boldsymbol{\sigma}}'}\Big|_{\hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} (\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \\ &= \mathbf{F}(\mathbf{s} - \boldsymbol{\sigma}) - \mathbf{F}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) = \mathbf{F}(\mathbf{s} - \hat{\boldsymbol{\sigma}}), \end{aligned} \quad (36)$$

where $(\mathbf{r} - \hat{\boldsymbol{\rho}})\Big|_{\mathbf{s}=\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} = \mathbf{0}$, $\frac{\partial (\mathbf{r} - \hat{\boldsymbol{\rho}})}{\partial \mathbf{s}'}\Big|_{\mathbf{s}=\boldsymbol{\sigma}} = \mathbf{F}$, $\frac{\partial (\mathbf{r} - \hat{\boldsymbol{\rho}})}{\partial \hat{\boldsymbol{\sigma}}'}\Big|_{\hat{\boldsymbol{\sigma}}=\boldsymbol{\sigma}} = -\mathbf{F}$. The nonzero elements of \mathbf{F} , $\partial(r_{ij} - \hat{\rho}_{ij})/\partial \hat{\sigma}_{lk}$ evaluated at $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}$, are $-\sigma_{ij}/(2\sigma_{ii}^{3/2}\sigma_{jj}^{1/2}) = \frac{\rho_{ij}}{2\sigma_{jj}}$ if $i \neq k, j = l$, $-\sigma_{ij}/(2\sigma_{ii}^{1/2}\sigma_{jj}^{3/2}) = \frac{\rho_{ij}}{2\sigma_{ii}}$ if $i = k, j \neq l$, and $1/(\sigma_{ii}^{1/2}\sigma_{jj}^{1/2})$ if $(i, j) = (k, l)$. Notice that \mathbf{F} can be written as $\mathbf{F} = \mathbf{J}\mathbf{G}$, where the nonzero elements of \mathbf{J} are $-\frac{\rho_{ij}}{2}$ if $i \neq k, j = l$ or $i = k, j \neq l$, and 1 if $(i, j) = (k, l)$.

Asymptotic Mean and Variance of the Overall Effect Size Estimators

The asymptotic mean and variance of $T = T_s$ or T_r under the parameter drift assumptions given in Eqs. (20) and (21) follow from standard results in quadratic forms of normal variates (e.g., Schott, 1997: Theorem 9.22).

The asymptotic mean and variance of \tilde{t} follow from standard results on Taylor expansions of moments of functions of random variables: Let T be a random variable with μ_T and σ_T^2 , then

$$E[f(T)] \approx f(\mu_T) + \frac{f''(\mu_T)}{2}\sigma_T^2, \quad (37)$$

$$\text{var}[f(T)] \approx [f'(\mu_T)]^2\sigma_T^2. \quad (38)$$

For the \tilde{t}_s based on the SRMR, $f(T) = \sqrt{\frac{T-c}{t}}$ and $f'(T) = \frac{1}{2t} \left(\frac{T-c}{t}\right)^{-\frac{1}{2}}$ and $f''(T) = -\frac{1}{4t^2} \left(\frac{T-c}{t}\right)^{-\frac{3}{2}}$. For the \tilde{t}_r based on the CRMR, $t - p$ is used instead of t . Equations (22) and (23) are obtained from (37) and (38) evaluating these derivatives at μ_T given in Eq. (20).

The asymptotic mean and variance of $\hat{\iota} = k^{-1}\tilde{\iota}$ are obtained from the asymptotic mean and variance of $\tilde{\iota}$ again using (37) and (38): $E[\hat{\iota}] = \iota$, $\text{var}[\hat{\iota}] = k^{-2}\text{var}[\tilde{\iota}]$.

Asymptotic Mean and Variance of the Sample SRMR and CRMR Under the Null Hypothesis of Exact Fit

Consider the statistic $T^* = \widehat{\text{SRMR}}$ or $\widehat{\text{CRMR}}$ given in Eqs. (7) and (8). We write T^* as $f(T) = \sqrt{T/t}$, where $T = T_s$ or T_r , and for T_r $t - p$ is used instead of t . Using (37) and (38) with $f'(T) = \frac{1}{2t} \left(\frac{T}{t}\right)^{-\frac{1}{2}}$ and $f''(T) = -\frac{1}{4T^2} \left(\frac{T}{t}\right)^{\frac{1}{2}}$,

$$E(T^*) = \sqrt{\frac{\mu_T}{t} \frac{8\mu_T^2 - \sigma_T^2}{8\mu_T^2}}, \quad \text{var}(T^*) = \frac{\sigma_T^2}{4t \times \mu_T^2}. \quad (39)$$

Now, under the null hypothesis of exact fit

$$E(T) = \mu_T = \text{tr}(\Xi), \quad \text{var}(T) = \sigma_T^2 = 2\text{tr}(\Xi^2). \quad (40)$$

Equations (28) and (29) follow from (39) and (40).

References

- Alden, D. L., Steenkamp, J. B. E. M., & Batra, R. (2006). Consumer attitudes toward marketplace globalization: Structure, antecedents and consequences. *International Journal of Research in Marketing*, 23(3), 227–239. doi:10.1016/j.ijresmar.2006.01.010.
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815–824. doi:10.1016/j.paid.2006.09.018.
- 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. (1995). EQS 5 [Computer Program]. Encino, CA: Multivariate Software Inc.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, 21, 235. doi:10.2307/270937.
- Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108–132. doi:10.1006/jmps.1999.1279.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge: Cambridge University Press.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x.
- Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean- and covariance-structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum.
- Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24(4), 445–455. doi:10.1207/s15327906mbr2404_4.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in rmsea test statistic in structural equation models. *Sociological Methods & Research*, 36(4), 462–494. doi:10.1177/0049124108314720.
- Coffman, D., & Millsap, R. (2006). Evaluating latent growth curve models using individual fit statistics. *Structural Equation Modeling*, 13(1), 37–41. doi:10.1207/s15328007sem1301_1.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1–22. doi:10.1037/1082-989X.12.1.1.
- Hildreth, L. (2013). *Residual analysis for structural equation modeling*. Ph.D. dissertation, Statistics Department, Iowa State University.
- 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.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
- Jöreskog, K. G., & Sörbom, D. (1981). Analysis of linear structural relationships by maximum likelihood and least squares methods. Research report 81-8, Uppsala, Sweden.
- Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7. A guide to the program and applications* (2nd ed.). Chicago, IL: International Education Services.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 333–351. doi:10.1207/S15328007SEM1003_1.
- Lee, S. Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44*(1), 99–113. doi:10.1007/BF02293789.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*(1), 185–199. doi:10.1037/0033-2909.114.1.185.
- Marsh, H., Hau, K., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.
- Marsh, H. W., Hau, K., & 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: A Multidisciplinary Journal, 11*(3), 320–341. doi:10.1207/s15328007sem1103_2.
- Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika, 71*(1), 57–77. doi:10.1007/s11336-005-0773-4.
- Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111–127). New York: Routledge.
- Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models: Standardized residual covariances and residual correlations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82. doi:10.1037/1082-989X.7.1.64.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115–132. doi:10.1007/BF02294210.
- Ogasawara, H. (2001). Standard errors of fit indices using residuals in structural equation modeling. *Psychometrika, 66*(3), 421–436. doi:10.1007/BF02294443.
- Raykov, T. (2000). On sensitivity of structural equation modeling to latent relation misspecifications. *Structural Equation Modeling, 7*(4), 596–607. doi:10.1207/S15328007SEM0704_4.
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5–26. doi:10.1177/0013164412449831.
- 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.
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 561–582. doi:10.1080/10705510903203433.
- Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54*(1), 131–151. doi:10.1007/BF02294453.
- Satorra, A. (2015). A comment on a paper by H. Wu and M. W. Browne (2014). *Psychometrika, 80*(3), 613–618. doi:10.1007/s11336-015-9455-z.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis. Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910–932. doi:10.1177/0013164412452564.
- Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum.
- Schott, J. R. (1997). *Matrix analysis for statistics*. New York: Wiley.

- Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH* [Computer program manual]. Evanston, IL: Systat Inc.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25(2), 173–180. doi:[10.1207/s15327906mbr2502_4](https://doi.org/10.1207/s15327906mbr2502_4).
- Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduck and Glaser. *Structural Equation Modeling*, 7(2), 149–162. doi:[10.1207/S15328007SEM0702_1](https://doi.org/10.1207/S15328007SEM0702_1).
- Steiger, J. H., & Fouladi, R. (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.
- Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Meeting of the Psychometric Society, Iowa.
- Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, 42(2), 233–239. doi:[10.1111/j.2044-8317.1989.tb00912.x](https://doi.org/10.1111/j.2044-8317.1989.tb00912.x).
- Wu, H., & Browne, M. W. (2015). Quantifying adventitious error in a covariance structure as a random effect. *Psychometrika*, 80(3), 571–600. doi:[10.1007/s11336-015-9451-3](https://doi.org/10.1007/s11336-015-9451-3).
- Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, 92, 767–774. doi:[10.1080/01621459.1997.10474029](https://doi.org/10.1080/01621459.1997.10474029).
- Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *The British Journal of Mathematical and Statistical Psychology*, 51(2), 289–309. doi:[10.1111/j.2044-8317.1998.tb00682.x](https://doi.org/10.1111/j.2044-8317.1998.tb00682.x).
- Yuan, K.-H., & Bentler, P. M. (1999). *F* tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, 24(3), 225–243. doi:[10.3102/10769986024003225](https://doi.org/10.3102/10769986024003225).
- Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods*, 15(4), 335–351. doi:[10.1037/a0020140](https://doi.org/10.1037/a0020140).
- Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology*, 33(1), 241–265. doi:[10.1111/j.0081-1750.2003.00132.x](https://doi.org/10.1111/j.0081-1750.2003.00132.x).

Manuscript Received: 1 AUG 2016

Final Version Received: 9 DEC 2016

Published Online Date: 7 FEB 2017