How I Evaluate Fit of Structural Equation Models

Rich Jones
Last update: April 29, 2024

Abstract

The first rule of evaluating fit of SEM models is, no single set of rules and cutoffs will be appropriate in all situations, and the most important consideration is the substantive interpretation of model results together with guiding theory regarding the structure of the model. I will interpret good fit of the model to the data when the sample size is at least 200 and the standardized root mean square residual (SRMR) ≤ 0.08 and there are no large (>.2)[1] standardized residuals. I will also report the RMSEA and CFI, and identify that it is desirable to have a confirmatory fit index (CFI) of 0.95 or greater and a root mean squared error of approximation (RMSEA) of < 0.05. I will primarily be motivated to make model modifications on the basis lack of fit as suggested by the individual residuals and/or the SRMR. If the sample size is less than 200, and in contexts where the inferences are exploratory rather than confirmatory (e.g., exploratory factor analysis), I will rely upon the CFI as the leading indicator of fit.

Table 1. Assessing SEM Model Fit

Statistic	Meaning	Interpretation	Caution
SRMR Standardized root mean square residual	The mean standardized residual among mean and covariance estimates for model-implied relative to observed values.	Infer "approximate fit" when SRMR is ≤ 0.08 and there are no "large" residuals. The computation is based on residuals only, and does not directly incorporate the sample size or model complexity.	Good models should have low SRMR and low residuals, but because SRMR does not incorporate model complexity, the SRMR will not help identify over-parameterized models. Should not be used if the sample size is small (<200); should not be interpreted without examination of individual residuals.
RMSEA Root mean squared error of approximation	The model discrepancy per degree of freedom	Infer good fit when less than 0.05.	The RMSEA is sample size dependent. It may not be appropriate when (a) the model has low degrees of freedom, (b) in small samples. The RMSEA will be lower in larger samples.
CFI Confirmatory fit index	Compares estimated model to a hypothetical null baseline model, usually defined as a model with all variables mutually uncorrelated.	Infer good fit when greater than 0.95.	Compares obtained model to a hypothetical model which may be known to be incorrect, and as such probably better for settings where the analysis is exploratory rather than confirmatory.

[1] I have no authoritative source for 0.2 identifying "large" residuals. This threshold needs additional research.

Introduction

The first rule of evaluating fit of a structural equation model (SEM) is: no single set of rules and cutoffs will be appropriate in all situations. This is because (a) fit statistics are variably influenced by many factors besides model-data fit and (b) conventions for judging fit have been derived from a limited and arguably simple set of models that may not be completely relevant to the goals of the current analysis, and (c) good analysis decisions are unlikely when applying – given the nearly infinite variety of data and model types – arbitrary thresholds for binary decision making on continuously distributed sample statistics (Greiff & Heene, 2017). The first and most important step in evaluating model fit is therefore a critical review of the substantive implication of model estimates together with guiding theory regarding the structure of the model. Critical and careful handling of data (addressing non-normality, sparseness, and collinearity or endogeneity) and visual inspection of substantive results and addressing obvious challenges in estimation (negative residual variances, Heywood cases) and careful stepwise building of complex models by separate evaluation of sub-model components may, with all due luck, avert challenges in model fit variably revealed in global fit statistics.

After the data are checked carefully, a model has been fit as guided by substantive theory, and estimates obtained have been evaluated for consistency with that theory, then it may be appropriate to evaluate the fit of a model with global fit statistics. To describe the fit of structural equation models, there are what have been called **absolute** and **incremental** (or sometimes called relative) fit indices <cite>. Absolute fit indices describe the degree to which the model results agree with observed data. Incremental or relative fit indices describe the degree to which the model results correspond to another model, one that makes minimal assumptions about the relationships among the variables (variously referred to as a baseline model, independence model, or null model). Although there are a great number of available fit indices available in structural equation modeling, I place emphasis on an absolute fit index computed directly from the residuals: the standardized root mean square residual, or SRMR. This is interpreted alongside the individual residuals, and another absolute fit index that accounts for model complexity (the root mean squared error of approximation, RMSEA), and an incremental fit index, the confirmatory fit index (CFI).

Absolute fit indices: SRMR, χ^2, RMSEA

Absolute fit indices include the model chi-square (χ^2), the root mean squared error of approximation (RMSEA), and the standardized root mean square residual (SRMR).[2]

SRMR

Conceptually, the SRMR conveys the mean absolute difference between observed and model estimated standardized means (if included in the model) and standardized covariances (i.e., correlations) (Beaujean, 2014). Values can be interpreted as the average population residual

[2] Asparouhov & Muthén (2018) describe the SRMR as an "approximate" fit index, along with the CFI, contrasting with the χ^2 which is a test of exact fit

correlation for the model being estimated. There are several different formulations of the SRMR in the literature (e.g., R/lavaan produces at least 7 different flavors of SRMR (Beaujean, 2014)). Describing the SRMR computed in the Mplus framework, Asparouhov & Muthén (2018) provide the formula for the SRMR in the all-continuous-dependent variable case SRMR as:

$$\text{SRMR} = \sqrt{\frac{\text{S}}{p + p(p+1)/2}}$$

where p is the number of variables in the model, and $p(p+1)/2$ provides the number of non-redundant elements in the variance-covariance matrix, and S is

$$\sum_{j=1}^{p}\sum_{k=1}^{j-1}\left(\frac{s_{jk}}{\sqrt{s_{jk}s_{kk}}} - \frac{\sigma_{jk}}{\sqrt{\sigma_{jk}\sigma_{kk}}}\right)^2 +$$

$$\sum_{j=1}^{p}\left(\frac{m_j}{\sqrt{s_{jj}}} - \frac{\mu_j}{\sqrt{\sigma_{jj}}}\right)^2 +$$

$$\sum_{j=1}^{p}\left(\frac{s_{jj} - \sigma_{jj}}{s_{jj}}\right)^2$$

Where s, σ are model-estimated and observed values for the variances and covariances, respectively. Further, m, μ are model-estimated and observed values for means, respectively. The SRMR is the mean squared error between the target model and all variables freely correlated model for correlations (standardized covariances), standardized means, and standardized residuals. The Mplus SRMR is distinct in that means and are explicitly incorporated into the computation (different software packages may provide different estimates for SRMR if the SRMR does not include the means).

The fit statistic does not incorporate model complexity, meaning there is no adjustment for model degrees of freedom. The fit statistic also does not directly include the model sample size.

Approximate fit. The ideal value is 0 (perfect fit). Hu & Bentler (1999) suggest values less than .08 be taken as indicating good fit. Asparouhov & Muthén (2018) recommend that approximate fit can be inferred when the SRMR is ≤ 0.08 and there are no individual residuals that are large. They caution that the SRMR should not be used when the sample size is small (N<200) and recommended when the sample size is large (N>500).

Model chi-square (χ^2)

The χ^2 statistic is computed based on the discrepancy of observed and model-implied mean and covariance matrices. As a test of the hypothesis that the specified model generated the data, the χ^2 is generally considered to be over-powered in sample sizes large enough to support stable estimates of covariance parameters. Therefore, the resulting P-value is rarely considered in a strict way, where larger χ^2 statistics and smaller P-values indicate greater discrepancy between the observed and model-implied mean and covariance matrices. Nevertheless, the χ^2 forms the basis of many SEM fit statistics, including the RMSEA and CFI (but not the SRMR).

<u>RMSEA</u>

The RMSEA which is based on the χ^2 and can be viewed as an index of model discrepancy per degree of freedom. It is computed as

$$\mathrm{RMSEA} = \frac{\sqrt{\chi^2 - df}}{\sqrt{df(N-1)}}$$

where df is the model degrees of freedom (a function of the number of variables and number of parameters estimated) and N is the sample size. This fit index is sensitive to the number of parameters and sample size. Ideal values approach 0 and when the computation results in a value is less than 0, the RMSEA is set to 0. Hu & Bentler (1999) suggests a cutoff value close to 0.06, MacCallum, Browne and Sugawara (1996) suggest a RMSEA value less than 0.05 indicates good fit, and MacCallum and colleagues (1986) suggest values less than 0.08 indicate mediocre fit. So: you could be strict and pick .05 and cite MacCallum et al or give yourself a little bit of a break and pick .06 and cite Hu and Bentler.

The RMSEA is not recommended for small degree of freedom models when the sample size is small. Kenny, Kaniskan, & McCoach (2015) write (p502): “we suggest not computing the RMSEA for very low df models that do not have a large sample size”. However, it is not really clear nor explicitly stated in Kenny et al what constitutes “low df” and “not large sample size”. The authors evaluated the power of the RMSEA to reject that the RMSEA was RMSEA = 0.05 when the data-generating model has an RMSEA of 0.08 over a grid of sample sizes (50, 100, 200, 400, 600, 1000) and model degrees of freedom (1, 2, 3, 5, 10, 20, 50). They find adequate power (>80%) in the following conditions (N(df): 1000(10,20,50); 400(50); 600(50)).

Relative or incremental fit: the CFI

The CFI is a relative or incremental fit measure, meaning the computation is based on comparing the fit of the target model to a null model or baseline model. It is computed as $1 - d_0/d_1$ where $d = \chi^2 - df$ and the subscript 0 indicates a baseline model and subscript 1 indicates a target model.[3] The most commonly used null or baseline model is one in which all variables are assumed to be uncorrelated (Rigdon, 1996). The most recent rules-of-thumb call for interpreting good fit when CFI values are at least 0.95 (Hu & Bentler, 1999).

---

[3] Note that if RMSEA = sqrt(d1)/sqrt(df(N-1)), then c2-df = (df(N-1))*RMSEA^2. Therefore, CFI = 1-

Cautions and caveats in interpreting fit statistics

Deviations from normality in observed data can inflate χ^2 and worsen associated fit statistics (CFI, RMSEA).

Models with more variables may have lower (better) RMSEA and lower (worse) CFI (Kenny & McCoach, 2003). Therefore, improving CFI might warrant considering reducing the number of variables (e.g., via omitting variables parcels). This practice might detract from the validity of the model (Bandalos, 2008) [see endnote 1, below] and is not recommended as a general approach.

Larger sample sizes tend to be associated with models with larger RMSEA and SRMR values (Kenny, 2015). If the RMSEA is poor and the degrees of freedom is small, it may be that the model is not one that is appropriate for using the RMSEA (Kenny et al., 2015).

If the SRMR indicates good fit but the CFI and/or the RMSEA do not, it could be that

- extraneous parameters being estimated (e.g., regressions or factor loadings that are not important),
- the SRMR as a summary fit statistic is masking a small number of sample statistics that are estimated with high residuals, or
- the observed variable residual variances are low, such as when observables are measured with high reliability (Browne, MacCallum, Kim, Andersen, & Glaser, 2002).[4]

When fit statistics do not agree, we can consider the advice of Rigdon (1996) who argued that the absolute indices (specifically the RMSEA) are more important in confirmatory modeling and the incremental (specifically the CFI) fit indices more important in exploratory contexts. The reason is that the CFI relies upon the assumption that the null or baseline model is reasonable in the population, which is almost always false. The CFI may be appropriate in the context of novel research questions involving smaller sample sizes, while the RMSEA (and we may assume other absolute measures of fit) more appropriate in more confirmatory research questions involving larger sample sizes. Rigdon (1996) provides an alternative expression of the RMSEA --

$$\mathrm{RMSEA} = \sqrt{\frac{\hat{\mathrm{F}}}{\mathrm{df}} - \frac{1}{(\mathrm{N}-1)}}$$

Where $\hat{\mathrm{F}} = (\mathrm{s} - \hat{\sigma})'\mathrm{W}^{-1}(\mathrm{s} - \hat{\sigma})$, a weighted (standardized) sum of squared deviations of the fitted s versus observed σ covariance matrix. This expression clearly reveals how the RMSEA is sample size dependent. With small N, a well-fitting model (by χ^2) may not reach the RMSEA threshold for "good fit". If the available sample is very large, the RMSEA may indicate good fit when the model is not a particularly good model.

[4] This is a cool paper that should be updated with modern fit statistics, and perhaps generated/simulated data reflecting larger sample sizes.

As would seem to support Rigdon's advice to prioritize the CFI in exploratory research questions, Garrido and colleagues (2016) compared the CFI (and TLI), RMSEA, and a (Mplus v6.11) SRMR in simulation study concerning identifying the number of factors in an exploratory factor analysis inference setting. A clear advantage in terms of power for making the correct inference was seen for the CFI (and TLI), followed by RMSEA, and SRMR was the least accurate. Greiff and Heene (2017) report similar findings although they do not report the statistical software used.

Endnotes

Note 1: Parcels

Use of parcels can lead to improved fit. However, whether or not this is the "right" approach was investigated by Bandalos (2008). She reports that (p234-235):

> *[d]istributed parceling methods, in which items that share secondary sources of variance are parceled together with items that do not, have been advocated on the grounds that they will result in better model fit, as measured by commonly used fit indexes in SEM. Parceling schemes in which nonnormally and normally distributed items are included in the same parcel have also been recommended as a way of obtaining more normally distributed indicators for use in covariance structure analyses, and this technique has been found to produce better model fit in empirical studies (Landis et al., 2000). In this study a parceling method that combined these two approaches was applied to four different models, misspecifying the model by effectively collapsing two correlated factors (Models 1 and 2) or by masking the presence of a method factor (Models 3 and 4). Solutions based on this parceling strategy evidenced good fit, outperforming solutions based on other types of parceling. However, it is not clear whether this is a help or a hindrance to those whose interest is in obtaining an accurate picture of parameter estimates and model fit. Although use of this technique did result in excellent fit, it did so for models that were misspecified, resulting in a high rate of Type II errors. Perhaps more important, use of this parceling technique resulted in unacceptable levels of bias for the structural parameters in the model.*

and go on to say (p 238)

> *Researchers analyzing coarsely categorized or nonnormally distributed data are therefore advised to make use of [the Mplus WLSMV] estimator rather than resorting to parceling methods if the unidimensionality of the variables to be parceled cannot be assured.*

These results would seem to argue in favor of using a measurement model for multiple-indicator scales rather than multiple item parcels even when fit is sub-optimal.

References

Asparouhov, T., & Muthén, B. (2018, 2 May). SRMR in Mplus. Retrieved from http://www.statmodel.com/download/SRMR2.pdf

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling, 15*(2), 211-240.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*: Routledge.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological methods, 7*(4), 403.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological methods, 21*(1), 93.

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European journal of psychological assessment, 33*(5), 313-317. doi:10.1027/1015-5759/a000450

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Kenny, D. A. (2015, 24 Nov). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological methods & research, 44*(3), 486-507.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333-351.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(4), 369-379.