David A. Kenny
January 11, 2014
Please send me your suggestions or corrections.

Measuring Model Fit

PLEASE DO NOT EMAIL ME FOR CITATIONS FOR STATMENTS ON THIS PAGE! I do provide some citations for claims made, but if you need more please search the literature yourself or just cite this page. Thank you.

Fit refers to the ability of a model to reproduce the data (i.e., usually the variance-covariance matrix). A good-fitting model is one that is reasonably consistent with the data and so does not necessarily require respecification. Also a good-fitting measurement model is required before interpreting the causal paths of the structural model.

It should be noted that a good-fitting model is not necessarily a valid model. For instance, a model all of whose parameters are zero is a "good-fitting" model. Additionally, models with nonsensical results (e.g., paths that are clearly the wrong sign) and models with poor discriminant validity or Heywood cases can be “good-fitting” models. Parameter estimates must be carefully examined to determine if one has a reasonable model as well as the fit statistics. Also it is important to realize that one might obtain a good-fitting model, yet it is still possible to improve the model and remove specification error. Finally, having a good-fitting model does not prove that the model is correctly specified. Conversely, it should be noted that a model all of whose parameters are statistically significant can be from a poor fitting model.

How Large a Sample Size Do I Need?

Rules of Thumb

- Ratio of Sample Size to the Number of Free Parameters
  - Tanaka (1987): 20 to 1, but that is unrealistically high.
  - Goal: Bentler & Chou (1987): 5 to 1
  - Several published studies do not meet this goal.
- Sample Size

200 is seen as a goal for SEM research

Lower sample sizes can be used for

Models with no latent variables

Models where all loadings are fixed (usually to one)

Models with strong correlations

Simpler models

Models for which there is an upper limit on N (e.g., countries or years as the unit), 200 might be an unrealistic standard.

Power Analysis

Best way to determine if you have a large enough sample is to conduct a power analysis.

Either use the Sattora and Saris (1985) method or conduct a simulation.

To test your power to detect a poor fitting model, you can use Preacher and Coffman's web calculator.

The Chi Square Test: χ^2

For models with about 75 to 200 cases, the chi square test is a reasonable measure of fit. But for models with more cases (400 or more), the chi square is almost always statistically significant. Chi square is also affected by the size of the correlations in the model: the larger the correlations, the poorer the fit. For these reasons alternative measures of fit have been developed. (Go to a website for computing *p* values for a given chi square value and df.)

Sometimes chi square is more interpretable if it is transformed into a *Z* value. The following approximation can be used:

$$Z = \sqrt{(2\chi^2)} - \sqrt{(2df - 1)}$$

An old measure of fit is the chi square to df ratio or χ^2/df. A problem with this fit index is that there is no universally agreed upon standard as to what is a good and a bad fitting model. Note, however, that two very popular fit indices, TLI and RMSEA, are largely based on this old-fashioned ratio.

The chi square test is too liberal (i.e., too many Type 1) errors when variables have non-normal distributions, especially distributions with kurtosis. Moreover, with small sample sizes, there are too many Type 1 errors.

Introduction to Fit Indices

The terms in the literature used to describe fit indices are confusing, and I think confused. I prefer the following terms (but they are unconventional): incremental, absolute, and comparative which are used on the pages that follow.

Incremental Fit Index

An incremental (sometimes called *relative*) fit index is analogous to R^2 and so a value of zero indicates having the worst possible model and a value of one indicates having the best possible. So my model is placed on a continuum. In terms of a formula, it is

$$\frac{\text{Worst Possible Model – My Model}}{\text{Worst Possible Model – Fit of the Best Possible Model}}$$

The worst possible model is called the *null* or *independence model* and the usual convention is to allow all the variables in the model to have variation but no correlation. (The usual null model is to allow the means to equal their actual value. However, for growth curve models, the null model should set the means as equal, i.e., no growth.) The degrees of freedom of the null model are k(k – 1)/2 where k is the number of variables in the model. Amos refers to the null model as the *independence model*. Note that a different null model needs to be fitted if the means are part of the model (e.g., a growth curve model).

Alternative null models might be considered (but almost never done). One alternative null model is that all latent variable correlations are zero and another is that all exogenous variables are correlated but the endogenous variables are uncorrelated with each other and the exogenous variables. O'Boyle and Williams (2011) suggest two different null models for the measurement and structural models.

Absolute Fit Index

An absolute measure of fit presumes that the best fitting model has a fit of zero. The measure of fit then determines how far the model is from perfect fit. These measures of fit are typically "badness" measure of fit in that bigger is worse.

Comparative Fit Index

A comparative measure of fit is only interpretable when comparing two different models. This term is unique to this website in that these measures are more commonly called absolute fit indices. However, it is helpful to distinguish absolute indices that do not require a comparison between two models.

Controversy about Fit Indices

Recently considerable controversy has flared up concerning fit indices. Some researchers do not believe that fit indices add anything to the analysis (e.g., Barrett, 2007) and only the chi square should be interpreted. The worry is that fit indices allow researchers to claim that a miss-specified model is not a bad model. Others (e.g., Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007) argue that cutoffs for a fit index can be misleading and subject to misuse. Most analysts believe in the value of fit indices, but caution against strict reliance on cutoffs.

Also problematic is the “cherry picking” a fit index. That is, you compute many fit indices and you pick the one index that allows you to make the point that you want to make. If you decide not to report a popular index (e.g., the TLI or the RMSEA), you need to give a good reason why you are not.

Finally, Kenny, Kaniskan, and McCoach (2013) have argued that fit indices should not even be computed for small degrees of freedom models. Rather for these models, the researcher should locate the source of specification error.

Catalogue of Fit Indices

There are now literally hundreds of measures of fit. This page includes some of the major ones currently used in the literature, but does not pretend to include all the measures. Though a bit dated, the book edited by Bollen and Long (1993) explains these indexes and others. Also a special issue of the *Personality and Individual Differences* in 2007 is entirely devoted to the topic of fit indices.

A key consideration in choice of a fit index is the penalty it places for complexity. That penalty for complexity is measured by how much chi square needs to change for the fit index not to change.

Bentler-Bonett Index or Normed Fit Index (NFI)

This is the very first measure of fit proposed in the literature (Bentler & Bonett, 1980) and it is an incremental measure of fit. The best model is defined as model with a χ^2 of zero and the worst model by the χ^2 of the null model. Its formula is:

$$\frac{\chi^2(\text{Null Model}) - \chi^2(\text{Proposed Model})}{\chi^2(\text{Null Model})}$$

A value between .90 and .95 is now considered marginal, above .95 is good, and below .90 is considered to be a poor fitting model. A major disadvantage of this measure is that it cannot be smaller if more parameters are added to the model. Its "penalty" for complexity is zero. Thus, the more parameters added to the model, the larger the index. It is for this reason that this measure is not recommended, but rather one of the next two is used.

Tucker Lewis Index or Non-normed Fit Index (NNFI)

A problem with the Bentler-Bonett index is that there is no penalty for adding parameters. The Tucker-Lewis index, another incremental fit index, does have such a penalty. Let χ^2/df be the ratio of chi square to its degrees of freedom, and the TLI is computed as follows:

$$\frac{\chi^2/\text{df(Null Model)} - \chi^2/\text{df(Proposed Model)}}{\chi^2/\text{df(Null Model)} - 1}$$

If the index is greater than one, it is set at one. It is interpreted as the Bentler-Bonett index. Note that for a given model, a lower chi square to *df* ratio (as long as it is not less than one) implies a better fitting model. Its penalty for complexity is χ^2/df. That is, if the chi square to df ratio does not change, the TLI does not change.

Note that the TLI (and the CFI which follows) depends on the average size of the correlations in the data. If the average correlation between variables is not high, then the TLI will not be very high. Consider a simple example. You have a 5-item scale that you think measures one latent variable. You also have 3 dichotomous experimental variables that you manipulate that cause those two latent factors. These three experimental variables create 7 variables when you allow for all possible interactions. You have equal N in the conditions, and so all their correlations are zero. If you run the CFA on just the 5 indicators, you might have a nice TLI of .95. However, if you add in the 7 experimental variables, your TLI might sink below .90 because now the null model will not be so "bad" because you now have added to the model 7 variables who have zero correlations with each other.

A reasonable rule of thumb is to examine the RMSEA for the null model and make sure that is no smaller than 0.158. An RMSEA for the model of 0.05 and a TLI of .90, implies that the RMSEA of the null model is 0.158. If the RMSEA for the null model is less than 0.158, an incremental measure of fit may not be that informative. So far as I know, this mathematical fact that a model whose null model RMSEA is less than 0.158 and whose RMSEA is 0.05 must have a TLI of less than .90 is something that has never been published but is in fact true.

Comparative Fit Index (CFI)

This incremental measure of is directly based on the non-centrality measure. Let d = χ^2 - *df* where *df* are the degrees of freedom of the model. The Comparative Fit Index or CFI equals

$$\frac{\text{d(Null Model) - d(Proposed Model)}}{\text{d(Null Model)}}$$

If the index is greater than one, it is set at one and if less than zero, it is set to zero. It is interpreted as the previous incremental indexes.

If the CFI is less than one, then the CFI is always greater than the TLI. CFI pays a penalty of one for every parameter estimated. Because the TLI and CFI are highly correlated only one of the two should be reported. The CFI is reported more often than the TLI, but I think the CFI’s penalty for complexity of just 1 is too low and so I prefer the TLI even though the CFI is reported much more frequently than the TLI.

Again the CFI should not be computed if the RMSEA of the null model is less than 0.158 or otherwise one will obtain too small a value of the CFI.

Root Mean Square Error of Approximation (RMSEA)

This absolute measure of fit is based on the non-centrality parameter. Its computational formula is:

$$\frac{\sqrt{(\chi^2 - df)}}{\sqrt{[df(N - 1)]}}$$

where N the sample size and *df* the degrees of freedom of the model. If χ^2 is less than df, then the RMSEA is set to zero. Like the TLI, its penalty for complexity is the chi square to df ratio. The measure is positively biased (i.e., tends to be too large) and the amount of the bias depends on smallness of sample size and df, primarily the latter. The RMSEA is currently the most popular measure of model fit and it now reported in virtually all papers that use CFA or SEM and some refer to the measure as the "Ramsey."

MacCallum, Browne and Sugawara (1996) have used 0.01, 0.05, and 0.08 to indicate excellent, good, and mediocre fit, respectively. However, others have suggested 0.10 as the cutoff for poor fitting models. These are definitions for the population. That is, a given model may have a population value of 0.05 (which would not be known), but in the sample it might be greater than 0.10. Use of confidence intervals and tests of PCLOSE can help understand the sampling error in the RMSEA. There is greater sampling error for small df and low N models, especially for the former. Thus, models with small df and low N can have artificially large values of the RMSEA. For instance, a chi square of 2.098 (a value not statistically significant), with a df of 1 and N of 70 yields an RMSEA of 0.126. For this reason, Kenny, Kaniskan, and McCoach (2013) argue to not even compute the RMSEA for low df models.

A confidence interval can be computed for the RMSEA. Ideally the lower value of the 90% confidence interval includes or is very near zero (or no worse than 0.05) and the upper value is not very large, i.e., less than .08. The width of the confidence interval is very informative about the precision in the estimate of the RMSEA.

p of Close Fit (PCLOSE)

This measure provides is one-sided test of the null hypothesis is that the RMSEA equals .05, what is called a *close-fitting model.* Such a model has specification error, but "not very much" specification error. The alternative, one-sided hypothesis is that the RMSEA is greater than 0.05. So if the *p* is greater than .05 (i.e., not statistically significant), then it is concluded that the fit of the model is "close." If the p is less than .05, it is concluded that the model's fit is worse than close fitting (i.e., the RMSEA is greater than 0.05). As with any significance test, sample size is a critical factor, but so also is the model df, with lower df there is less power in this test.

You can use a Preacher and Coffman webpage to test any null hypothesis about the RMSEA. Note that the standard chi square test takes as the null hypothesis that the RMSEA equals zero.

Standardized Root Mean Square Residual (SRMR)

The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. It is a positively biased measure and that bias is greater for small N and for low df studies. Because the SRMR is an absolute measure of fit, a value of zero indicates perfect fit. The SRMR has no penalty for model complexity. A value less than .08 is generally considered a good fit (Hu & Bentler, 1999).

Akaike Information Criterion (AIC)

The AIC is a comparative measure of fit and so it is meaningful only when two different models are estimated. Lower values indicate a better fit and so the model with the lowest AIC is the best fitting model. There are somewhat different formulas given for the AIC in the literature, but those differences are not really meaningful as it is the difference in AIC that really matters:

$$\chi^2 + k(k + 1) - 2df$$

where *k* is the number of variables in the model and *df* is the degrees of freedom of the model. (If means are included in the model, then replace k(k + 1) with k(k + 3).) Note that k(k + 1) - 2df equals the twice the number of free parameters in the model. The AIC makes the researcher pay a penalty of two for every parameter that is estimated.

One advantage of the AIC, BIC, and SABIC measures is that they can be computed for models with zero degrees of freedom, i.e., just-identified models.

Bayesian Information Criterion (BIC)

Two other comparative fit indices are the BIC and the SABIC. Whereas the AIC has a penalty of 2 for every parameter estimated, the BIC increases the penalty as sample size increases

$$\chi^2 + \ln(N)[k(k + 1)/2 - df]$$

where ln(N) is the natural logarithm of the number of cases in the sample. (If means are included in the model, then replace k(k + 1)/2 with k(k + 3)/2.) The BIC places a high value on parsimony (perhaps too high).

The Sample-Size Adjusted BIC (SABIC)

Like the BIC, the Sample-size adjusted BIC or SABIC places a penalty for adding parameters based on sample size, but not as high a penalty as the BIC. The SABIC is not given in Amos, but is given in Mplus. Several recent simulation studies (Enders & Tofighi, 2008; Tofighi, & Enders, 2007) have suggested that the SABIC is a useful tool in comparing models. Its formula is

$$\chi^2 + \ln[(N + 2)/24][k(k + 1)/2 - df]$$

Again if means are included in the model, then replace k(k + 1)/2 with k(k + 3)/2.

GFI and AGFI (LISREL measures)

These measures are affected by sample size. The current consensus is not to use these measures (Sharma, Mukherjee, Kumar, & Dillon, 2005).

Hoelter Index

The index states the sample size at which chi square would not be significant (alpha = .05), i.e., that is how small one's sample size would have to be for the result to be no longer significant. The index should only be computed if the chi square is statistically significant. Its formula is:

$$[(N - 1)\chi^2(crit)/\chi^2] + 1$$

where *N* is the sample size, χ^2 is the chi square for the model and χ^2(crit) is the critical value for the chi square. If the critical value is unknown, the following approximation can be used:

$$\frac{[1.645 + \sqrt{(2df - 1)}]^2 + 1}{2\chi^2/(N - 1) + 1}$$

where df are the degrees of freedom of the model. For both of these formulas, one rounds *down* to the nearest integer value. Hoelter recommends values of at least 200. Values of less than 75 indicate very poor model fit.

The Hoelter only makes sense to interpret if N > 200 and the chi square is statistically significant. It should be noted that Hu and Bentler (1998) do not recommend this measure.

Factors that Affect Fit Indices

Number of Variables

Anecdotal evidence that models with many variables do not fit
Kenny and McCoach (2003) show
RMSEA improves as more variables are added to the model
TLI and CFI are relatively stable, but tend to decline slightly
We do not understand why it is that models with more variables tend to have relatively poor fit.

Model Complexity

How much chi square needs to change per df for the fit index not to change:

	Theoretical Value	A&M*	Reis**
Bentler and Bonett	0	0	0
CFI	1	1	1
AIC	2	2	2
SABIC	ln[(N +2)/24]	1.96	1.76
Tucker Lewis	χ^2/df	2.01	1.46
RMSEA	χ^2/df	2.01	1.46
BIC	ln(N)	5.13	4.93

*A&M: Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology*, *22*, 453-474.
**Reis: Reisenzein, R. (1986). A structural equation analysis of Weiner's attribution-affect model of helping behavior. *Journal of Personality and Social Psychology, 50*, 1123-1133.

Note that
Each changes by a constant amount, regardless of the df change.

Larger values reward parsimony and smaller values reward complexity. The BIC rewards parsimony most, and the CFI (after the Bentler and Bonett), the least. The SABIC is comparable to the RMSEA and the TLI.

Sample Size

Bentler-Bonett fails to adjust for sample size: models with larger sample sizes have smaller values. The TLI and CFI do not vary much with sample size. However, these measures are less variable with larger sample sizes.

The RMSEA and the SRMR are larger with smaller sample sizes.

Normality

Non-normal data (especially high kurtosis) inflates chi square and absolute measures of fit. Presumably, incremental and comparative measures of fit are less affected.

References

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Bentler, P. M., & Chou, C. P. (1987) Practical issues in structural modeling. *Sociological Methods & Research, 16*, 78-117.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-600.

Bollen, K. A., & Long, J. S., Eds. (1993). *Testing structural equation models*. Newbury Park, CA: Sage

Enders, C.K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 75-95.

Hayduk, L., Cummings, G. G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two three – Testing the theory in structural equation models! *Personality and Individual Differences, 42*, 841-50.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

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

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2013). *The performance of RMSEA in models with small degrees of freedom*. Unpublished paper, University of Connecticut.

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*, 333-3511.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149.

O'Boyle, E. H., Jr., Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology, 96*, 1-12.

Satorra, A., & Saris,W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W.R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*, 935-43.

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

Tofghi, D., & Enders, C. K. (2007). Identifying the correct number of classes in mixture models. In G. R. Hancock & K. M. Samulelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Greenwich, CT: Information Age.

Go to the next SEM page.

Go to the main SEM page.