Some Clarifications and Recommendations on Fit Indices

Tanaka (1993), Maruyama (1998), and others distinguish between several types of fit indices: *absolute fit indices*, *relative fit indices*, *parsimony fit indices*, and those based on the *noncentrality* parameter.

Absolute Fit Indices (χ^2, GFI, AGFI, Hoelter's CN, AIC, BIC, ECVI, RMR, SRMR)

Absolute fit indices do not use an alternative model as a base for comparison. They are simply derived from the fit of the obtained and implied covariance matrices and the ML minimization function. Chi-square (χ^2, sometimes referred to as T) is the original fit index for structural models because it is derived directly from the fit function [f_{ML}(N-1)]. Because chi-square is the original fit index and because it is the basis for most other fit indices, it is routinely reported in all SEM results sections.

In practice, however, chi-square is not considered to be a very useful fit index by most researchers, because it is affected by the following factors (1) sample size—larger samples produce larger chi-squares that are significant even with very small discrepancies between implied and obtained covariance matrices. On the other hand, small samples may be too likely to accept poor models (Type II error). Based on my experience, it is difficult to get a nonsignificant chi-square when samples sizes are much over 200 or so. (2) model size also has an increasing effect on chi-square values. Models with more variables tend to have larger chi-squares. (3) Chi-square is affected by the distribution of variables. Highly skewed and kurtotic variables increase chi-square values. This has to do with the multivariate normality assumption that we will discuss later in the class. (4) There may be some lack of fit because of omitted variables. Omission of variables may make it difficult to reproduce the correlation (or covariance) matrix perfectly.

There are several other indices that fall into the category of absolute indices, including the Goodness-of-fit index (GFI, also known as gamma-hat or $\hat{\gamma}$), the adjusted goodness of fit index (AGFI), χ^2/df ratio, Hoelter's CN ("critical N"), Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Expected Cross-validation Index (ECVI), the root mean square residual (RMR), and the standardized root mean square residual (SRMR). Most of these indices, with the possible exception of the SRMR, have similar problems to those of the chi-square, because they are simple transformations of chi-square. As one example, the AIC (as given by Tanaka, 1993) is $\chi^2 + 2(p)$, where p is the number of free parameters (the number counted in calculating df).

Relative Fit Indices (IFI, TLI, NFI)
Relative fit indices compare a chi-square for the model tested to one from a so-called *null model* (also called a "baseline" model or "independence" model). The null model is a model tested that specifies that all measured variables are uncorrelated (there are no latent variables). The null model should always have a very large chi-square (poor fit). Although

other baseline models could be used, this is not often seen in practice.[1] There are several relative fit indices (which are not explicitly designed to be provide penalties for less parsimonious models), including Bollen's Incremental Fit Index (IFI, also called BL89 or Δ_2), the Tucker-Lewis Index [TLI, Bentler-Bonett Nonnormed Fit Index (NNFI or BBNFI), or ρ_2], and the Bentler-Bonett Normed Fit Index (NFI). Most of these fit indices are computed by using ratios of the model chi-square and the null model chi-square and dfs for the models. All of them have values that range between approximately 0 and 1.0. Some of these indices are "normed" so that their values cannot be below 0 or above 1 (e.g., NFI, CFI described below). Others are considered "nonnormed" because, on occasion, they may be larger than 1 or slightly below 0 (e.g., TLI, IFI). In the past, these indexes have generally been used with a conventional cutoff in which values larger than .90 are considered good fitting models, but there seems to be growing consensus that this value should be increased to approximately .95 (based largely on Hu & Bentler, 1999).

Parsimonious Fit Indices (PGFI, PNFI, PNFI2, PCFI)
These fit indices are relative fit indices that are adjustments to most of the ones above. The adjustments are to penalize models that are less parsimonious, so that simpler theoretical processes are favored over more complex ones. The more complex the model, the lower the fit index. Parsimonious fit indices include PGFI (based on the GFI), PNFI (based on the NFI), PNFI2 (based on Bollen's IFI), PCFI (based on the CFI mentioned below). Mulaik et al. (1989) developed a number of these. Although many researchers believe that parsimony adjustments are important, there is some debate about whether or not they are appropriate. My own perspective is that researchers should evaluate model fit independent of parsimony considerations, but evaluate alternative theories favoring parsimony. With that approach, we would not penalize models for having more parameters, but if simpler alternative models seem to be as good, we might want to favor the simpler model.

Noncentrality-based Indices (RMSEA, CFI, RNI, CI)
The concept of the *noncentrality parameter* is a somewhat difficult one. The rationale for the noncentrality parameter is that our usual chi-square fit is based on a test that the null hypothesis is true ($X^2 = 0$). This gives a distribution of the "central" chi-square. Because we are hoping *not* to reject the null hypothesis in structural modeling, it can be argued that we should be testing to reject the alternative hypothesis (H_a). A test that rejected the alternative hypothesis, H_a, would make statistical decisions using the "noncentral" chi-square distribution created under the case when H_a is assumed to be true in the population (i.e., an incorrect model in the population). This approach to model fit uses a chi-square equal to the df for the model as having a perfect fit (as opposed to chi-square equal to 0). Thus, the noncentrality parameter estimate is calculated by subtracting the df of the model from the chi-square ($\chi^2 - df$). Usually this value is adjusted for sample size and referred to as the rescaled noncentrality parameter:

$$d = \frac{\chi^2 - df}{N - 1}$$

---

[1]   The uncorrelated null model is not fully universal. In fact, Mplus has introduced an alternative null model under some circumstances (correlations among non-latent exogenous variables are included). However, past application and statistical work has been based on the uncorrelated null model approximately 99% of the time.

A population version is often referred to as δ and is computed by dividing by N rather than N-1. Noncentrality-based indices include the Root Mean Square Error of Approximation (RMSEA)—not to be confused with RMR or SRMR, Bentler's Comparative Fit Index (CFI), McDonald and Marsh's Relative Noncentrality Index (RNI), and McDonald's Centrality Index (CI). Because the noncentrality parameter is simply a function of chi-square, df, and N, several of the formulas for the relative fit indices described above can be algebraically manipulated to include the noncentrality parameter. For example the TLI can also be presented as:

$$TLI = \frac{(d_0 / df_0) - (d_{model} / df_{model})}{d_0 / df_0}$$

Where d_{model} and df_{model} are the noncentrality parameter and the degrees of freedom for the model tested and d_0 and df_0 are the noncentrality parameter and df for the null model. Recent work by Raykov (2000, 2005) shows that noncentrality parameter sample estimates are biased and that this problem may affect fit indices computed based on noncentrality.

Sample Size Independence

Many of the relative fit indices (and the noncentrality fit indices) are affected by sample size, so that larger samples are seen as better fitting (i.e., have a higher fit index value). Bollen (1990) made a very useful distinction between fit indices that can be shown to explicitly include N in their calculation and those that are dependent on sample size empirically. That is, even though a fit index may not include N in the formula, or even attempt to adjust for it, does not mean that the fit index will really turn out to be independent of sample size. He also showed that the TLI and IFI are relatively unaffected by sample size (see also Anderson & Gerbing, 1993; Hu & Bentler, 1995; Marsh, Balla, & McDonald, 1988).

$$TLI = \frac{\chi^2_{null} / df_{null} - \chi^2_{model} / df_{model}}{\chi^2_{null} / df_{null} - 1}$$

$$IFI = \frac{\chi^2_{null} - \chi^2_{model}}{\chi^2_{null} - df_{model}}$$

This is one reason why I tend to favor Bollen's IFI. If you are interested in adjusting for parsimony, you might consider the Mulaik et al.'s PNFI2 which is a parsimony adjusted version of the IFI. One can make an argument about parsimony adjustment similar to Bollen's argument about sample size. It might be important to differentiate between fit indices that are explicitly adjusting for parsimony and ones that are empirically affected by model complexity. The TLI is a example of an index that adjusts for parsimony, even though that was not its original intent.

Recommendations

Every researcher and every statistician seems to have a favorite index or set of indices. You should be prepared for reviewers to suggest the addition of one or two of their favorite indices, but it would not be fair to yourself or others to pick the index that is most optimistic about the fit of your model. In recent years, there has been concern that the recommended cutoff values for relative fit indices of .90 are too low and that higher values, such as .95

should be used. Hu and Bentler (1999) empirically examine various cutoffs for many of these measures, and their data suggest that to minimize Type I and Type II errors under various conditions, one should use a combination of one of the above relative fit indexes and the SRMR (good models < .08) or the RMSEA (good models < .06). These values should not be written in stone (in fact there have been some recent concerns raised; e.g., Fan & Sivo, 2005; Marsh et al., 2004) but I believe this is useful work and hope it will helpful for establishing a more concrete basis for conventional cutoff values in the future. Based on the IFI's independence of sample size and the data from Hu and Bentler, I usually prefer to report the IFI in combination with the SRMR in my work. Most importantly, researchers should decide a priori about fit criteria, state those criteria in their reports, and consider reporting more than one fit index (Jackson, Gillaspy, & Purc-Stephenson, 2009).

References

Bollen, 1990, Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*, 256-259.

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

Gerbing, D.W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen, & J.S. Long (eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural Equation Modeling. Concepts, Issues, and Applications* (pp. 76-99). London: Sage.

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.

Jackson, D.L., Gillaspy, J.A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. Psychological Methods, 14, 6–23.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

Raykov, T. (2000). On the large-sample bias, variance, and mean squared error of the conventional noncentrality parameter estimator of covariance structure models. *Structural Equation Modeling, 7*, 431-441.

Raykov, T. (2005). Bias-corrected estimation of noncentrality parameters of covariance structure models. *Structural Equation Modeling, 12*, 120-129.

Steiger, J.H. (1989). EZPATH: *A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen, & J.S. Long (eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Suggested Further Readings

Bollen, 1990, Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*, 256-259.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Gerbing, D.W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen, & J.S. Long (eds.), *Testing structural equation models*. Newbury Park, CA: Sage.

Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural Equation Modeling. Concepts, Issues, and Applications* (pp.76-99). London: Sage.

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

Sivo, S.A, Fan, X., Witta, E.L., & Willse, J.T. (2006). The Search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*, 267–288