CHAPTER 13

# Model Fit and Model Selection in Structural Equation Modeling

**Stephen G. West**
**Aaron B. Taylor**
**Wei Wu**

One of the strengths of structural equation modeling (SEM) is the ability to test models that represent a complex set of theoretical hypotheses. The set of hypothesized relationships is specified and commonly represented graphically in the compact form of a path diagram. The model and its associated path diagram contain one or more of three components. It may contain a hypothesized measurement component that relates the observed (measured) variables to underlying constructs (Figure 13.1A). It may contain a structural (path) component that portrays the hypothesized causal relationships between the constructs (Figure 13.1B). It may contain a hypothesized mean component that portrays similarities and differences in the level of the constructs, potentially as a function of other variables (Figure 13.1C). Once a path model is specified, an important question arises: How well does the hypothesized model fit observed data on each of the variables?

The path model diagram implies a set of algebraic equations whose parameters (e.g., factor loadings in $\Lambda_y$, factor variances and covariances in $\Psi$) are estimated, typically through maximum likelihood (ML) or generalized least squares (GLS) estimation procedures. For the confirmatory factor analysis (CFA) model in Figure 13.1A,

$$\Sigma = \Lambda_y \Psi \Lambda_y' + \Theta_\varepsilon \qquad (13.1)$$

where $\Sigma$ is the population covariance matrix of the observed variables, $\Lambda_y$ is the matrix of factor loadings, $\Psi$ is the matrix of factor covariances, and $\Theta_\varepsilon$ is the covariance matrix of residuals. The parameters estimated for the specified model, in turn, provide the machinery for calculating what the variances, covariances, and means of the variables would be, *if in fact the model were true* (model-implied estimates). The key question for assessing the overall fit of the model is how well the estimates implied by the model match the variances, covariances, and means of the observed data.

This chapter addresses two related but different questions. First, we may wish to answer the question of model fit: Does the hypothesized model provide an adequate fit to the data? Second, we may wish to answer the question of model selection: If multiple competing models have been proposed, which of these models provides the best account of the data? Or, alternatively, which competing model is most likely to replicate in another sample drawn from the same population? We focus on the model fit question in the initial part of the chapter, returning to brief consideration of the model selection question at the end of the chapter. We also briefly consider other key aspects of model evaluation beyond those of overall model fit.

We begin by reviewing the properties of the chi square ($\chi^2$) test statistic and several "practical" indices

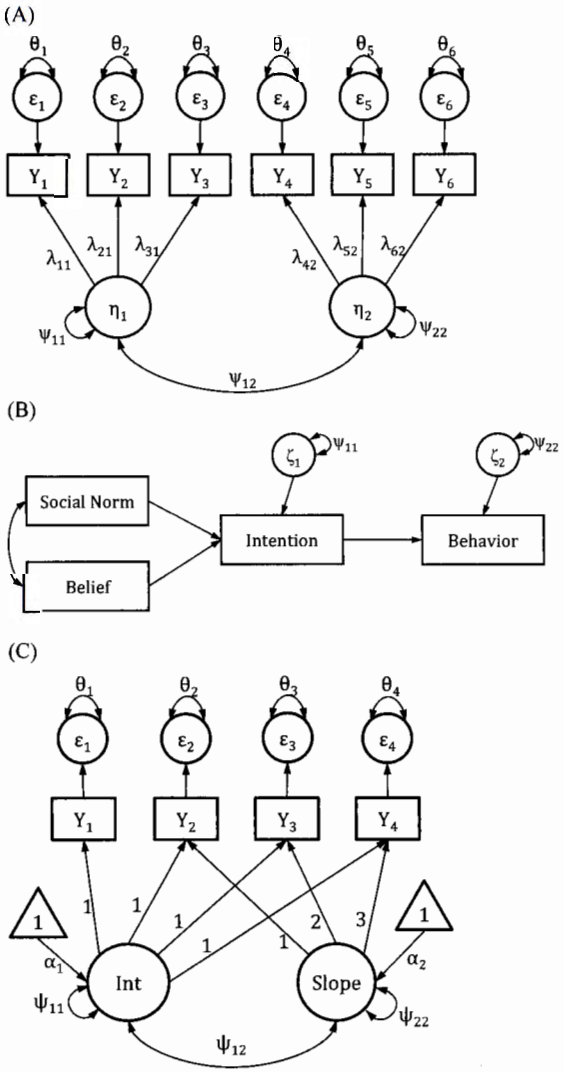


**FIGURE 13.1.** (A) Two-factor confirmatory factor analysis model. (B) Path model with four measured variables (Fishbein–Azjen model). (C) Linear growth model with four time points.

of overall model fit, focusing on those that are currently being reported in journals by researchers. In the first part of our review we emphasize lack of sensitivity to sample size in estimation, the criterion that dominated the evaluation of fit indices in the last part of the 20th century. We then consider other desiderata for good fit indices, discovering that other model-related factors can make it difficult to establish a threshold for good fit. Most existing work has only studied the performance of fit indices in simple CFA (measurement) models; we initially follow this precedent but later consider the use of fit indices with other, more complex models, such as growth models and multilevel models. We consider evaluating the fit of different model components as well as overall global fit. We also consider other approaches to evaluating the adequacy of a model. Finally, we consider model selection indices useful for selecting the best of a set of competing models.

## ASSESSING OVERALL MODEL FIT: THE CHI-SQUARE TEST AND PRACTICAL FIT INDICES

Most of the practical fit indices involve the chi-square ($\chi^2$) test statistic for the hypothesized model, sometimes in conjunction with same test statistic for a baseline model (Yuan, 2005). A summary of some of the equations, original sources, and key properties of several commonly used fit indices is presented in Table 13.1.

For covariance structure models, we use the following notation. The number of observed variables being modeled is denoted $p$, and their covariance matrix, based on a sample size of $N$, is $\mathbf{S}$. The corresponding population covariance matrix is $\mathbf{\Sigma}$. The covariance matrix reproduced by the model using $q$ estimated parameters is $\hat{\mathbf{\Sigma}}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a vector of free parameters estimated by the model (factor loadings in $\mathbf{\Lambda}$; factor variances and covariances in $\mathbf{\Psi}$; unique variances and covariances in $\mathbf{\Theta}$). Each of the covariance matrices ($\mathbf{S}, \mathbf{\Sigma}, \hat{\mathbf{\Sigma}}(\boldsymbol{\theta})$), has $p^*$ nonredundant elements, where $p^* = p(p + 1)/2$. The model estimation procedure attempts to minimize a discrepancy function $F$, which achieves a minimum value $f$. A general form of the discrepancy function is presented in Equation 13.2 (Browne, 1974):

$$F = (\mathbf{s} - \hat{\boldsymbol{\sigma}}(\boldsymbol{\theta}))' \mathbf{W}^{-1} (\mathbf{s} - \hat{\boldsymbol{\sigma}}(\boldsymbol{\theta})) \qquad \textbf{(13.2)}$$

where $\mathbf{s}$ is a vector containing the $p^*$ nonredundant elements in the sample covariance matrix, $\hat{\boldsymbol{\sigma}}(\boldsymbol{\theta})$ is a vector

containing the $p^*$ nonredundant elements in the model implied covariance matrix, and $\mathbf{W}$ is a weight matrix. Equation 13.3 presents the most commonly used discrepancy function for the ML estimation procedure (Jöreskog, 1967):

$$\hat{F} = \log|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})| + \text{tr}|\mathbf{S}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1}| - \log|\mathbf{S}| - p \quad \textbf{(13.3)}$$

where tr is the trace of the matrix.

## Chi-Square (Likelihood Ratio) Test

For standard ML estimation (Equation 13.3), under the null hypothesis that the model-implied covariance matrix equals $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, the population covariance matrix $\boldsymbol{\Sigma}$, the test statistic $T = (N - 1)f$ follows a central $\chi^2$ distribution with degrees of freedom ($df$) equal to $p^* - q$. $f$ is the minimum of $\hat{F}$. $q$ is the number of parameters to be estimated. Important assumptions underlying this test statistic are that (1) the observed variables have a multivariate normal distribution, (2) $N$ is sufficiently large, and (3) none of the tested parameters is at a boundary (e.g., variance = 0). We refer to this expression as the $\chi^2$ test (although other such tests are possible; Hu & Bentler, 1995). If the observed $\chi^2$ exceeds the critical value given the $df$ and the nominal Type I error rate (typically $\alpha = .05$), the null hypothesis that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}$ is rejected. This means that the null hypothesis of perfect fit in the population is false, the assumptions are wrong, or both. As we discuss below, this $\chi^2$ test has limitations and is not always the final word in assessing fit.

This $\chi^2$ test can be considered a special case of the likelihood ratio (LR) test for nested models. A model is nested within another if its estimated parameters are a subset of the estimated parameters in the other model (see Bentler & Satorra, 2010). Typically, this occurs when a parameter is set equal to a fixed value (e.g., $\psi_{21} = 1$) or two or more parameters are set equal (e.g., $\lambda_{11} = \lambda_{42}$, setting the factor loadings of indicators 1 and 4 on their respective factors equal; see Figure 13.1A). The null hypothesis is that the model estimating fewer parameters (Fewer) fits no worse in the population than the model estimating more parameters (More). The LR test statistic is presented in Equation 13.4:

$$\Delta\chi^2 = \chi^2_{\text{Fewer}} - \chi^2_{\text{More}} \quad \Delta df = df_{\text{Fewer}} - df_{\text{More}} \quad \textbf{(13.4)}$$

Given that previous assumptions (1), (2), and (3) are met and that the two tested models are not too discrepant from the true model in the population (Steiger, Shapiro, & Browne, 1985), $\Delta\chi^2$, the difference between the two tested models' $\chi^2$ values, follows a $\chi^2$ distribution under the null hypothesis, with $df$ equal to $\Delta df$ (Bentler & Bonett, 1980). The $\chi^2$ test of overall model fit tests the null hypothesis that the tested model fits no worse than a saturated model, which estimates $p^*$ parameters and fits the data perfectly. The saturated model has a $\chi^2$ value of 0 with $df = 0$. A saturated model exists for all covariance structure models; however, some more complex models do not have a known saturated model or the standard saturated model is incorrect.

Jöreskog (1969), who introduced the $\chi^2$ test of fit in the context of covariance structure models, also noted its limitations (see also Bentler & Bonett, 1980; James, Mulaik, & Brett, 1982; Tucker & Lewis, 1973). A major problem with the $\chi^2$ test is that as $N$ increases, its power to detect even trivial differences between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{S}$ approaches 1.0. A model that accounts for the major sources of covariance in the data, even if it ignores what Jöreskog termed "minor factors," can still be of practical value—"all models are wrong, some are useful" (Box, 1979, p. 202). Models may be considered to be approximations of reality a priori, so the null hypothesis of exact fit is *not* expected to be retained (Cudeck & Henly, 1991; Jöreskog & Sörbom, 1981; MacCallum, Widaman, Preacher, & Hong, 2001; Steiger & Lind, 1980). In short, the null hypothesis of exact overall fit tested by the $\chi^2$ test is often not of general interest.

Other problems with the $\chi^2$ test have also been raised. Because researchers hope to retain the null hypothesis (thus supporting the theoretically hypothesized model), the use of the $\chi^2$ test statistic encourages the use of small samples (Bentler & Bonett, 1980; Meehl, 1967). Small samples, in turn, potentially obscure poor fit and yield less precise estimates of the free (estimated) parameters in a model. The test statistic $T$ is not likely to follow a $\chi^2$ distribution when the observed variables are not multivariate normal and or when $N$ is small (Bentler, 1990; Jöreskog & Sörbom, 1981). Even when its assumptions are met, the $\chi^2$ test tends to reject true models at higher than the nominal rate in small samples (Boomsma, 1982); conversely, the $\chi^2$ test often has low power to detect meaningful levels of model misspecification in small samples (Gallini & Mandeville, 1984). Researchers have developed practical fit indices in an attempt to overcome some of these problems. Special emphasis has historically been placed on the criterion that the value of fit indices for correctly specified or slightly misspecified models should not be affected by sample size (e.g., Marsh, Balla, & McDonald, 1988).

**TABLE 13.1. Fit Indices for Covariance Structure Models**

| Equation No. | Fit index | | Reference | Goodness- or badness-of-fit index | Theoretical range | Cutoff criterion | Sensitive to $N$ | Penalty for model complexity? |
|---|---|---|---|---|---|---|---|---|
| T1 | $\chi^2 = (N-1)f$ | | Jöreskog (1969) | Badness | $\geq 0$ | $p < .05$ | Yes | No |
| T2 | $\chi^2 / df$ | (8) | Jöreskog (1969) | Badness | $\geq 0$ | $< 5$[d] | Yes | Yes |
| T3 | $\mathrm{GFI} = 1 - \frac{\mathbf{e'We}}{\mathbf{s'Ws}}$ | (10) | Jöreskog & Sörbom (1981) | Goodness | 0–1[a] | $> .95$[d] | Yes | No |
| T4 | $\mathrm{AGFI} = 1 - \frac{p^*}{df}(1-\mathrm{GFI})$ | (6) | Jöreskog & Sörbom (1981) | Goodness | 0–1[a] | N/A[d,e] | Yes | Yes |
| T5 | $\mathrm{GFI}^* = \frac{p}{p+2\left(\frac{\chi^2 - df}{N-1}\right)}$ | (0) | Maiti & Mukherjee (1990); Steiger (1989) | Goodness | 0–1[a] | $> .95$ | No | No |
| T6 | $\mathrm{AGFI}^* = 1 - \frac{p^*}{df}(1-\mathrm{GFI}^*)$ | (0) | Maiti & Mukherjee (1990); Steiger (1989) | Goodness | 0–1[a] | N/A[e] | No | Yes |
| T7 | $\mathrm{RMR} = [p^{*-1}(\mathbf{e'Ie})]^{1/2}$ | (4) | Jöreskog & Sörbom (1981) | Badness | $> 0$ | N/A[e,f] | Yes | No |
| T8 | $\mathrm{SRMR} = [p^{*-1}(\mathbf{e'W_s e})]^{1/2}$ | (13) | Bentler (1995) | Badness | $> 0$ | $< .08$ | Yes | No |
| T9 | $\mathrm{RMSEA} = \sqrt{\frac{\hat{\lambda}_N}{df}} = \sqrt{\frac{\max(\chi^2 - df, 0)}{df(N-1)}}$ | (42) | Steiger & Lind (1980) | Badness | $> 0$ | $< .06$ | Yes to small $N$ | Yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T10 | $TLI^c = \frac{\chi_0^2/df_0 - \chi_k^2/df_k}{\chi_0^2/df_0 - 1}$ | (22) | Tucker & Lewis (1973) | Goodness | 0–1 [a, b] | > .95 | No | Yes |
| T11 | $NFI = \frac{f_0 - f_k}{f_0} = \frac{\chi_0^2 - \chi_k^2}{\chi_0^2}$ | (7) | Bentler & Bonett (1980) | Goodness | 0–1 | > .95[d] | Yes | No |
| T12 | $IFI = \frac{\chi_0^2 - \chi_k^2}{\chi_0^2 - df_k}$ | (3) | Bollen (1989); Marsh et al. (1988) | Goodness | > 0[b] | > .95 | Yes to small *N* | Yes |
| T13 | $RNI = \frac{(\chi_0^2 - df_0) - (\chi_k^2 - df_k)}{(\chi_0^2 - df_0)}$ | (3) | Bentler (1990); McDonald & Marsh (1990) | Goodness | > 0[b] | >.95 | No | Yes |
| T14 | $CFI = \frac{max(\chi_0^2 - df_0, 0) - \max(\chi_k^2 - df_k, 0)}{\max(\chi_0^2 - df_0, 0)}$ | (42) | Bentler (1990) | Goodness | 0–1 | > .95 | No | Yes |

*Note.* $\chi^2$, chi-square test statistic; GFI = goodness-of-fit index; AGFI, adjusted goodness-of-fit index. GFI*, revised GFI; AGFI*, revised AGFI; RMR, root mean square residual; SRMR, standardized root mean square residual; RMSEA, root mean square error of approximation; TLI, Tucker–Lewis index; NFI, normed fit index; IFI, incremental fit index; RNI, relative noncentrality index; CFI, comparative fit index; *f*, minimized discrepancy function; *o*, baseline model; *k*, tested or hypothesized model; *df*, degrees of freedom; *N*, sample size; $p^*$, the number of nonduplicated elements in the covariance matrix; **e**, a vector of residuals from a covariance matrix; **s**, a vector of the $p^*$ nonredundant elements in the observed covariance matrix; **I**, an identify matrix; **W**, a weight matrix; $\mathbf{W}_s$, a diagonal weight matrix used to standardize the elements in a sample covariance matrix; $\lambda_N$, noncentrality parameter, normed so that it is not negative. The numbers in parentheses in the "Fit indices" column represent the number out of 55 articles on structural equation models in substantive American Psychological Association journals in 2004 that reported each of the practical fit indices described here (see Taylor, 2008). No other practical fit indices were reported.

[a] Can be negative. Negative value indicates an extremely misspecified model. [b] When exceeds 1, the fit index indicates extremely well-fitting model. [c] Also called non-normed fit index (NNFI). [d] Fit index is affected by sample size. [e] No cutoff criteria have been proposed for this index. [f] Not standardized, so will be affected by size of elements in covariance matrix.

## Practical Fit Indices

The decade of the 1980s was the heyday of the development of new fit indices, and—with apologies to songwriter Paul Simon—there must be 50 ways to index your model's fit (see Marsh, Hau, & Grayson, 2005, for a list of 40). In this section we focus on several practical fit indices commonly reported in published articles. Table 13.1 reports the fit indices identified based on a computer and manual search of American Psychological Association journals (Taylor, 2008; see also Jackson, Gillapsy, & Purc-Stephenson, 2009). Good (and bad) reasons exist for the use of these particular indices, such as the precedent of use by other researchers, their routine computation by SEM software, and positive evaluations in reviews (e.g., Hu & Bentler, 1998).

Following McDonald and Ho (2002), we distinguish between absolute and comparative fit indices. Absolute fit indices are functions of the test statistic $T$ or of the residuals (Yuan, 2005). In contrast, comparative fit indices assess the improvement in fit of the hypothesized model relative to a baseline model. The most restricted model that is "theoretically defensible" (Bentler & Bonett, 1980) has become the standard baseline model estimated by most SEM software packages (e.g., EQS, LISREL, Mplus). This independence model estimates a variance for each measured variable but permits no covariances between measured variables (see Figure 13.2A). This standard baseline model is not always appropriate for more complex SEM models (McDonald & Marsh, 1990; Widaman & Thompson, 2003; see Figure 13.2B). Other baseline models may be justified in some research contexts, even for CFA models (e.g., Sobel & Bohrnstedt, 1985). Another distinction is between goodness- and badness-of-fit indices. Goodness-of-fit indices increase (often to a maximum value of 1) with improving fit. Badness-of-fit indices decline (often to 0) with improving fit. All comparative fit indices are goodness-of-fit indices; absolute fit indices can be either goodness- or badness-of-fit indices.

Of the fit indices presented in Table 13.1, the root mean square error of approximation, the standardized root mean square residual, the goodness-of-fit index, the $\chi^2/df$ ratio, the adjusted goodness-of-fit index, and the root mean square residual are absolute indices; the comparative fit index, the Tucker–Lewis index, the normed fit index, the relative noncentrality index, and the incremental fit index are comparative fit indices. We consider the absolute indices first, followed by the comparative indices, with each group presented in roughly

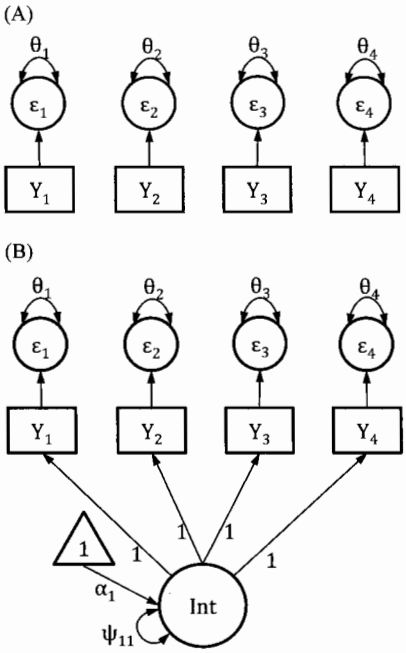


**FIGURE 13.2.** (A) Baseline model for a confirmatory factor analysis model with four indicators. (B) Baseline (intercept only) model for linear growth model with four time points.

their order of introduction in the literature. Not all of these fit indices are currently recommended; all continue to appear with some frequency in published SEM applications. We note commonly used cutoff values proposed for those indices that are not affected by $N$.

### $\chi^2/df$ Ratio

The $\chi^2/df$ ratio was never formally introduced as a fit index but appears to have evolved as an easily computed, ad hoc measure of fit. Jöreskog (1969), in his consideration of limitations of the $\chi^2$ test of overall fit, suggested that the $\chi^2$ value be used more descriptively in the evaluation of model fit, with the $df$ acting as a standard of comparison. The rationale for the $\chi^2/df$ ratio[1] is that the expected value of the $\chi^2$ for a correct model equals the $df$. Wheaton, Muthén, Alwin, and Summers (1977) explicitly introduced the $\chi^2/df$ ratio with little comment except to indicate that their experience suggested that a value of 5 or less indicated good fit; this proposed

reference value is heavily influenced by $N$ (Marsh et al., 1988). Given a fixed $N$, smaller values of the $\chi^2/df$ ratio indicate better fit; it is a badness-of-fit index. The $\chi^2/df$ ratio has a minimum of 0, which occurs when a model with positive $df$ has a $\chi^2$ value of 0. Saturated models, which by definition fit perfectly, have 0 $df$; therefore, they have an undefined $\chi^2/df$. There is no theoretical maximum for the $\chi^2/df$ ratio.

Unlike $\chi^2$, which can only remain constant or improve as parameters are added to a model, the $\chi^2/df$ ratio can potentially get worse. The $\chi^2/df$ ratio penalizes model complexity. If added parameters fail to reduce a model's $\chi^2$ appreciably, the $\chi^2/df$ ratio will get worse because adding parameters reduces the model's $df$. The $\chi^2/df$ ratio suffers from the same problem as the $\chi^2$ test—its value is dependent on sample size for misspecified models (Marsh et al., 1988).

## Goodness-of-Fit and Adjusted Goodness-of-Fit Indices

Jöreskog and Sörbom (1981) introduced the goodness-of-fit (GFI) and adjusted goodness-of-fit (AGFI) indices. They described these indices as proportions of variance accounted for, but their formulas did not make this interpretation transparent. Bentler (1983, Equation 3.5) later reexpressed the GFI formula, clarifying this interpretation (see Table 13.1, Equation T3). Equation T3 uses a weight matrix $\mathbf{W}$ that is computed from the elements of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})^{-1}$ for ML and $\mathbf{S}^{-1}$ for GLS. Thus, GFI is calculated using the weighted sum of squared residuals from a covariance matrix and weighted sums of squared variances and covariances. It is similar to the familiar $R^2$ measure used in ordinary least squares (OLS) regression, which can be expressed as

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \tag{13.5}$$

The major difference[2] between Equation T3 and Equation 13.5 is the GFI's use of the weight matrix $\mathbf{W}$. This matrix, which appears in the fit function, relates the GFI directly to the estimation procedure, which is typically a desirable property for a fit measure (Menard, 2000).

Jöreskog and Sörbom (1981) presented the AGFI as an adjustment to the GFI based on a model's $df$ (Table 13.1, Equation T4). The goal of the adjustment was to penalize model overfitting, in which additional parameters are estimated with small resulting improvement in fit. Equation 13.6 reexpresses Equation T4 to make the relationship between the GFI and AGFI clearer:

$$\frac{(1 - \text{GFI})}{(1 - \text{AGFI})} = \frac{df}{p^*} \tag{13.6}$$

Equation 13.6 shows that the AGFI will be smaller than the GFI for all realistic models in which at least one parameter is estimated ($df < p^*$). The AGFI will approach the GFI as fewer parameters are estimated (as $df$ approaches $p^*$).

Both the GFI and the AGFI are goodness-of-fit indices, increasing with improving fit. They are proportions that conceptually have a range of 0 to 1, but can potentially be negative (Jöreskog & Sörbom 1981; Maiti & Mukherjee, 1990). The GFI will be negative if $\mathbf{e}'\mathbf{We} > \mathbf{s}'\mathbf{Ws}$ (see Equation T3 in Table 13.1), meaning that the weighted squared residuals are actually larger than the weighted squared covariances in $\mathbf{S}$! This result is theoretically possible, but only in extremely badly misspecified models that would never be entertained by researchers. In contrast, the AGFI can become negative whenever GFI $< (p^* - df)/p^* = q/p^*$. In other words, the AGFI will be negative whenever the proportion of variance accounted for by a model, as measured by the GFI, is smaller than the proportion of the $p^*$ observed covariances used to estimate parameters.

Mulaik and colleagues (1989) noted that the relationship between the AGFI and the GFI is analogous to the relationship between $R^2$ and adjusted $R^2$ (Wherry, 1931) in OLS regression. They critiqued the AGFI because, as noted earlier, it can fall below 0 (as can adjusted $R^2$). Given that the AGFI is in a proportion metric, negative values are mathematically uninterpretable, although such values could only occur with an extremely misspecified model. Mulaik and colleagues also questioned the penalty used by the AGFI to choose more parsimonious models: The GFI is not very sensitive to changes in a model's $df$ when the model has a large $df$ to begin with, particularly as the GFI approaches 1.

Maiti and Mukherjee (1990, Equation 19) and Steiger (1989, Equation 51) suggested a revised index known as GFI* (a.k.a., gamma hat) that improves on the properties of the GFI (Table 13.1, Equation T5). Steiger demonstrated that although the GFI and GFI* asymptotically estimate the same quantity, the GFI is biased and the GFI* is unbiased in smaller samples. An unbiased estimate of the AGFI, the AGFI*, can also be calculated by substituting the GFI* for the GFI in Equation T4, yielding Equation T6 in Table 13.1. In contrast to

GFI and AGFI, which are affected by sample size, GFI* and AGFI* are expected to have the desirable erty of not being affected by *N*. They are promis- global fit indices (see Hu & Bentler, 1998; Taylor, 3), but to date have been little used in practice.

## t Mean Square Residual and Standardized Root n Square Residual

skog and Sörbom (1981) also introduced the root n square residual (RMR), which is the square root ne average of the squared residuals (Table 13.1, ation T7). Recall that residuals are differences veen observed covariances and model-implied co- ances ($\mathbf{s} - \hat{\sigma}(\theta)$) rather than differences between erved scores and predicted scores ($Y - \hat{Y}$). Equation n Table 13.1 clarifies the relationship between the R and the GFI. Rather than using a weight matrix he RMR uses the identity matrix **I**. RMR depends in unweighted rather the weighted function of the duals.

he RMR's use of unweighted residuals can be a ue, particularly for observed measures with little surement error (e.g., some cognitive or biological sures). Browne, MacCallum, Kim, Andersen, and ser (2002) demonstrated that the weighting of re- als in the ML and GLS fit functions (see Equation ?) can severely overstate a model's badness of fit en the measured variables have small unique vari- es. The RMR does not weight the residuals in its ulation, so it is unaffected by this problem; all other ndices discussed in this chapter (except for the stan- dized root mean square residual, discussed immedi- y below) are affected. The GFI and the AGFI use the ne weighting as the ML or GLS fit functions; other ndices incorporate the fit function through their use $^2$, which is equal to $(N-1)f$ in their equations.

he RMR is a badness-of-fit index; it approaches 0 he fit of a model improves. Unfortunately, its scal- can impede interpretation as it diverges from 0. RMR will tend to be larger for covariance matri- with larger elements than for matrices with smaller ments, precluding comparisons across data sets. tler (1995) introduced the standardized root mean are residual (SRMR) to address this comparison blem. The SRMR converts the residuals into a stan- dized metric. Each standardized residual that goes the calculation of the SRMR is the raw residual a proportion of the element of **S** being estimated. ause of this standardized metric, SRMR values can be meaningfully compared across models fit to different data sets. The calculation of the SRMR is similar to the calculation of the RMR (Table 13.1, Equation T8), except that it uses a diagonal weight matrix $\mathbf{W}_s$ to standardize the elements in **S**, whereas the RMR uses an identity matrix, leaving the elements unstandardized. Each diagonal element of $\mathbf{W}_s$ is the reciprocal of the square root of the product of the variances on which the corresponding element of **S** is based. For example, for $s_{12}$ and $\hat{\sigma}_{12}$, the corresponding diagonal element of $\mathbf{W}_s$ has a value of $\sqrt{s_{11}s_{22}}$.

The SRMR's weight matrix $\mathbf{W}_s$ is diagonal, whereas the weight matrix **W** in the fit function in Equation 13.2 is, in general, not diagonal. Although the SRMR differs from the RMR in that it is standardized, it is like the RMR in that its weighting of the residuals ignores the possible covariance of the elements in **S** or $\hat{\boldsymbol{\Sigma}}$, taken into account by the ML or GLS fit function. This outcome implies that it is also like the RMR in being immune to the problem discussed earlier of overstating misfit when several manifest variables in a model have small unique variances (see Browne et al., 2002).

Like the RMR, the SRMR is a badness-of-fit index. It has a minimum of 0 for a perfectly fitting model. In practice, the SRMR will be less than 1, typically far less. An SRMR of 1 would indicate that the residuals were, on average, as large as the elements of **S** being estimated, an extremely poorly fitting model that no researcher would seriously consider.

## Root Mean Square Error of Approximation

The root mean square error of approximation (RMSEA; Steiger, 1989, 1990; Steiger & Lind, 1980) is based on the insight that although $(N-1)f$ asymptotically follows the familiar (central) $\chi^2$ distribution under the null hypothesis, it asymptotically follows a *noncentral* $\chi^2$ distribution under the alternate hypothesis. The noncentrality parameter ($\lambda$) of this distribution depends on how badly the model fits, so it can be used to construct a fit index. Since the expected value of a noncentral $\chi^2$ distribution is $df + \lambda$, Steiger (1989) pointed out that the noncentrality parameter could be estimated as

$$\hat{\lambda} = (\chi^2 - df)/(N-1) \qquad (13.7)$$

To keep this estimated noncentrality parameter from taking on an unrealistic negative value, Steiger suggested that $\lambda$ be given a lower bound of 0.

$$\hat{\lambda}_N = \max(\chi^2 - df,\ 0)/(N-1) \quad \textbf{(13.8)}$$

where the N subscript indicates that $\hat{\lambda}_N$ has been normed to keep it non-negative.

Steiger and Lind (1980) suggested two adjustments to $\hat{\lambda}_N$ to improve the RMSEA's interpretation. First, they added a penalty function to discourage researchers from overfitting models, dividing $\hat{\lambda}_N$ by its *df*. Second, they took the square root of this result, so that the RMSEA is in the same metric as the weighted residuals (see Equation T9 in Table 13.1). Steiger (1989; Steiger & Lind, 1980; see also Browne, 1974) showed that the population noncentrality parameter being estimated by $\hat{\lambda}$ could be considered as a weighted sum of squared residuals (see Equation 13.2)

$$\hat{\lambda} = \mathbf{e}'\mathbf{W}\mathbf{e} \quad \textbf{(13.9)}$$

The residuals are then weighted in the same manner as in the ML or GLS estimation procedure because the weight matrix **W** is the same.

The RMSEA is a badness-of-fit index, declining with improving fit. The RMSEA is bounded at a lower value of 0. It has no theoretical maximum. Browne and Cudeck (1993) suggested that a model with an RMSEA of .10 is unworthy of serious consideration.

A confidence interval (CI) for the RMSEA is provided by most computer programs. An iterative procedure is used to find limits of a CI for $\hat{\lambda}_N$, and then these limits are substituted into the left formula of Equation T9 in Table 13.1. Steiger and Lind (1980) advocated using a 90% CI. Browne and Cudeck (1993) extended the use of this CI to a test of close fit. Noting that in their experience, RMSEA values of .05 or less indicated "close fit," they constructed a test of the null hypothesis that the true value of the RMSEA $\leq$ .05, now implemented in many SEM software packages. This null hypothesis that the model closely fits the data is retained if the lower limit of the RMSEA's confidence interval falls at or below .05. Alternatively, an RMSEA whose upper limit exceeded .08 or .10 could be deemed unacceptable. RMSEA underestimates fit at small sample sizes ($N < 200$; see Curran, Bollen, Chen, Paxton, & Kirby, 2003).

## Tucker–Lewis Index

Tucker and Lewis (1973) noted that the fit function $F$ (Equation 13.2) is a sum of squares that when divided by *df* yields a mean square $M$. For exploratory factor analysis they proposed the Tucker–Lewis index (TLI), which compares $M_k$ for the hypothesized model to $M_0$ for a baseline, independence model. (In this and subsequent equations for comparative fit indices, quantities subscripted with 0 come from the baseline model and quantities subscripted with $k$ come from the hypothesized model.)

Bentler and Bonett (1980) generalized the TLI to the covariance structure analysis context and labeled it the non-normed fit index (NNFI), although the TLI designation remains more common. They formulated the TLI in terms of $\chi^2/df$ ratios (see Table 13.1, Equation T10). Their formulation makes clear that the TLI is conceptually in a proportion metric. In terms of $\chi^2/df$ ratios, it gives the distance between the baseline and target models as a proportion of the distance between the baseline model and a true model. The 1 in the denominator is the expected value of the $\chi^2/df$ ratio for a true model.

Although the TLI is conceptually in a proportion metric, it can potentially fall below 0 or above 1. TLI can occasionally exceed 1 if $\chi^2_k / df_k < 1$. By contrast, TLI can be negative if the denominator is negative and the numerator is positive. Both conditions under which the TLI becomes mathematically negative, $\chi^2_k / df_k < \chi^2_0 / df_0 < 1$ and $1 < \chi^2_0 / df < \chi^2_k / df_k$, require the baseline model to fit the data very well, a condition that is unlikely to occur in practice.

The TLI penalizes models that estimate many parameters. McDonald and Marsh (1990) showed that it could be rewritten in terms of James, Mulaik and Brett's (1982; see also Mulaik et al., 1989) parsimony ratio (PR): $PR = df_k/df_0$. Thus, PR is the proportion of the number of parameters fixed in the hypothesized model relative to the proportion of the number of parameters fixed in the baseline independence model. McDonald and Marsh's reexpression of the TLI is given in Equation 13.10:

$$TLI = 1 - \frac{(\chi^2_k - df_k)/(\chi^2_0 - df_0)}{df_k / df_0} = 1 - \frac{\hat{\lambda}_k / \hat{\lambda}_0}{PR} \quad \textbf{(13.10)}$$

Given equal model fit, models with larger PRs yield larger TLI values. Bollen (1986), in an early critique, argued that the TLI would be affected by sample size; however, Monte Carlo studies (e.g., Marsh et al., 1988) have consistently found that the TLI is not affected by sample size (see also Balderjahn, 1988).

## ed Fit Index

r and Bonett (1980) also introduced the normed ex (NFI), which compares the fit of a target model fit of a baseline model. Rather than use $\chi^2/df$ ra- the TLI does, it uses either fit function values alues (Table 13.1, Equation T11). The expression the NFI indicates the improvement in fit realized ving from the baseline model to a hypothesized , as a proportion of the baseline model's fit. The pression for the NFI (Equation T11 in Table 13.1) so be used even when the fit function is not re- o the $\chi^2$ distribution. The NFI cannot fall below bove 1. The NFI cannot fall below 0 because the ne model must be nested within the hypothesized , so the hypothesized model cannot have a worse r) $\chi^2$. It cannot exceed 1 because the minimum of the hypothesized model's $\chi^2$ value is 0, which the maximum NFI equal to $\chi_0^2 / \chi_0^2 = 1$. The NFI ot include a penalty function to penalize overfit- nd is affected by $N$, with small sample sizes pro- underestimates of the true NFI (Marsh et al.,

## nental Fit Index

(1989; see also Marsh et al., 1988) introduced remental fit index (IFI) in an attempt to improve NFI. The NFI does not approach 1 for correct s in small samples (Bentler, 1990). The key prob- that the expected value of a model's $\chi^2$ for cor- odels does not equal zero as the NFI assumes, tead equals the model's $df$. The IFI subtracts the esized model's $df$ in the denominator, as this is pected value of a model's $\chi^2$ if the model is cor- able 13.1, Equation T12). The IFI is theoretically oportion metric, but it can potentially exceed 1. do so under precisely the same circumstances as I: when the hypothesized model's $\chi^2$ is less than Also like the TLI, the IFI can be negative, but $\chi_0^2 < df_k$, again suggesting a remarkably good the baseline model. Like the NFI, the IFI's nu- r cannot be negative: The baseline model must ted in the hypothesized model, so the baseline s $\chi^2$ cannot be smaller than that of the hypoth- model.

ough proposed as an improvement to the NFI, introduced new problems. First, McDonald and (1990) showed that the IFI will tend to overesti- s asymptotic value in small samples; this over- estimation will be more severe as the misspecification of the hypothesized model increases, as indexed by its noncentrality parameter $\lambda$. The IFI's positive bias in small samples is probably a greater concern than the NFI's negative bias, as positive bias leads to conclusions that a model fits better than it actually does. Negative bias can have the virtue of encouraging conservative conclusions about model fit (Marsh, Balla, & Hau, 1996).

Second, the inclusion of the model's $df$, which should act as a penalty function for overly complex models like that of the TLI, actually works in the wrong direction (Marsh, 1995; Marsh et al., 1996). If a superfluous parameter is added to a model, the model's $df$ will be reduced by 1, but its $\chi^2$ will not decrease, meaning that the IFI's denominator will decrease while its numerator will remain unchanged, resulting in a larger IFI value. Marsh and colleagues (1996) refer to this as a "penalty for parsimony," noting that it runs counter to the more desirable behavior of the TLI, which penalizes for unnecessary complexity.

## Comparative Fit Index and Relative Noncentrality Index

Bentler (1990) and McDonald and Marsh (1990) independently introduced two virtually identical fit indices. McDonald and Marsh introduced the relative noncentrality index (RNI), which uses the noncentrality parameter as an index of lack of fit just as the RMSEA does. The noncentrality parameter is estimated using Equation 13.7, just as it is for the RMSEA. The RNI then takes a form similar to that of the other comparative fit indices, giving the reduction in noncentrality realized by moving from the baseline to the hypothesized model, as a proportion of the baseline model's noncentrality (Table 13.1, Equation T13). The RNI converges asymptotically to the same value as do the NFI and the IFI, but has the desirable property of being unaffected by sample size. The RNI can exceed 1 under the same unlikely circumstances that the TLI and the IFI do: when the hypothesized model's $\chi^2$ is smaller than its $df$.

In defining the CFI, Bentler used the same logic as Steiger and Lind (1980) with the RMSEA and fixed the estimated noncentrality parameter to have a minimum of 0. Doing this replaces $\hat{\lambda}$ from Equation 13.7 with $\hat{\lambda}_N$ from Equation 13.8 and yields the CFI's formula (Table 13.1, Equation T14). In models for which the $\chi^2$ is larger than the $df$, which likely includes the great majority of

models tested in psychological research, the CFI and RNI take on identical values. The RNI and CFI will differ only when a model's $\chi^2$ is smaller than its *df*, characteristic of extremely well fitting models. Under such circumstances, the RNI exceeds 1, whereas the CFI is bounded at the maximum theoretical value of 1. Goffin (1993) pointed out that the RNI and the CFI estimate the same population quantity, but this difference means that they have different strengths. The RNI is a less biased estimator than the CFI because it does not truncate its distribution at 1. The CFI is a more efficient estimator (smaller standard error) because its truncated distribution discards values that the population index cannot possibly take on. Goffin suggested that these qualities make the RNI preferable for comparing competing models, and the CFI preferable for reporting the fit of a single model. Both the CFI and RNI are straightforward to interpret and are not affected by *N*.

### Summary

Our review thus far has considered the characteristics of commonly used practical fit indices and their performance in simple CFA models in which each factor has a small number of measured indicators. Researchers have strongly preferred fit indices whose mean values in simulation studies are independent of *N* (e.g., Marsh et al., 1988). This preference parallels psychology's increasing use of effect sizes that are independent of *N* rather than *p*-values, which are strongly related to *N* (Wilkinson and Task Force on Statistical Inference, 1999). Other desirable unique properties of a specific fit index (e.g., the confidence interval of the RMSEA; the proportion of variance interpretation of GFI[3]) may argue for its use so long as a minimum sample size is exceeded that makes bias in its estimation trivially small. A second important issue is ease of interpretation. Indices in a proportion fit metric or standardized metric that is unaffected by the scaling of the measured variables will be easier to interpret than indices without these qualities. Using these criteria to cull the fit indices reviewed earlier, the fit indices commonly reported in the literature that are worthy of consideration are the SRMR (given its standardized metric), RMSEA (for sample sizes over 200), TLI, and CFI/RNI. The TLI and CFI/RNI are goodness-of-fit indices in a proportion fit metric, whereas the RMSEA and SRMR are badness-of-fit indices that are not in a proportion metric. Other evaluations of more extensive sets of fit indices (Hu & Bentler, 1998; Marsh et al., 2005) also provide favorable evaluations of these fit indices, as well as others with which there is far less practical experience.

## Proposed Cutoff Values

Most researchers focus on the first question posed at the beginning of this chapter: Does the hypothesized model provide an adequate fit to the data? Higher values on goodness-of-fit indices and lower values on badness-of-fit indices indicate better overall fit of the model to the data. But, what is an "adequate" fit? Researchers ideally desire a comparison standard that specifies a single criterion value that defines adequate fit.

Bentler and Bonett (1980) originally suggested a standard of .90 for the NFI and TLI (NNFI), fit indices in the proportion metric (also including the CFI/RNI reviewed earlier). Hu and Bentler (1995) proposed a criterion of <.05 for what they termed "good fit" and from .05 to .10 for "acceptable fit" for the SRMR. Browne and Cudeck (1993) suggested for the RMSEA that a value of .05 represented what they termed a "close fitting model" and .08 represented an "adequate" fitting model. These recommendations were based on the researchers' practical experience with the fit indices in the evaluation of many CFA models. Hu and Bentler (1999) later took another approach, conducting a simulation study that addressed the ability of fit indices to distinguish between correctly specified and misspecified models. Based on this study, they proposed a criterion of .95 for the TLI and CFI, a criterion of .06 for the RMSEA, and a criterion of .08 for the SRMR. Thus, Hu and Bentler proposed replacing the initial ad hoc practical guidelines with standards based on the results of a simulation study using a small set of correctly specified and misspecified covariance structure models. The rationale for their proposed standards, which focuses on the acceptance versus rejection of hypothesized models, has been questioned by Marsh, Hau, and Wen (2004) because it implicitly reintroduces sample size as a determinant of the outcome.

We believe that the proposed cutoff values can be guidelines about the overall fit of the model to the data, but we caution readers that the reification of specific cutoff standards for the acceptance versus rejection of a hypothesized model can be fraught with peril. The next section examines three important issues related to the use of cutoff values for fit indices.

plications. Under these conditions, the GLS weight and ML weight matrices computed from the elements $\mathbf{S}^{-1}$ and $\hat{\Sigma}(\theta)^{-1}$, respectively, will typically vary appreciably. Of note, LISREL uses GLS estimation for its baseline model, whereas most other SEM packages use the same procedure (typically ML in practice) as is used to estimate the hypothesized model. Values of comparative fit indices estimated using different estimation procedures will differ, perhaps substantially (Tanaka, 1993).

A second issue identified by Widaman and Thompson (2003) is that the baseline model must be nested within the hypothesized model. When mean structures are included, or certain restrictions are placed on the model, modified baseline models must be used in the calculation of comparative fit indices, or the value of the fit indices will be incorrect, sometimes appreciably so. Wu and colleagues (2009) discuss this and specify an acceptable baseline model (see Figure 13.2B) in the context of growth curve models.

Finally, some models do not have a proper saturated (0 *df*) model in which the number of estimated parameters matches the number of observed means and covariances in the model. Growth curve models in which each individual is measured at a different set of time points (so-called "random time models") and models with certain patterns of missing data do not have a saturated model (Wu et al., 2009). In addition, the standard saturated model used for linear structural equation models is not appropriate for models with interactions or quadratic effects of latent variables. The standard $\chi^2$ test statistic and all practical fit indices based on the $\chi^2$ reported by computer programs will be incorrect. Klein and Schermelleh-Engel (2010) provide a method of estimating $\chi^2$ based on an appropriate saturated model for these cases.

Each of these issues illustrates the need for careful attention to the baseline and saturated models in the calculation of comparative fit indices.

## Encouragement of Poor Practices

Reliance on fixed comparison standards for fit indices can also encourage poor practice by researchers. First, researchers, despite hypothesizing a model based on prior theory and research, may add and delete paths and factor loadings based on modification indices until the prescribed threshold standard for adequate fit is met. Traditional post hoc model modification, particularly

1990; MacCall
Necowitz, 1992
40, this volume
promising). It is
pothesized mod
post hoc modifi
only to find that
gain in fit over t
perhaps a worse
is used). Modific
the epistemologi
confirmatory to
has a clear mean
on the basis of t
acknowledgment
it also requires fa
statistics and hyp
tion of the model
Reporting of fit in
as the TLI or RM
nate, this problem
ships in the data.

Second, fixed s
tion of constructs
theory, we would
items that covers
provides a relativ
score across the
bretson & Reise,
only a small nun
Marsh and colleag
(2003) found, fit
ber of indicators p
model is properly
models will also t
hypothesized cons
become more diss
1969). As Marsh
importance typica
one model against
naturally extingui
using larger numb
part why so many
[indicators per fac

In sum, attemp
equate fit encoura

## ISSUES WITH PRACTICAL FIT INDICES

### Model Characteristics and Standards for Fit

We earlier summarized the results of an extensive body of simulation research attempting to identify practical fit indices whose estimates are not affected by sample size. Unfortunately, much less research has investigated the effect of other model characteristics on fit. The available results suggest that other model and data characteristics can substantially affect the performance of fit indices. Within CFA models, Chen, Curran, Bollen, Kirby, and Paxton (2008; see also Savalei, 2011) showed that model specification and *df* can affect the performance of the RMSEA. Marsh, Hau, Balla, and Grayson (1998) have found that as the number of indicators per factor increases, models showed decreased fit to simulated data with properly specified models. Kenny and McCoach (2003) found that all fit indices examined, with the exception of the RMSEA, showed decreased fit as more indicators were added to a single-factor model. Marsh and colleagues note that "this apparent decline in fit associated with larger [number of indicators per factor] must reflect problems in the standards used to evaluate model fit rather than misspecification in the approximating model" (p. 217). Saris, Satorra, and van der Veld (2009) have found that given a constant magnitude of misspecification and sample size, the numerical value of other parameters in a model can affect the value of fit indices, with, for example, higher factor loadings leading to poorer fit index values. Davey, Savla, and Luo (2005) found that the values of fit indices for slightly misspecified CFA models increased as the proportion of missing data increased. Adding random error to a model may improve its apparent fit! This is not a desirable property.

As mean structures are added to models, other issues arise. The SRMR as commonly calculated only addresses the discrepancies between the model's implied and observed covariances; the mean structure is ignored. For practical fit indices based on the $\chi^2$ test statistic, the fit function adds another term to capture the discrepancy between the observed and model implied means. A general discrepancy function extends Equation 13.2 to mean and covariance structures (Browne & Arminger, 1995):

$$F = [\mathbf{s} - \hat{\sigma}(\theta)]\mathbf{W}^{-1}[\mathbf{s} - \hat{\sigma}(\theta)] \qquad (13.11)$$

where $\mathbf{W}$ and $\mathbf{V}$ are weight matrices, and $\bar{X}$ and $\hat{\mu}(\theta)$ are the vectors of observed means and model implied means, respectively. The first term assesses fit in the covariance structure; the second assesses fit in the mean structure. Wu, West, and Taylor (2009) note the complexity of assessing fit for growth models given that misspecification in one structure can affect the other structure. In addition, the metrics of fit in the two structures may be quite different: Taking a traditional standard for the GFI of .90 or .95 for a CFA model in a large sample may be an appropriate proportion of variance for which to account (cf. Tanaka & Huba, 1985), but do we also expect our model to account for 90% or 95% of the variance in the latent means? Experience with other models such as analysis of variance with reliably measured outcomes would not lead us to expect such high values. Based on Wu and West's (2010) study of the effects of different types of model misspecification and different data characteristics (e.g., the ratio of the Level 1 to Level 2 variances) on the fit of growth curve models, Wu attempted to develop standards for fit indices. She abandoned this effort because standards following Hu and Bentler's (1999) accept–reject criterion varied dramatically as a function of the type of misspecification and data characteristics.

Taken together, these results suggest that appropriate cutoff standards may be specific to particular models and data sets. Current standards for interpreting acceptable model fit are only rough guidelines; they become increasingly less reasonable as they are extrapolated to models and data further from the CFA models with complete data studied by Hu and Bentler (1999).

### Baseline and Saturated Models

Several smaller but easily overlooked issues relate to baseline and saturated models. For comparative fit indices, most commonly used SEM programs use the baseline model proposed by Bentler and Bonett (1980), which estimates a model in which each variable has a variance, but in which there are no covariances between variables (see Figure 13.2A). Because of the different weight matrices used in estimating the baseline model, comparative fit indices based on different estimation methods (e.g., GLS, ML) will differ (Sugawara & MacCallum, 1993). GLS and ML do produce the same results in large samples if data have the typically assumed multivariate normal distribution and the hypothesized

plications. Under these conditions, the GLS weight and ML weight matrices computed from the elements $\mathbf{S}^{-1}$ and $\hat{\Sigma}(\theta)^{-1}$, respectively, will typically vary appreciably. Of note, LISREL uses GLS estimation for its baseline model, whereas most other SEM packages use the same procedure (typically ML in practice) as is used to estimate the hypothesized model. Values of comparative fit indices estimated using different estimation procedures will differ, perhaps substantially (Tanaka, 1993).

A second issue identified by Widaman and Thompson (2003) is that the baseline model must be nested within the hypothesized model. When mean structures are included, or certain restrictions are placed on the model, modified baseline models must be used in the calculation of comparative fit indices, or the value of the fit indices will be incorrect, sometimes appreciably so. Wu and colleagues (2009) discuss this and specify an acceptable baseline model (see Figure 13.2B) in the context of growth curve models.

Finally, some models do not have a proper saturated (0 *df*) model in which the number of estimated parameters matches the number of observed means and covariances in the model. Growth curve models in which each individual is measured at a different set of time points (so-called "random time models") and models with certain patterns of missing data do not have a saturated model (Wu et al., 2009). In addition, the standard saturated model used for linear structural equation models is not appropriate for models with interactions or quadratic effects of latent variables. The standard $\chi^2$ test statistic and all practical fit indices based on the $\chi^2$ reported by computer programs will be incorrect. Klein and Schermelleh-Engel (2010) provide a method of estimating $\chi^2$ based on an appropriate saturated model for these cases.

Each of these issues illustrates the need for careful attention to the baseline and saturated models in the calculation of comparative fit indices.

## Encouragement of Poor Practices

Reliance on fixed comparison standards for fit indices can also encourage poor practice by researchers. First, researchers, despite hypothesizing a model based on prior theory and research, may add and delete paths and factor loadings based on modification indices until the prescribed threshold standard for adequate fit is met. Traditional post hoc model modification, particularly when undertaken on an atheoretical basis, is unlikely to [illegible] of an improved model (Kaplan, 1990; MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992; but see Marcoulides & Ing, Chapter 40, this volume, for newer approaches that are more promising). It is a sobering exercise to modify one's hypothesized model and then to test the hypothesized and post hoc modified model in a new replication sample, only to find that modified model does not lead to any gain in fit over the originally hypothesized model (and perhaps a worse fit, if a fit index with a penalty function is used). Modification of hypothesized models changes the epistemological status of the tested model from confirmatory to exploratory. The LR ($\chi^2$) test no longer has a clear meaning once the model has been modified on the basis of the data. The change requires explicit acknowledgment of the model's new exploratory status; it also requires far more tentative reporting of model fit statistics and hypothesis tests in the absence of replication of the model in a new sample (see Diaconis, 1985). Reporting of fit indices with a parsimony penalty such as the TLI or RMSEA can reduce, but does not eliminate, this problem of capitalization on chance relationships in the data.

Second, fixed standards can lead to poor representation of constructs. From the standpoint of modern test theory, we would like to have a measure with multiple items that covers the full content of the construct and provides a relatively precise estimate of each person's score across the full range of the construct (e.g., Embretson & Reise, 2000). Yet most CFA models utilize only a small number of indicators per construct. As Marsh and colleagues (1998) and Kenny and McCoach (2003) found, fit indices tend to decrease as the number of indicators per construct increases, even when the model is properly specified. Fit indices for hypothesized models will also tend to decrease as the coverage of the hypothesized construct improves because the items will become more dissimilar (cf. Tucker, Koopman, & Linn, 1969). As Marsh and colleagues note, "Because of the importance typically placed . . . on evaluating the fit of one model against a fixed standard . . . , this bias would naturally extinguish the possibly desirable strategy of using larger numbers of indicators. This may explain in part why so many published CFA studies are based on [indicators per factor] = 2 or 3" (p. 217).

In sum, attempting to meet cutoff standards for adequate fit encourages post hoc model modification and the use of a relatively small number of indicators of each latent construct, practices which are often nonoptimal from a scientific viewpoint.

## OTHER STRATEGIES FOR EVALUATING FIT

### Fit of Model Components

The fit indices considered so far provide information about overall model fit. In models with several components, researchers may place differential importance on the fit of each of the different components. In combining measures of the fit of each component to produce a measure of overall fit, no guarantee exists that the researcher's theoretically desired weights will match those produced by the computer software. Model components with extremely good or extremely poor fit, even if they are of little theoretical interest, may swamp the contribution of other model components in the calculation of global fit indices.

In their consideration of CFA models and structural path models between latent variables, Anderson and Gerbing (1988) originally proposed a two-step approach. The first step involved satisfactory specification of the measurement model by estimating a CFA model (saturating the $\Psi$ matrix of the covariances of latent factors). However, the challenge remained of assessing the fit of the structural model—the weight of the measurement structure in determining the fit of the overall model could potentially make it difficult to detect misspecification in the path model. In the context of multilevel SEM, Ryu and West (2009; see also Yuan & Bentler, 2007) proposed procedures for separately examining the Level 2 (between groups) and Level 1 (participants within groups) components of fit. Ryu and West showed that in the fit function, the between-groups component has far less weight (reflecting the number of groups) than the within-subjects component (reflecting the number of cases), giving the latter component disproportionate importance in determining overall model fit. They showed that improved results could be obtained using procedures that provided separate fit statistics for the Level 1 and Level 2 models. Wu and colleagues (2009) found that misfit in latent growth curve models can result from failure to reproduce the marginal means, the conditional means, the within-persons covariance structure, or the between-persons covariance structure. Wu and West (2010) presented methods for appropriately saturating different components of the model to provide more appropriate examination of the fit of the other components in the model. Such strategies can be useful in isolating the source(s) of model misfit, which is particularly important when some components of the model (e.g., marginal means; between-persons covariance structure) are of particular theoretical interest and other portions (e.g., within-persons covariance structure) are of far less theoretical interest.

### Examination of Individual Standardized Residuals

McDonald (1999, 2010) and McDonald and Ho (2002) have advocated an even more fine-grained analysis of fit—the separate examination of the each standardized residual in the covariance structure and, if applicable, the mean structure in the model. Models in which all residuals are not large are deemed to fit the data adequately. Models in which there are one or more large residuals indicate problems with model fit. By examining the individual standardized discrepancies between the observed and model-implied covariance or mean, "it becomes possible to judge whether a marginal or low index of fit is due to a correctable misspecification of the model, or to a scatter of discrepancies, which suggests that the model is possibly the best available approximation to reality" (McDonald & Ho, 2002, p. 73).

In sum, indices that assess the fit of theoretically important model components and examine individual residuals in the covariance and mean structures can provide a richer, more fine-grained understanding of the strengths and limitations of the hypothesized model in accounting for the data set that complements the use of global fit indices.

## WHAT IF THERE ARE ALTERNATIVE MODELS?

So far we have focused solely on cases in which there is assumed to be only one theoretically hypothesized model. However, often one of two other cases will occur. First, there may be alternative a priori theoretical models whose fit the researcher wishes to compare with that of the target model. This case can be particularly informative about the strengths and weaknesses of competing theoretical models in accounting for the data. Second, there may be other exploratory models proposed during the model fitting process that the researcher wishes to compare with the originally hypothesized model. Fit indices and model selection indices can be used to make these comparisons, again with the caveat that the second case requires appropriate acknowledgment of its exploratory status. Many discussions of model comparison emphasize the criterion of

parsimony—given similar overall model fit, the model with fewer parameters will be preferred over an alternative model with more parameters (e.g., Mulaik et al., 1989; Preacher, 2006). As considered in a later section, fit indices and model selection indices (see Table 13.2) that penalize model complexity will often, but not always, be preferred.

## Comparing Nested Models

As noted earlier, when the researcher wishes to compare two nested models, the likelihood ratio (LR) test discussed earlier can be used to determine whether the imposition of the restrictions on the more restricted model (yielding fewer parameters estimated) makes a statistically significant difference in model fit. Unfortunately, the same issues arise with the LR test for comparison of nested models that arose earlier for the overall $\chi^2$ fit test statistic. Small $N$'s can yield nonsignificant LR tests, masking important differences between the models. Conversely, very large $N$'s can produce statistically significant $\chi^2$ values even when the discrepancies between the two models are trivial.

Recognizing this issue, some researchers have argued that a change in practical fit indices that is less than some cutoff criterion may provide the desired information about differences in the fit between nested models. In the context of measurement invariance, which involves testing a series of nested measurement models (Widaman & Reise, 1997), Cheung and Rensvold (2002) suggested that changes in selected fit indices, including the CFI or GFI* among the fit indices reviewed earlier, appear to provide good performance in the assessment of measurement invariance. They proposed specified cutoff criteria for the change in fit between nested measurement invariance models (e.g., .01 for ΔCFI and .001 for ΔGFI*).

Note that neither of these fit indices includes a penalty for complexity, presumably because it is more difficult to compare corrected fit indices against an absolute standard for change. In addition, many of the same issues noted earlier with respect to the evaluation of fit indices against a cutoff criterion also appear to apply in this context (Fan & Sivo, 2009). Chen (2007) noted that different fit indices tended to be differentially sensitive to different types and amounts of invariance.

The LR test and changes in fit indices provide methods for comparing nested models. Model selection indices, which are the focus of the next section, allow comparison of both nested and non-nested models.

## Model Selection Indices

We consider four model selection indices (and some of their variants) whose properties have received analytical and empirical evaluation that have been formally proposed for the comparison of either nested or nonnested models (see Table 13.2 for a summary). The general goal of these indices is to select the model with highest generalizability to samples with the same $N$ drawn from the same population. According to Cudeck and Henly (1991), this model should have the smallest expected discrepancy between the fitted model covariance matrix and the population covariance matrix.

### Akaike Information Criterion

The Akaike information criterion (AIC) was proposed by Akaike (1973) to measure the expected discrepancy between the true model and the hypothesized model. The first term of AIC (Table 13.2, Equation T13) is a measure of lack of fit; the second term reflects model complexity, penalizing more complex models. Of importance, the AIC only considers the number of free parameters in determining model complexity. The model with the smallest AIC is selected. Although commonly used, the AIC favors too complex models at small $N$ due to the fact that it fails to take into account the effect of $N$ on model selection. To solve the problem, alternatives to the AIC have been proposed that downweight sample size and therefore may have better performance in these contexts (e.g., bootstrapped information criterion [EIC]; Ishiguro, Sakamoto, & Kitagawa, 1997). For example, Bozdogan (1987) developed a consistent Akaike information criterion (CAIC, Table 13.2, Equation T16). The CAIC performs better than AIC at small $N$ and with a large number of parameters. It does not necessarily favor models with more parameters, unless $N$ is sufficiently large.

### Bayesian Information Criterion

The Bayesian information criterion (BIC) aims to select the model that is most likely to have generated the data in the "Bayesian sense" (Myung & Pitt, 2004; Raftery, 1995). The BIC is in fact a large-sample approximation of the Bayesian model selection procedure that we describe below. The first term in the BIC (see Table 13.2, Equation T17) is the same lack of fit measure used by the AIC. The second term is a measure of model complexity, which is the product of the number of free

**TABLE 13.2. Model Selection Indices for Non-Nested Models**

| Equation No. | Index | Rationale | Measure of model complexity | Pros | Cons |
|---|---|---|---|---|---|
| T15 | $AIC = f + 2k$ | Selects the model that had least expected discrepancy from the true model. | Number of parameters. | Easy to calculate. Performs well at large sample sizes. | Tends to select too complex models. Bad in recovering true model at small sample *N*. |
| T16 | $CAIC = f + [1 + \ln(N)]k$ | Selects the model that had least expected discrepancy from the true model. | Number of parameters and sample size. | Easy to calculate. Performs better than AIC at small sample size and large number of parameters. | |
| T17 | $BIC = f + k\ln(N)$ | Selects the model that is most likely to have generated the data in the Bayesian sense. | Number of parameters and sample size. | Easy to calculate. Performs well under large sample size. | Tends to select a model with too few parameters. Bad in recovering true model at small *N*. |
| T18 | $CV = -\ln\int(y_{val} \mid \hat{\theta}_{cal})$ | Selects the model that has more generalizability to the sample from the population. | Complexity penalty is implicit. | Easy to calculate. More consistent with the implication of generalizability. | Requires sample split. Estimates are often unreliable, especially for small sample size. |
| T19 | $ECVI = f + \frac{2k}{N}$ | Expected value of CV. | Number of parameters and sample size. | Can be calculated using one sample. More consistent with the implication of generalizability. | Assumes multivariate normality. |
| T20 | $BMS = \ln\int_{\Theta} f(y \mid \hat{\theta})\pi(\theta)d\theta$ | Selects the model with the highest mean likelihood of the data over the parameter space. | Number of parameters, sample size, and functional form of a model. | Includes more accurate measure on model complexity. Leads to more accurate model selection. | Computational burden. Hard to specify and calculate for most SEM models. |

*Note.* AIC, Akaike information criterion; BIC, Bayesian information criterion; CV, cross-validation index; BMS, Bayesian model selection; *f*, minimized discrepancy function; *k*, the number of fre parameters of the model; *N*, sample size; ln, natural logarithm; $y_{val}$, the validation sample; $y_{cal}$, the calibration sample; $\hat{\theta}_{cal}$, the parameter values estimated by the calibration sample; $\pi(\theta)$, the prio density of the parameters; $\Theta$, the parameter space.

parameters and the natural logarithm of $N$. Unlike the AIC, the estimation of additional parameters will have a decreasing impact (penalty) in model selection as sample size increases. The model with the smallest BIC is selected. Alternatives to the BIC that may have better performance in some cases have been proposed (e.g., Sclove, 1987).

## Cross-Validation Index

Browne and Cudeck (1993) proposed the cross-validation index (CV) as a means of estimating the generalizability of the estimate of model fit in a new sample from the same population. The CV involves two sequential steps: (1) first fitting a model to a calibration sample ($y_{cal}$) and (2) fitting the same model to a validation sample ($y_{val}$) with the parameter values fixed at those estimated in the first step. The resulting fit in the validation sample estimates the generalizability of the model to a new sample (see Table 13.2, Equation T18). Cudeck and Henly (1991) also noted that the CV is a measure of overall discrepancy between the fitted model and population covariance matrices.

In practice, the calibration and validation samples are often obtained by randomly splitting the observed data into two subsamples of equal size, which becomes impractical when the available sample is small. Browne and Cudeck (1989, 1993) proposed the expected cross-validation index (ECVI; see Table 13.2, Equation T19) based on a single sample under the assumption of multivariate normality. Conceptually, one can interpret the ECVI as the average discrepancy in the fitted covariance matrices between two samples of equal sample size across all possible combinations of two samples from the same population. Because it considers all possible combinations, it is expected to give more stable estimates than the CV. However, the ECVI can provide misleading information about model selection when the multivariate normality assumption is severely violated.

## Bayesian Model Selection

Bayesian model selection (BMS), an approach developed in the Bayesian statistical framework, is theoretically useful but difficult to implement in many contexts (Pitt, Kim, & Myung, 2003; Wu, Myung, & Batchelder, 2010). BMS attempts to select the model with the highest mean likelihood of producing the data. To achieve this goal, BMS takes the logarithm of the mean likelihood, averaged across the full range of parameter values and weighted by the prior density (Table 13.2, Equation T20). BMS assumes that there exists (1) a true known probability distribution (prior density[4]) from which the data were sampled and (2) a known parameter space that represents the potential values that each of the parameters may take on. BMS represents the Bayesian posterior probability of the model being correct given the data. BMS is potentially of particular value for comparing models that have different functional forms, but which have the same number of freely estimated parameters. In the context of SEM, an example in which this would occur is the comparison of a latent interaction model, $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1\xi_2$, with a latent quadratic model, $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1^2$. To date, BMS has not been implemented in SEM, except in contexts involving the analysis of correlation rather than covariance structures (see Preacher, 2006, for an example).

## Comparison of the Model Selection Indices

Table 13.2 compares six model fit indices for non-nested models. The AIC, CAIC, and BIC all measure model complexity using a number of parameters. The CAIC and BIC also include a weight for sample size, whereas the AIC does not, so that the CAIC and BIC tend to select simpler models than does the AIC at smaller $N$. BMS considers functional form in addition to the number of free parameters and sample size. Therefore, BMS is expected to perform more accurately in model selection than the AIC, CAIC, and BIC when the competing models have different functional forms. Given the pervasiveness of linear functional forms in SEM, this feature would only rarely be an advantage in practice. The chief disadvantage of BMS is the need to specify a known probability distribution and parameter space for the model, which is very difficult in practice. These can be specified for some well-defined problems, but even then the computational burden can be enormous. The CV implicitly builds model complexity into its calculation procedure. The definition of the CV also seems more consistent with the implication of generalizability. However, it often leads to an unreliable estimate of generalizability in small samples. ECVI provides more stable estimates than the CV but assumes multivariate normality. Wicherts and Dolan (2004) noted that the ECVI has a linear relationship with the AIC. Thus, it leads to the same rank-ordering of competing models as the AIC.

In contrast to fit indices, *N* plays an important role in choosing among competing models. More complex models are often preferred for large *N*s, in the sense that they are more likely to replicate in a new sample, even though a simpler model might provide a better approximation to the population. In contrast, for small *N*s, simpler models are typically preferred because there are not enough data to support the estimation of parameters with sufficient precision in a more complex model (MacCallum, 2003). Cudeck and Henly (1991) argued that this effect of sample size should be deemed not as undesirable but as *fundamental* to any statistical decision. It is important to choose the model that performs best in practice given the specified sample size.

## SENSITIVITY OF KEY MODEL PARAMETERS

Beyond the consideration of fit, researchers are concerned about producing unbiased and precise estimates of key parameters in their hypothesized models. In some models, such as the Fishbein–Azjen model presented in Figure 13.1B, researchers may consider each of the paths to be equally important. In other models, such as the growth model presented in Figure 13.1C, researchers may believe that some parameters (e.g., the mean intercept and slope) are key parameters, whereas other parameters, such as the values of autocovariances between the Level 1 residuals, are only of importance to the extent they may produce more accurate estimates of the values of the key parameters or their standard errors (Kwok, West, & Green, 2007).

Several approaches have been proposed for probing the sensitivity of the estimates of the model parameters to changes in the model. Saris and colleagues (2009) proposed focusing on key model parameters. They suggest identifying a value that would reflect a meaningful change in that key parameter, and then conducting a series of simulation studies in which the values of other model parameters are varied within plausible ranges. This approach can be used to investigate each key parameter separately in turn, but it has not yet been extended to permit the simultaneous investigation of multiple parameters as a function of changes in other parts of the model. Millsap (in press) has extended this approach to permit the examination of alternative models that fall within a small specified range of the target model on the RMSEA. An alternative approach has been proposed by MacCallum, Lee, and Browne (in press) that more easily allows for the examination of the joint sensitivity of multiple parameter estimates. They propose allowing a small, nonconsequential increase of *F* from its minimum *f*, followed by an examination of the range of values of the key parameters that are permissible given this slight reduction in fit. They find that all potential parameter estimates that satisfy the criterion will fall within an ellipsoid, with one dimension for each key parameter. Analysts can choose to consider dimensions corresponding to two or three of the key parameters simultaneously to permit visualization of the acceptable parameter space. In some cases, the range of parameter estimates will be reasonable and little difference in the conclusions of the model could result. In other cases, the range of potentially acceptable parameter estimates will be large, even permitting analysts to conclude that the direction of the parameter estimates is uncertain, providing the researcher with little confidence in the conclusions based on fitting the model.

## DISTINGUISHING BETWEEN EQUIVALENT MODELS

Even when a model fits the data well, other equivalent models that fit the data equally well typically exist. Figure 13.3 presents three path models each having 1 *df* that provide an equally good fit to a 3 × 3 covariance structure. These models are data equivalent (see Williams, Chapter 15, this volume) but have distinctly different substantive interpretations since the directions of the paths vary. Figure 13.3A presents a model depicting the full mediation of an effect of *X* through *M*, which in turn influences *Y*. Figure 13.3B presents a model of the reverse causal effect, where *Y* affects *M*, which in turn affects *X*. Finally, Figure 13.3C presents a model in which *M* is a common cause of both *X* and *Y*. These models cannot be distinguished with cross-sectional data. Batchelder and Riefer (1999) proposed the use of model validation to distinguish between models. In the present example of a mediational model, developing manipulations that separately target the $X \rightarrow M$ path and the $M \rightarrow Y$ path could provide experimental data that would help the researcher to distinguish between the three models (Spencer, Zanna, & Fong, 2005). Alternatively, a longitudinal panel study (Cole & Maxwell, 2003) in which *X*, *M*, and *Y* were measured at Times 1, 2, and 3 could provide evidence that allows the researcher to establish the temporal precedence of the effects, thereby helping to rule out the alterna-

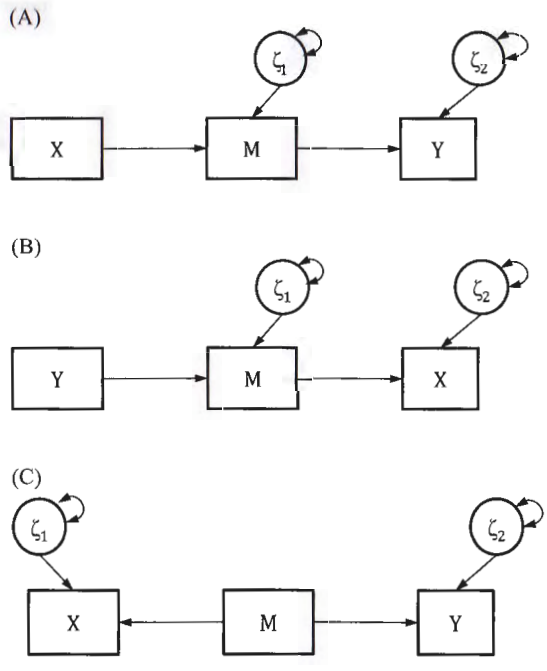


**FIGURE 13.3.** Three data-equivalent path models with cross-sectional data. (A) Mediational model: *X* causes *M*, which in turn causes *Y*. (B) Mediational model: *Y* causes *M*, which in turn causes *X*. (C) Common cause model: *M* causes both *X* and *Y*.

tive models. Conceivably, model validation strategies could even be extended to CFA models, for example, by developing a manipulation expected to affect (1) only the first factor ($\eta_1$) but not the second factor ($\eta_2$), or (2) only the second factor but not the first factor, as a means of helping to clarify the measurement structure (see Figure 13.1A). Although such strategies have been used widely in other areas (e.g., multinomial tree models), they have not seen much usage with structural equation models.

## PARSIMONY REVISITED

The focus on achieving accurate estimates of key parameters raises another issue, the value of parsimony. Other things being equal, science clearly prefers (1) models with fewer parameters and (2) models that make more precise estimates, so that they place more restrictions on the range of data structures they will fit (Preacher, 2006). The value of parsimony is clearly evident when models involving exploratory components are being compared. On the other hand, other things are not always equal. Cole, Ciesla, and Steiger (2007) have shown that key parameters in structural equation models can be seriously misestimated if theoretically justified residual correlations are not included in the model, even though minimal effects on overall model fit are observed. Marsh and Hau (1996) have shown that the failure to estimate relatively small correlated residuals in longitudinal confirmatory factor analysis models could have effects on estimates of the key stability parameters, even though model fit is little affected. In part for this reason, Bentler (1992) has argued that the assessment of fit and parsimony may often best be kept separate. This strategy lessens the possibility of deleting model parameters that provide important correction for artifacts that may plague the model, a practice that leads to distortion of key model parameters.

## SUMMARY AND CONCLUSION

The assessment of fit provides researchers with an overall perspective on how well the theoretical model is able to reproduce the observed data. The $\chi^2$ test statistic provides a statistical test of whether the residuals between the model-implied values and actual data are greater than would be expected on the basis of sampling error, assuming adequate sample size and multivariate normality. However, given that the hypothesized model is an approximation to the unknown true model, and that the $\chi^2$ test statistic is affected by sample size, researchers have sought to develop alternative practical fit indices that provide measures of fit that are not related to sample size. A large number of these fit indices have been proposed, and the properties of several of the more widely used indices are presented in Table 13.1. Several of these (e.g., CFI/RNI, RMSEA, TLI, SRMR, GFI*) have desirable properties, and their estimates are *not* related to sample size. However, the quest for a standard cutoff criterion for each of the fit indices has proven to be elusive. Fit indices are affected by other model properties, such as the number of indicators and magnitudes of factor loadings. As fit indices are applied beyond CFA models to more complex models—multiple group models, multilevel models, growth models, and so forth—this quest for a single, standard cutoff cri-

terion becomes increasingly chimerical and alternative strategies are needed.

One alternative strategy involves separate examination of the fit of each part of the model, for example, in multilevel models and growth curve models. This strategy can be extended even down to examination of the reproduction of the individual means or covariances to identify problem spots in the model. A second strategy involves hypothesizing multiple competing models and then using model selection indices to identify the optimal model. A third strategy involves close examination of key parameter estimates and their sensitivity to other aspects of the model. A fourth strategy involves the use of model validation procedures that can potentially help researchers distinguish between data-equivalent models or even models that produce similar values of practical fit indices. Nearly 20 years ago Bollen and Long (1993) wrote, "The test statistics and fit indices are very beneficial, but they are no replacement for sound judgment and substantive expertise" (p. 8). This advice remains true today, but sound judgment is now aided by several alternative strategies that provide supplemental information on the adequacy of the hypothesized model in accounting for the observed data.

## ACKNOWLEDGMENTS

Stephen G. West was supported by a Forschungspreis from the Alexander von Humboldt Foundation during the writing of this chapter. We thank Peter Bentler, Patrick Curran, Edgar Erdfelder, Robert MacCallum, Roger Millsap, Victoria Savalei, editor Rick Hoyle, and an anonymous reviewer for comments on an earlier version of this chapter.

## NOTES

1. A value of $\chi^2/df < 1$ can occur due to sampling variation, particularly in small samples.
2. Another difference between the GFI and $R^2$ is that structural equation models attempt to reproduce observed covariances, so the GFI is based on $p^*$ residual covariances, whereas OLS regression attempts to reproduce observed scores on a dependent variable, so $R^2$ is based on $N$ residual dependent variable scores.
3. In the case of the GFI, the GFI* appears to have the desired proportion of variance interpretation without the bias of underestimating the true value at small sample sizes.
4. The prior density of the parameter is the probability distribution over the parameter space, prior to seeing the data. The prior density represents the researcher's prior belief or prior assumptions about the probabilities of different parameter values.

## REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411–423.

Balderjahn, I. (1988). A note to Bollen's alternative fit measure. *Psychometrika, 53*, 283–285.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review, 6*, 57–86.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika, 48*, 493–517.

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

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the *Bulletin*. *Psychological Bulletin, 112*, 400–404.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods, 15*, 111–123.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika, 51*, 375–377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*, 303–316.

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

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I, pp. 149–173). Amsterdam: North Holland.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building, In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8*, 1–24.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. A. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum Press.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445–455.

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.

Browne, M. W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7*, 403–421.

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 and Research, 36*, 462–494.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12*, 381–398.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*, 512–519.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods and Research, 32*, 208–252.

Davey, A., Savla, J., & Luo, Z. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling, 12*, 578–597.

Diaconis, P. (1985). Theories of data analysis. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Ed.), *Exploring data tables, trends, and shapes* (pp. 1–36). New York: Wiley.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indices in assessing mean structure invariance. *Structural Equation Modeling, 16*, 54–69.

Gallini, J. K., & Mandeville, G. K. (1984). An investigation of the effects of sample size and specification error on the fit of structural equation models. *Journal of Experimental Education, 53*, 9–19.

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. *Multivariate Behavioral Research, 28*, 205–214.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

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.

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics, 49*, 411–434.

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

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago: International Educational Services.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*, 237–155.

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*, 333–351.

Klein, A. G., & Schermelleh-Engel, K. (2010). A measure for detecting poor fit due to omitted nonlinear terms in SEM. *Advances in Statistical Analysis, 94*, 157–166.

Kwok, O.-M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42*, 557–592.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*, 107–120.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*, 113–139.

MacCallum, R. C., Lee, T., & Browne, M. W. (in press). Fungible parameter estimates in latent growth curve models.

In M. Edwards & R. C. MacCallum (Eds.), *Current issues in the theory and application of latent variable models.* New York: Routledge.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611–637.

Maiti, S. S., & Mukherjee, B. N. (1990). A note on the distributional properties of the Jöreskog–Sörbom fit indices. *Psychometrika, 55,* 721–726.

Marsh, H. W. (1995). The Δ2 and $\chi^2$ 12 fit indices for structural equation models: A brief note of clarification. *Structural Equation Modeling, 2,* 246–254.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Mahwah, NJ: Erlbaum.

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.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, E. (1998). Is more ever too much?: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005), Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275–340). Mahwah, NJ: Erlbaum.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675–686.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *American Statistician, 54,* 17–24.

Millsap, R. E. (in press). A simulation paradigm for evaluating approximate fit. In M. Edwards & R. C. MacCallum (Eds.), *Current issues in the theory and application of latent variable models.* New York: Routledge.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

Myung, J. I., & Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology, 383,* 351–366.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin and Review, 10,* 29–44.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227–259.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–196.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16,* 583–601.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16,* 561–582.

Savalei, V. (2011). *The relationship between RMSEA and model misspecification in CFA models.* Unpublished manuscript, Psychology Department, University of British Columbia, Vancouver, BC, Canada.

Sclove, L. S. (1987). Application of model-selection criteria for some problems in multivariate analysis. *Psychometrika, 52,* 333–343.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152–178). San Francisco: Jossey-Bass.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89,* 845–851.

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–264.

Sugawara, H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement, 17*, 365–377.

Tanaka, J. S. (1993). Multifaceted conceptions of fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38*, 197–201.

Taylor, A. B. (2008). *Two new methods of studying the performance of SEM fit indices.* Doctoral dissertation, Arizona State University, Tempe, AZ.

Tucker, L., Koopman, R., & Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*, 421–459.

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

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8*, 84–136.

Wherry, R. J., Sr. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*, 440–457.

Wicherts, J. M., & Dolan, C. V. (2004). A cautionary note on the use of information fit indexes in covariance structure modeling with means. *Structural Equation Modeling, 11*, 45–50.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*, 16–37.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology, 54*, 291–303.

Wu, W., & West, S. G. (2010). Sensitivity of SEM fit indices to misspecifications in growth curve models: A simulation study. *Multivariate Behavioral Research, 45*, 420–452.

Wu, W., West, S. G., & Taylor, A. B. (2009). Growth curve modeling: Evaluating model fit and model selection. *Psychological Methods, 14*, 183–201.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*, 115–148.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology, 37*, 53–82.