# Assessing Goodness of Fit in Confirmatory Factor Analysis

Jun Sun

*The author identifies 3 main purposes of conducting confirmatory factor analysis (CFA), and their different requirements on goodness-of-fit assessment. For a better understanding of fit indices, he proposes a hierarchical classification scheme based on J. S. Tanaka's (1993) multifaceted conceptions and discusses how to assess goodness of fit for different purposes in CFA.*

In counseling and education, researchers often use confirmatory factor analysis (CFA) to evaluate and compare the hypothesized factor structure of scores obtained from various measurement instruments. Dozens of fit indices are available to assess goodness of fit in CFA, but only a portion have been frequently used. A survey of the articles published in *Measurement and Evaluation in Counseling and Development (MECD)* from 1996 to 2002 revealed at least 16 articles in which authors had conducted CFA for three purposes, and they reported using 12 different fit indices. These three purposes were construct validity evaluation, response pattern comparison, and competing model comparison. Table 1 categorizes these articles according to the main purposes for conducting CFA and lists the reported fit indices in the note.

Because no single fit index assesses every aspect of goodness of fit (Thompson, 2000), all the aforementioned articles reported the use of multiple fit indices. Some fit indices (e.g., comparative fit index [CFI]) had been used more frequently than others (e.g., Akaike information criterion [AIC]). To the best of my knowledge, however, there were no structured criteria, either stated in these articles or available somewhere else, regarding how to select appropriate fit indices for different purposes. Thus, the selection of fit indices seemed to be arbitrary. As pointed out by Byrne (1998), "Assessment of model adequacy must be based on multiple criteria that take into account theoretical, statistical, and practical considerations" (p. 119). Failure to do so may lead to one or more of the following:

- Incomplete picture of goodness of fit
- Selection of indices based on value, not on theory
- Difficulty for others to cross-validate the result due to some undesirable characteristics of reported fit indices (e.g., sensitivity to sample size)

To take multiple criteria for selecting fit indices into account simultaneously, it may help to classify fit indices in a structured way to facilitate the understanding and comparison of them. In an attempt to accomplish this, I propose a hierarchical classification scheme of fit indices based on Tanaka's (1993) multifaceted conceptions and discuss the use of fit indices in CFA.

This article is organized as follows: The first section discusses the three purposes of conducting CFA and their different requirements on the goodness-of-fit assessment. The second section describes a hierarchical classification scheme organized according to how fit

*Jun Sun, Department of Information and Operations Management, Mays Business School, Texas A&M University.* *The author thanks Bruce Thompson and Victor L. Willson for their guidance.* *Correspondence concerning this article should be addressed to Jun Sun, Department of Information and Operations Management, Mays Business School, Texas A&M University, College Station, TX 77843-4217 (e-mail: john_sun@tamu.edu).*

## TABLE 1

**Example Studies That Conducted Confirmatory Factor Analysis for Different Purposes (Published in *Measurement and Evaluation in Counseling and Development* from 1996 to 2002)**

| CFA Purpose | Example Studies[a] |
|---|---|
| Construct validity evaluation | Cokley & Helm (2001); Kirisci & Moss (1997); McCoach (2002) |
| Response pattern comparison | Chan & Lin (1996); Denzine & Kowaliski (2002); O'Rourke & Cappeliez (2001); Rogers, Abbey-Hines, & Rando (1997); Rogers & Hanlon (1996); Slaney, Rice, Mobley, Trippi, & Ashby (2001) |
| Competing model comparison | Beasley, Long, & Natali (2001); Cockley, Bernard, Cunningham, & Motoike (2001); Erford, Peyrot, & Siska (1998); Fuertes, Miville, Mohr, Sedlacek, & Gretchen (2000); Loo (2001); Thomas, Donnell, & Buboltz (2001); Utsey (1999) |

*Note.* CFA = confirmatory factor analysis.

[a]The following fit indices were used in studies that conducted CFA and were published in *Measurement and Evaluation in Counseling and Development* from 1996 to 2002: $\chi^2$, $\chi^2/df$, Akaike information criterion, goodness-of-fit index, adjusted goodness-of-fit index, normed fit index, Tucker-Lewis index (non-normed fit index), comparative fit index, relative noncentrality index, root mean square error of approximation, root mean square residual, and standardized root mean square residual.

indices are constructed, and the next section categorizes fit indices based on this scheme. The fourth section discusses the characteristics of fit indices, and the last section discusses how to select and use fit indices for different purposes.

## MAIN PURPOSES FOR CONDUCTING CFA

### Construct Validity Evaluation

When developing an instrument to evaluate some psychometric constructs of individuals in certain populations, it is important to validate the evaluation by checking the interpretation of obtained scores. A measurement model specifies a statistical way to interpret scores, and each factor, measured by multiple indicators, represents a construct. By fitting scores to the factor structure of a measurement model, CFA can evaluate two aspects of construct validity: *discriminant validity* and *convergent validity*. In CFA, discriminant validity refers to "the distinctiveness of the factors measured by different sets of indicators" (Kline, 1998, p. 60), and convergent validity refers to the cohesiveness of a set of indicators in measuring their underlying factor (rather than something else). With a given estimated model, there is evidence for discriminant validity if different factors are not excessively correlated with each other (e.g., > 0.85), and there is evidence for convergent validity if a set of indicators all have relatively high structure coefficients with the factor that they are specified to measure (Kline, 1998).

In counseling and education, therefore, CFA is often used to evaluate the construct validity of scores from measurement instruments. For example, McCoach (2002) used CFA to examine how well the School Attitude Assessment Survey evaluated four dimensions of school attitudes. Cokley and Helm (2001) conducted CFA to check the construct validity of scores from the Multidimensional Inventory of Black Identity (Sellers, Rowley, Chavous, Shelton, & Smith, 1997). This type of study features stand-alone evaluation of model fit in which fit indices indicate how well the scores from a single group fit the theoretical factor structure as evidence of construct validity.

### Response Pattern Comparison

When researchers use measurement instruments to collect scores from individuals in various populations or various segments of a population, they may be concerned about whether the response

patterns of different groups vary significantly from each other. Therefore, CFA is often used to test the invariance of the same factor structure and parameter estimates of a measurement model across multiple groups. Such information about the suitability and stability of score interpretation can help researchers and professionals make better decisions on instrument usage and result inference. For example, Slaney, Rice, Mobley, Trippi, and Ashby (2001) used CFA to examine the sample difference in responses to a perfectionism scale across groups from different regions. O'Rourke and Cappeliez (2001) used CFA to test whether the male and female participants responded differently to marital satisfaction and aggrandizement scales. To cross-validate Good et al.'s (1995) study on their Gender Role Conflict Scale, Rogers, Abbey-Hines, and Rando (1997) collected scores from other groups and compared their CFA result with those obtained by Good et al.

A study that uses CFA to compare the response patterns of different groups may require as many as three steps (Byrne, 2001). The first step is to test for invariant factor structure across multiple groups. If scores from different groups fit the same hypothesized factor structure reasonably well, the next step is to test for invariant model parameters across these groups. If the hypothesis of invariant parameters is rejected, the third step is to identify which parameters are variant. The last two steps usually involve placing and releasing some constraints on certain parameters in a measurement model and testing whether such modifications cause significant change in model fit. A model with constraints on certain parameters (e.g., set parameters to zero or non-zero values, or to be invariant across groups) is referred to as *nested* in the original model. Therefore, studies that compare response patterns often not only examine the overall goodness of fit of a certain factor structure across multiple groups but also compare the goodness of fit between two models with one nested in the other.

## Competing Model Comparison

For some measurement instruments, there can be different score interpretations regarding the underlying factor structure. Researchers often conduct CFA to compare competing measurement models in order to identify the most appropriate score interpretations under certain circumstances. For example, Cokley, Bernard, Cunningham, and Motoike (2001) compared 1-, 2-, 3-, 5-, and 7-factor models of the Academic Motivation Scale (Vallerand et al., 1992) based on different theories, and their results favored the 7-factor model. Beasley, Long, and Natali (2001) compared 1-, 2-, and 4-factor models of the Mathematics Anxiety Scale for Children (Chiu & Henry,1990), and the unidimensional model was supported in their study.

In some cases, researchers are interested in comparing models that have different sets of indicators. For example, Thomas, Donnell, and Buboltz (2001) used CFA to compare two candidate models underlying the Hong Psychological Reactance Scale (Hong & Faedda, 1996), which had the same number of factors but a different number of indicators. Models that share the same set of indicators are usually easier to compare regarding goodness of fit than are models that have different sets of indicators.

## CFA Purposes and Fit Index Use

In summary, the assessment of goodness of fit in a CFA study usually involves one factor structure and one group for construct validity evaluation, one factor structure but multiple groups for response pattern comparison, and different factor structures for competing model comparison. Fit indices assess goodness of fit through certain types of comparison; however, each type of comparison is somewhat unique in its assumption, baseline, and emphasis. Moreover, fit indices vary in their appropriateness for specific circumstances in which goodness of fit is assessed. Therefore, different fit indices may be suitable for different purposes and circumstances, and the choice of fit indices may sometimes affect the interpretation of results.

Two issues may be relevant to the use of fit indices:

1. How is each type of fit index constructed? This issue is related to specific aspects of goodness of fit regarding the assumption, baseline, and emphasis that each type of fit index is supposed to assess.
2. What are the characteristics of each fit index? This issue is related to the comparison of fit indices regarding their appropriateness under certain circumstances for assessing goodness of fit.

The way in which fit indices are constructed "determines" their characteristics to some extent. Therefore, a clear and systematic classification scheme of fit indices based on their formation can be helpful in enabling a better understanding, selection, and interpretation of them.

## A HIERARCHICAL CLASSIFICATION SCHEME OF FIT INDICES

The difficulty in comparing and selecting fit indices is due, to a large extent, to their multifaceted nature. Tanaka (1993) identified six dimensions that categorize fit indices dichotomously: (1) population-based versus sample-based, (2) simplicity versus complexity, (3) normed versus nonnormed, (4) absolute versus relative, (5) estimation method free versus estimation method specific, and (6) sample size independent versus sample size dependent. Among these dimensions, three (Dimensions 1, 2, and 4) are closely related to how fit indices are constructed, and the others (Dimensions 3, 5, and 6) concern some of the characteristics of fit indices. These dimensions are not totally independent from each other because the characteristics of fit indices are more or less related to their structure.

Tanaka's (1993) conceptions allow an easy comparison of fit indices along one dimension at a time. However, selecting fit indices for certain purposes often requires comparing them along multiple dimensions simultaneously, which can be very complicated. Tanaka did not provide clear guidance on this issue, nor have other researchers. A better strategy is to classify fit indices hierarchically through different levels according to how they are constructed. Using the three relevant dimensions in Tanaka's conceptions as a basis, this article classifies fit indices into different categories through three levels: discrepancy assumption level, model involvement level, and complexity adjustment level. These levels form a decision tree that classifies fit indices in a mutually exclusive way, which facilitates the discussion of the characteristics of each type. Also, such a classification scheme clarifies different aspects of goodness of fit that each index is supposed to assess and, to some degree, prevents the arbitrary selection of fit indices. Therefore, this classification scheme may facilitate the interpretation, comparison, and selection of fit indices.

To understand how fit indices are constructed, a basic knowledge about model estimation in CFA is helpful. The free parameters of measurement models are estimated by fitting scores to the factor structure with certain estimation methods. Commonly used estimation methods include maximum likelihood (ML), generalized least squares (GLS), and asymptotically distribution-free (ADF). These methods estimate the free parameters of measurement models by minimizing a discrepancy function between the sample covariance matrix and the covariance matrix reproduced from the hypothesized model, $F(\mathbf{S}; \Sigma(\theta))$. Different estimation methods use different discrepancy functions and yield different values of fit indices.

The number of distinct items in the sample covariance matrix can be regarded as the total degrees of freedom for model estimation. If the number of indicators included in a measurement model is $p$, the total degrees of freedom is

$$p^* = p(p + 1)/2. \quad (1)$$

The estimation of a free parameter "consumes" one degree of freedom. Therefore, the available degrees of freedom for estimating a model, or model *df,* is the difference between the total degrees of freedom, $p^*$, and the number of free parameters. When a measurement model has more parameters to be estimated, it has a smaller *df*. Thus, a model *df* indicates the complexity of a measurement model for a fixed number of indicators: A larger *df* indicates a simpler model, and a smaller *df* indicates a more complex model. When a model *df* is less than zero, the structural equations corresponding to the model are not solvable. When a model *df* is equal to zero, the model is *just-identified,* and there is at most one solution for parameter estimation. If there is one, the reproduced covariance matrix must be identical to the sample covariance matrix. Because all possible parameters are freed to be estimated, such a measurement model is *saturated* and has little theoretical value in CFA. When a model *df* is greater than zero, the model is *over-identified,* and there may be an infinite number of solutions. Most estimation methods, including ML, GLS, and ADF, determine the optimal solution through minimizing a discrepancy function. It should be noted that for a given set of data, there can be different models that result in the same level of goodness of fit (MacCallum, Wegener, Uchino, & Fabrigar, 1993). Therefore, researchers should conduct CFA on predetermined measurement models derived from theories and/or empirical evidence, rather than in an exploratory way.

The minimized value of discrepancy functions for ML, GLS, and ADF can be denoted as $F_{min}$. Under a set of standard assumptions, $(N-1) \times F_{min}$ follows a central chi-square distribution, with the degrees of freedom equal to the model *df,* and this statistic is usually called the chi-square statistic ($\chi^2$). The expected value of the chi-square statistic is equal to the model *df,* that is, $E(\chi^2/df) = 1$. A key assumption is that the model is correctly specified, and thus the reproduced covariance matrix is assumed to be identical to the true population covariance matrix, that is, $\Sigma(\theta) = \Sigma$. However, researchers (e.g., Cudeck & Henly, 1991) questioned this assumption and argued that almost all hypothesized models are more or less misspecified, and thus two matrices cannot be identical, that is, $\Sigma(\theta) \neq \Sigma$.

Steiger, Shapiro, and Browne (1985) showed that given certain standard assumptions, the chi-square statistic of a slightly or moderately misspecified model has a noncentral chi-square distribution, with the noncentrality parameter lambda and the model *df* as the degrees of freedom. The expected value of the chi-square statistic is equal to the sum of the model *df* and lambda, that is, $E(\chi^2) = df + \lambda$. Because $E(\chi^2 - df) = \lambda$, the sample estimate of lambda is

$$l = \chi^2 - df. \tag{2}$$

Almost all fit indices evaluate model fit by assuming either (a) that a hypothesized model is correctly specified so that there is no discrepancy between the reproduced covariance matrix and the true population covariance matrix (i.e., $\Sigma(\theta) = \Sigma$) or (b) that a hypothesized model is unavoidably misspecified so that there is discrepancy between two matrices (i.e., $\Sigma(\theta) \neq \Sigma$). Therefore, at the so-called discrepancy assumption level, fit indices can be divided into two general categories, the sample-based and the population-based, consistent with Tanaka's (1993) conceptions (Dimension 1). Sample-based fit indices are based on the observed difference between the reproduced covariance matrix and the sample covariance matrix, the former assumed to be the same with the true population covariance matrix to test model adequacy. Population-based fit indices are based on the estimated difference between the reproduced covariance matrix and the unknown population covariance matrix.

Fit indices can be further divided into absolute and relative ones, corresponding to Tanaka's (1993) fourth dimension, according to whether or not their calculation involves comparison with another model. Thus, this second level can be termed *model involvement level.* The calculation of absolute indices only involves the hypothesized model, whereas that of relative indices involves another baseline model. The most common baseline model is the independence model, with all pattern coefficients among indicators set to be zero. The *df* of an independence model

is equal to $p^* - p$, because there are $p$ variances to be estimated, one for each indicator. The standard baseline model may not be optimal in all cases, and Widaman and Thompson (2003) have discussed how to specify better baseline models for relative fit indices. However, such discussions are beyond the scope of this article, and unless some other baseline models are specified, all relative fit indices discussed here use the independence models of the original as their baseline models.

Some fit indices yield better results when there are more free parameters to be estimated, whereas others impose some penalty on model complexity. Corresponding to the second dimension in Tanaka's (1993) conceptions, the third level (complexity adjustment level) classifies fit indices on whether or not they are adjusted for model complexity. There are different ways of adjustment, such as linearly combining the chi-square statistic with a weighted model *df*, dividing the chi-square statistic with the model *df*, and multiplying a relative fit index with a parsimony index.

Therefore, these three levels classify fit indices according to how they are constructed: (a) discrepancy assumption level, whether or not a fit index is based on the assumption that $\Sigma(\theta) = \Sigma$; (b) model involvement level, whether or not a fit index involves another baseline model; and (c) complexity adjustment level, whether or not a fit index is adjusted for model complexity. Of these levels, the discrepancy assumption level is the most fundamental, whereas the complexity adjustment level is the least fundamental. As shown in Figure 1, fit indices can be classified hierarchically into different types according to these levels, forming a decision tree.

## STRUCTURE OF FIT INDICES

### Sample-Based Absolute Fit Indices

According to the hierarchical classification scheme, root mean square residual (RMSR), standardized root mean square residual (SRMR), the chi-square statistic, $\chi^2/df$, AIC, goodness-of-fit index (GFI), and adjusted goodness-of-fit index (AGFI) are sample-based absolute fit indices. Among them, RMSR and SRMR are not based on the chi-square statistic. RMSR (Jöreskog & Sörbom, 1981) is the square root of the average squared difference between the reproduced and sample covariance matrices. SRMR (Bentler, 1995) is the standardized version of RMSR. When the reproduced and sample covariance matrices are identical, the values of both RMSR and SMSR are equal to zero, indicating a perfect fit.

As mentioned, the value of the chi-square statistic is equal to $(N-1) \times F_{min}$. Its probability value can be used to test the hypotheses $H_0$: $\Sigma(\theta) = \Sigma$ versus $H_a$: $\Sigma(\theta) \neq \Sigma$. The null hypothesis indicates a perfect fit, and it is desirable not to reject it. Because a statistical test is available, the chi-square statistic seems to be ideal for the evaluation of goodness of fit. Unfortunately, there are two major drawbacks of the chi-square statistic: It is sensitive to sample size and favors complex models. First, the calculation of the chi-square statistic depends explicitly on the sample size *(N)*. When the sample size is large, even a slight difference between the reproduced and sample covariance matrices can be magnified to be significant. As Bentler and Bonett (1980) pointed out, when the sample size is large enough, the null hypothesis, $\Sigma(\theta) = \Sigma$, will always be rejected. Because of that, Jöreskog (1969) pointed out that the chi-square statistic is more a descriptive index of fit than a statistical test. The chi-square statistic, however, can be useful to compare different measurement models for a given data set, especially when a model is nested in another (Jöreskog & Sörbom, 1989). Therefore, when comparing a constrained model with its original model, as is often the case in response pattern comparison, the chi-square test can be used to determine whether the constraints result in significantly different goodness of fit.

The second drawback of the chi-square statistic is that it favors complex models, that is, its value always decreases when more parameters are freed for estimation until the model becomes saturated. There are two ways to adjust the chi-square statistic for model complex-

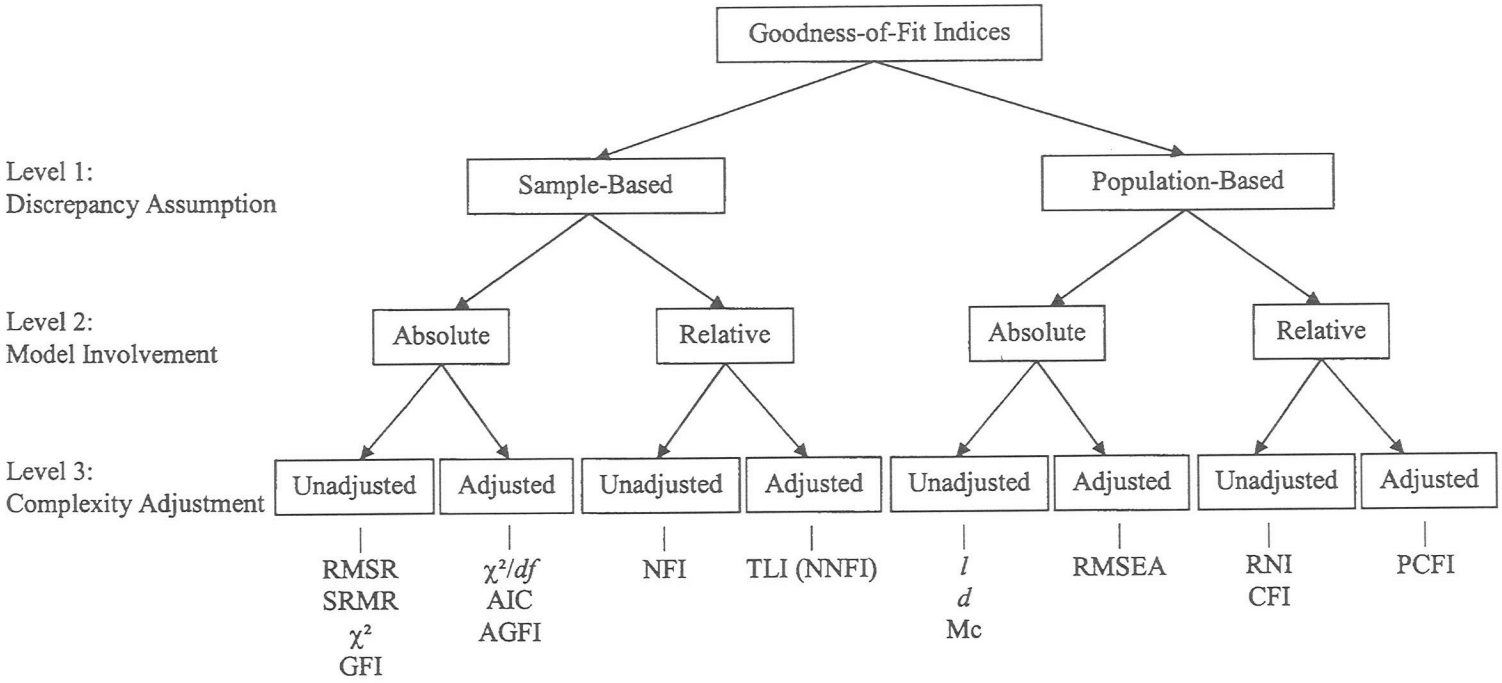


**FIGURE 1**

**A Hierarchical Classification of Goodness-of-Fit Indices**

*Note.* RMSR = root mean square residual; SRMR = standardized root mean square residual; GFI = goodness-of-fit index; AIC = Akaike information criterion; AGFI = adjusted goodness-of-fit index; NFI = normed fit index; TLI = Tucker-Lewis index; NNFI = non-normed fit index; *l* = the estimate of noncentrality parameter; *d* = the estimate of minimized population discrepancy function; Mc = McDonald's centrality index; RMSEA = root mean square error of approximation; RNI = relative noncentrality index; CFI = comparative fit index; PCFI = parsimony comparative fit index.

ity. First, researchers have provided several composite fit indices by linearly combining the chi-square statistic with a weighted model *df*, that is, $\chi^2 + k \times df$, where $k$ is a constant (usually negative). Because these fit indices are adjusted for model complexity directly, they are particularly useful in comparing models of different complexity. The simplest composite fit index is AIC (Akaike, 1973), expressed as

$$\text{AIC} = \chi^2 - 2df. \tag{3}$$

Another way to adjust the chi-square statistic for model complexity is to divide it by the model *df*. Because $E(\chi^2/df) = 1$ for a "true" model, then the value of $\chi^2/df$ being close to 1 indicates a good model.

As an alternative to the chi-square statistic that depends on sample size explicitly, Jöreskog and Sörbom (1984) proposed the GFI,

$$\text{GFI} = 1 - \chi^2/\min[F(\mathbf{S}; \Sigma(\mathbf{0}))], \tag{4}$$

where $\min[F(\mathbf{S}; \Sigma(\mathbf{0}))]$ is the minimum value of the discrepancy function with all elements in the population covariance matrix assumed to be zero, that is, $\mathbf{S} = \mathbf{0}$, as if there is no model at all. Although in the seemingly relative form, GFI is still an absolute fit index because it does not involve any model other than the hypothesized one. Because there is an exact monotonic relationship between GFI and the chi-square statistic (Maiti & Mukherjee, 1990), GFI can be regarded as the normed chi-square statistic, with the value ranging from zero to 1. Like the chi-square statistic, the value of GFI can always be improved by freeing more parameters. To adjust GFI for model complexity, Jöreskog and Sörbom (1984) proposed the AGFI,

$$\text{AGFI} = 1 - (1 - \text{GFI})\, p^*/df. \tag{5}$$

Although the calculation of GFI does not involve the sample size explicitly, its sampling distribution is still subject to the sample size (Jöreskog, 1993). Therefore, the value of both GFI and AGFI are affected by sample size.

## Sample-Based Relative Fit Indices

Relative fit indices compare a hypothesized model with a baseline model (usually the corresponding independence model) with the chi-square statistic or related statistics. Here are some notations: $\chi_1^2$ and $\chi_0^2$ denote the chi-square statistic of the hypothesized model and the independence model respectively; $df_1$ and $df_0$ denote the model *df* of the hypothesized model and the independence model, respectively. According to the hierarchical classification scheme, the normed fit index (NFI; Bentler & Bonett, 1980) and Tucker-Lewis index (TLI; Tucker & Lewis, 1973), also known as Bentler and Bonett's (1980) non-normed fit index (NNFI), are sample-based relative fit indices. The NFI, given by

$$\text{NFI} = 1 - \chi_1^2/\chi_0^2, \tag{6}$$

is the simplest relative fit index. It indicates how much a model improves the goodness of fit from the independence model by directly comparing the chi-square statistics of the two models. Some other fit indices compare the $\chi^2/df$ of a hypothesized model with that of its baseline model. Because $E(\chi^2/df) = 1$ for a true model, these indices compare a hypothesized model with its baseline model in the deviation of the chi-square statistic from its expected value under the assumption that the hypothesized model is correctly specified. The TLI, also known as NNFI, is one such index, given by

$$\text{TLI or NNFI} = (\chi_0^2/df_0 - \chi_1^2/df_1)/(\chi_0^2/df_0 - 1). \tag{7}$$

It actually compares the hypothesized model with a "true" model in the ability to correct the chi-square deviation from the independence model.TLI (NNFI) is only approximately normed because, although its value typically falls between zero and 1, it is not completely constrained to the zero-to-1 range.

## Population-Based Absolute Fit Indices

Steiger and Lind (1980) gave the population discrepancy function in the form of $F(\Sigma; \Sigma(\theta))$, which "compares" the reproduced covariance matrix with the population covariance matrix instead of the sample covariance matrix. The minimized value of the population discrepancy function is

$$\delta = F_{\min}(\Sigma; \Sigma(\theta)). \tag{8}$$

Steiger et al. (1985) showed that under standard assumptions, the noncentrality parameter is equal to the product of the sample size and the minimized population discrepancy function, that is,

$$\lambda = N \times \delta. \tag{9}$$

Therefore, the estimate of minimized population discrepancy function is

$$d = l/N = (\chi^2 - df)/N. \tag{10}$$

A good model is supposed to reproduce a covariance matrix close to the population covariance matrix, and, therefore, a value of $d$ closer to zero is desired. The calculation of all population-based fit indices is based on $d$ in some form. The normed version of $d$ is McDonald's (1989) centrality index (Mc), given by,

$$\text{Mc} = \exp(-{}^1/_2 d), \tag{11}$$

and its value typically lies within the zero-to-1 range (but can exceed 1).

Like the chi-square statistic, $d$ favors complex models. The root mean square error of approximation, called RMS by Steiger and Lind (1980) or RMSEA by Browne and Cudeck (1993), takes model complexity into consideration by dividing $d$ with the model $df$ and then taking the square root:

$$\text{RMSEA} = \text{SQRT}(d/df). \tag{12}$$

The value of RMSEA closer to zero indicates a better model. Both $d$ and RMSEA have known sampling distributions, and most estimation tools give their confidence intervals.

## Population-Based Relative Fit Indices

McDonald and Marsh's (1990) relative noncentrality index (RNI) and Bentler's (1990) CFI are similar population-based relative fit indices. If $d_1$ and $d_0$ are denoted as the minimized value of the population discrepancy function for the hypothesized model and for the independence model, respectively, then RNI is given by

$$\text{RNI} = 1 - d_1/d_0 = 1 - (\chi_1^2 - df_1)/(\chi_0^2 - df_0), \tag{13}$$

and CFI is expressed as

$$\text{CFI} = 1 - \max(d_1, 0)/\max(d_0, d_1, 0) = 1 - \max(\chi_1^2 - df_1, 0)/\max(\chi_0^2 - df_0, \chi_1^2 - df_1, 0). \tag{14}$$

CFI can be viewed as a normed version of RNI because its value is bounded between zero and 1, whereas the value of RNI is not. When the value of RNI falls into the zero-to-1 range, RNI and CFI are the same. RNI or CFI indicates how much the hypothesized model corrects the noncentrality of the chi-square distribution from the independence model. Because RNI and CFI are relative and do not assume $\Sigma(\theta) = \Sigma$, some researchers have suggested that they should be fit indices of choice (e.g., Byrne, 1998).

However, RNI and CFI favor complex models. Now, the only available population-based relative fit indices adjusted for complexity are the parsimony versions of RNI and CFI. James, Mulaik, and Brett (1982) suggested using the parsimony index (PI) to evaluate the complexity of the hypothesized model against the independence model:

$$PI = df_1/df_0. \tag{15}$$

The parsimony version of RNI or CFI can be obtained by multiplying them times PI, for example,

$$PCFI = CFI \times PI. \tag{16}$$

However, there are no clear rules about how to interpret such indices, and some researchers (e.g., Marsh & Hau, 1996; Williams & Holahan, 1994) have questioned their usefulness.

## CHARACTERISTICS OF FIT INDICES

In addition to the three dimensions relating to the characteristics of fit indices as conceived by Tanaka (1993), other dimensions have also been identified. In this section, I discuss some of the characteristics of fit indices that are particularly related to their use in CFA.

### Availability of Cutoff Criteria

Cutoff criteria may be available for fit indices if the value typically falls within a certain range. A value close to either the typical lower or upper bound indicates a good fit. The former includes SRMR and RMSEA that have a minimum value of zero, and the latter includes normed or approximately normed fit indices, such as GFI, AGFI, Mc, and relative fit indices, with the typical upper bound of 1. For most fit indices, there are some rules of thumb. The rule of thumb for (approximately) normed fit indices is that a value greater than 0.9 indicates an acceptable fit and a value greater than 0.95 indicates a good fit. Browne and Cudeck (1993) suggested that an RMSEA less than 0.08 indicates an acceptable model.

Hu and Bentler (1999) examined some cutoff criteria using simulations, and they found that the cutoff criteria should be slightly more strict than conventional rules of thumb for most fit indices in model evaluation and selection. In particular, they suggested that the cutoff values required in order to conclude that there is a relatively good fit should be close to 0.95 for TLI (NNFI), CFI, and RNI; 0.90 for Mc; 0.08 for SRMR; and 0.06 for RMSEA.

Basic statistics, including the chi-square statistic ($\chi^2$), model *df,* and sample size *(N),* should also be reported to facilitate future cross-validation studies. The ratio $\chi^2/df$ can be calculated, and a value close to 1 indicates a good model, but there is not a clear cutoff criterion. From their experiences, some researchers have proposed that a $\chi^2/df$ ratio less than 2 or 3 probably indicates an acceptable model (e.g., Byrne, 1989; Carmines & McIver, 1981).

### Sensitivity to Sample Size

Sample size can have two effects on the value of fit indices: an *inflation* effect for large sample size and a *bias* effect for small sample size. Inflation occurs when the value of a fit

index increases systematically when the sample size becomes larger. Bias occurs when a fit index tends to reject "good" models or accept "bad" models when the sample size is small. These sample size effects are related to how fit indices are constructed.

Most sample-based absolute fit indices, such as the chi-square statistic, $\chi^2/df$, and AIC, are relatively vulnerable to sample size effects because their calculations involve sample size. Although the calculation of GFI does not involve sample size explicitly, its value is still subject to sample size as mentioned. Because the minimized population discrepancy function *(d)* uses sample size *(N)* as the divisor in its calculation to mitigate the inflation effect, population-based absolute fit indices (which are based on *d*) are more robust, as supported by empirical studies (e.g., Fan, Thompson, & Wang, 1999). However, they may still be vulnerable to the small sample bias effect, because *d*, based on the chi-square statistic, is sensitive to the violation of normality.

Compared with sample-based absolute fit indices, relative fit indices are less affected by the change in sample size because they basically compare model fit at the same sample size. However, some fit indices are still subject to the sample size effects systematically due to the way they are constructed. Using the relationship $\delta = (\chi^2 - df)/N$ under certain standard assumptions, the population versions of the sample-based relative fit indices can be derived by replacing $\chi^2$ with $N\delta+df$ in the calculation of these indices (Bentler, 1990; McDonald & Marsh, 1990):

$$\text{population NFI} = [(\delta_0 - \delta_1)+(df_0 - df_1)/N]/(\delta_0 + df_0/N) \quad (17)$$

$$\text{population TLI (NNFI)} = (\delta_0/df_0 - \delta_1/df_1)/(\delta_0/df_0). \quad (18)$$

The population version of NFI still depends on sample size, but the population version of TLI (NNFI) does not. Therefore, researchers (e.g., Hu & Bentler, 1995) do not recommend NFI, although it is simple and normed, but recommend TLI (NNFI) because it is relatively robust to the large sample inflation effect. However, when sample size is small, the standard assumptions can be violated, and the bias effect may occur. On the other hand, RNI and CFI are already population-based and should be relatively robust to both sample size effects. Consistent with the reasoning just outlined, Hu and Bentler (1998, 1999) found, using simulations, that RMSEA, Mc, TLI (NNFI), RNI, and CFI are relatively robust to the large sample inflation effect, but TLI (NNFI), Mc, and RMSEA are somewhat vulnerable to the small sample bias effect and, thus, are not preferred when sample size is small.

## Sensitivity to Model Misspecification

In CFA, there are two types of model misspecifications: misspecified factor covariance and misspecified factor loading. The former is related to the discriminant validity and the latter is related to the convergent validity of score interpretation. Hu and Bentler (1998) found that SRMR is most sensitive to factor covariance misspecification and that TLI (NNFI), RNI, CFI, Mc, and RMSEA are most sensitive to the factor loading misspecification. Most population-based fit indices are sensitive to model misspecification because their formations take model misspecification into account at the discrepancy assumption level.

## Sensitivity to Estimation Methods

Some fit indices perform more consistently across different estimation methods than do others. Hu and Bentler (1998) found that the performance of SRMR, TLI (NNFI), RNI, and CFI were relatively stable across all three methods. However, the results of Mc and RMSEA were only consistent for ML and GLS, but not for ADF. They also found that fit indices obtained from ML outperformed those obtained from GLS and ADF in assessing goodness of fit. Table 2 summarizes the characteristics of fit indices as discussed in the preceding section.

## TABLE 2

### Summary of the Characteristics of Fit Indices Related to Their Use in Confirmatory Factor Analysis

| Fit Index | Category | Cutoff Criteria Available | Robust to | | Sensitive to Model Misspecification | | Consistent for Major Estimation Methods |
|---|---|---|---|---|---|---|---|
| | | | Small *N* | Large *N* | Factor Covariance | Factor Loading | |
| $\chi^2$ | SAU | | | | | | |
| RMSR | SAU | | | | | | |
| SRMR | SAU | yes | | | yes | | yes |
| GFI | SAU | somewhat | | | | | |
| $\chi^2/df$ | SAA | somewhat | | | | | |
| AIC | SAA | | | | | | |
| AGFI | SAA | somewhat | | | | | |
| NFI | SRU | somewhat | | | | | |
| TLI (NNFI) | SRA | yes | | yes | | yes | yes |
| *l* | PAU | | | | | | |
| *d* | PAU | | | | | | |
| Mc | PAU | yes | | yes | | yes | ML & GLS, not ADF |
| RMSEA | PAA | yes | | yes | | yes | ML & GLS, not ADF |
| RNI or CFI | PRU | yes | yes | yes | | yes | yes |
| PCFI | PRA | | | | | | |

*Note.* RMSR = root mean square residual; SRMR = standardized root mean square residual; GFI = goodness-of-fit index; AIC = Akaike information criterion; AGFI = adjusted goodness-of-fit index; NFI = normed fit index;TLI = Tucker-Lewis index; NNFI = non-normed fit index; *l* = the estimate of noncentrality parameter; *d* = the estimate of minimized population discrepancy function; Mc = McDonald's centrality index; RMSEA = root mean square error of approximation; RNI = relative noncentrality index; CFI = comparative fit index; PCFI = parsimony comparative fit index; SAU = sample-based, absolute and unadjusted (for complexity); SAA = sample-based, absolute and adjusted (for complexity); SRU = sample-based, relative and unadjusted; SRA = sample-based, relative and adjusted; PAU = population-based, absolute and unadjusted; PAA = population-based, absolute and adjusted; PRU = population-based, relative and unadjusted; PRA = population-based, relative and adjusted; ML = maximum likelihood; GLS = generalized least squares; ADF = asymptotically distribution-free.

## GOODNESS-OF-FIT ASSESSMENT

As mentioned, CFA can be used for three main purposes: construct validity evaluation, response pattern comparison, and competing model comparison. Different types of fit indices assess different aspects of goodness of fit and may be appropriate for different purposes. To select fit indices that perform better than others under particular circumstances, individual characteristics of different types of fit indices must also be compared.

### Construct Validity Evaluation

To validate score interpretation with stand-alone evaluation of model fit, it is necessary to use fit indices that have well-established cutoff criteria, or at least some rules of thumb, available. Other characteristics of fit indices need to be taken into account as well, especially the sensitivity to sample size and model misspecification. Combining these considerations, the following fit indices are recommended: SRMR, TLI (NNFI), Mc, RMSEA, and CFI (or RNI). Among these fit indices, Mc and RMSEA are suitable for ML and GLS estimation methods but not for ADF, and others work for all three methods.

These fit indices evaluate different aspects of goodness of fit. SRMR is a sample-based absolute fit index, and it measures the extent of discrepancy between the reproduced and sample covariance matrices. Unlike other fit indices that are based on either the central (for sample-based indices) or noncentral (for population-based indices) chi-square distribution, SRMR makes no assumptions about the form of the underlying sampling distribution. This makes SRMR particularly desirable because "essentially nothing is known about the theoretical sampling distribution of the various estimators" (Bentler, 1990, p. 245). Mc and RMSEA are population-based absolute fit indices, and they estimate the chi-square noncentrality that results from model misspecification. TLI (NNFI) is a sample-based relative fit index, and CFI (or RNI) is a population-based relative fit index. They use the independence model as a base from which to estimate how the hypothesized model improves the chi-square statistic or chi-square noncentrality. Whereas SRMR, Mc, and CFI (or RNI) favor complex models, TLI (NNFI) and RMSEA are adjusted for model complexity.

Together these fit indices comprehensively evaluate the construct validity, particularly the convergent and discriminant validity, of scores from instruments. As mentioned, SRMR is most sensitive to factor covariance misspecification, and Mc, RMSEA, TLI (NNFI), and CFI (or RNI) are most sensitive to the factor loading misspecification. Therefore, SRMR is good for the evaluation of discriminant validity, and the others are good for the evaluation of convergent validity.

## Response Pattern Comparison

The first step in comparing the patterns of response to a measurement instrument is to test for the invariant factor structure across different groups. In the studies surveyed, researchers often conducted separate CFA for each group and compared the goodness-of-fit results. Many mainstream estimation tools now allow users to fit scores to a measurement model for multiple groups simultaneously. Compared with single-group analyses, the simultaneous multigroup analysis is more efficient and appropriate because it assesses the goodness of fit once, but for all groups. A set of joint versions of fit indices, the same as recommended for stand-alone evaluation of model fit, can evaluate the overall model fit across different samples.

If the same factor structure is reasonably adequate across different groups, the next step is to test whether the parameter estimates can be held invariant as well. It is necessary to differentiate two kinds of situations regarding the availability of what is known as a *reference group*. When one sample is drawn from a whole population while another is drawn from a subset, the former sample can be regarded as the reference group and the latter can be regarded as the *target group*. To test whether the parameter estimates are invariant from the reference group to the target group, it is possible to fit the reference group to the factor structure, constrain model parameters with estimates, and then fit the target group to the constrained model. Because the constrained model is nested in the original model, the chi-square test—which examines the significance of the change in the chi-square statistic against the change in model *df*—can be used to check whether there is significant deterioration in goodness of fit for the target group. If there is, it suggests that at least one of the parameters is variant across groups. To identify the variant parameters, it is a common practice to free one or more parameters of interest (e.g., factor correlations, structure coefficients, measurement error terms) at a time and conduct additional chi-square tests to see whether the model fit improves significantly.

In many cases, however, a reference group is unavailable because no single group can represent the whole population. In testing gender difference, for example, neither the male group nor the female group can be regarded as the reference group for the other. This situation calls for the use of simultaneous multigroup analysis. To test for the invariance of parameters and identify those that are variant, all or some of the parameters in the measurement model can be set invariant across different groups. In this case, the chi-square test examines the significance of change in the joint chi-square statistic against the change in model *df* due to the modifications on model parameters (i.e., placing or releasing constraints on parameters).

Because the chi-square test is based on the assumption of central chi-square distribution, it would be advisable to check this assumption with noncentrality estimators. There are known sampling distributions for the noncentrality parameter estimate *(l)* and the minimized population discrepancy function estimate *(d),* and most estimation tools give their confidence intervals. In the first step to test the invariance of factor structure across groups, possible violation of the assumption can be checked by examining whether zero is included in the confidence interval of *l* or *d*. If not, it indicates that the factor structure itself is inadequate, and caution should be taken in using the chi-square test in the steps that would follow. If no severe violation is detected, constraints can be placed on model parameters, and it can be determined whether there is a significant shift in noncentrality. If the confidence intervals of *d* or *l* for original and constrained models do not overlap, it indicates that these constraints lead to a significant deterioration in goodness of fit and further chi-square testing is not necessary.

## Competing Model Comparison

When competing models share the same set of indicators, there is at least one model (i.e., independence model) nested in all candidate models to serve as a baseline for comparison. Using the method suggested by Widaman and Thompson (2003), it is possible to identify or generate a baseline model that is better than the independence model. With a common baseline model available, candidate models in goodness of fit can be compared by using relative fit indices. TLI (NNFI) and CFI (or RNI) are preferred because they are relatively robust to sample size effects and tend to yield consistent results at different sample sizes (however, if the sample size is small, TLI or NNFI may not be appropriate). Together, they assess different aspects of goodness of fit: TLI (NNFI) is sample-based and adjusted for complexity and CFI (or RNI) is population-based but not adjusted for complexity. Both are sensitive to the misspecification of factor loading structure. If a common baseline model other than the independence model is adopted, it is necessary to manually calculate the revised versions of TLI (NNFI) and CFI (or RNI) with the chi-square statistic and *d* values obtained from separate estimation of the baseline model and candidate models.

In many cases, competing models do not share the same set of indicators, and there are no common baseline models. Comparing such models is less straightforward but more subjective because direct comparison of relative fit indices is not meaningful. Some researchers (e.g., Burnham & Anderson, 2002) have recommended AIC for the comparison of candidate models, and better models should have a lower value of AIC. AIC retains most of the original model fit information while it is adjusted for model complexity. However, AIC may not perform very consistently across different sample sizes. In CFA, it may be helpful to further compare models with other fit indices that indicate different aspects of goodness of fit, as in stand-alone evaluation of model fit. Appropriate indices should be relatively robust to sample size effects; sensitive to model misspecification; and, preferably, adjusted for complexity. Therefore, SRMR, RMSEA, TLI (NNFI), and CFI (or RNI) are recommended in addition to AIC. When the model comparison yields consistent results, a conclusion can be reached with confidence. Otherwise, the judgment that one model is better than another should be made cautiously. Table 3 lists the suggestions regarding goodness-of-fit assessment for different purposes in CFA.

## CONCLUSION

In this article, I first identified three main purposes of conducting CFA on the scores from measurement instruments: construct validity evaluation, response pattern comparison, and competing model comparison. With Tanaka's (1993) multifaceted conceptions as the basis, I proposed a hierarchical classification scheme according to how fit indices are constructed. Using this scheme, I discussed the characteristics of fit indices. Finally, I reviewed how to assess goodness of fit for different purposes in CFA.

## TABLE 3

### Suggestions Regarding Goodness-of-Fit Assessment for Different Purposes in Confirmatory Factor Analysis (CFA)

| CFA Purpose | Practice/Situation | Goodness-of-Fit Assessment | Suggested Indices |
|---|---|---|---|
| Construct validity evaluation | • Stand-alone evaluation of model fit | • Evaluation of model fit from multiple aspects | • SRMR, TLI (NNFI), Mc, RMSEA, CFI (or RNI) |
| Response pattern comparison | • Test for invariant factor structure | • Simultaneous evaluation across multiple groups | • Joint versions of above indices; CI of *l* or *d* |
| | • Test for invariant model parameters | • Chi-square test (single-group or multigroup) | • $\Delta\chi^2$ against $\Delta df$ (due to parameter contraints) |
| Competing model comparison | • Common baseline models available | • Standard or revised relative fit indices | • TLI (NNFI), CFI (or RNI) |
| | • Common baseline models unavailable | • Primary comparison | • AIC |
| | | • Refined comparison in multiple aspects | • SRMR, TLI (NNFI), RMSEA, CFI (or RNI) |

*Note.* SRMR = standardized root mean square residual; TLI = Tucker-Lewis index; NNFI = nonnormed fit index; Mc = McDonald's centrality index; RMSEA = root mean square error of approximation; CFI = comparative fit index; RNI = relative noncentrality index; CI = confidence interval; *l* = the estimate of noncentrality parameter; *d* = the estimate of minimized population discrepancy function; $\Delta\chi^2$ = change in the value of the chi-square statistic; $\Delta df$ = change in degrees of freedom available for the model estimation; AIC = Akaike information criterion.

The focus of this article is on the systematic discussion of the properties and characteristics of fit indices and how they can be related to the goodness-of-fit assessment in CFA. This article only surveyed the typical uses of CFA found in 16 articles in the *MECD* journal that were published between 1996 and 2002. It is hoped that this discussion can help researchers to select fit indices on a less arbitrary and more theoretical basis and that it will enhance further discussion.

## REFERENCES


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

Beasley, T. M., Long, J. D., & Natali, M. (2001). A confirmatory factor analysis of the Mathematics Anxiety Scale for Children. *Measurement and Evaluation in Counseling and Development, 34,* 14–26.

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

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.

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.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information theoretic approach.* New York: Springer-Verlag.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models.* New York: Springer-Verlag.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum

Carmines, E. G., & McIver, J. P.(1981). Analyzing models with unobserved variables. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65–115). Beverly Hills, CA: Sage.

Chan, D. W., & Lin, W.-Y. (1996). The two- and three-dimensional models of the HK-WISC: A confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 28,* 191–199.

Chiu, L. H., & Henry, L. I. (1990). Development and validation of the Mathematics Anxiety Scale for Children. *Measurement and Evaluation in Counseling and Development, 23,* 121–127.

Cokley, K. O., Bernard, N., Cunningham, D., & Motoike, J. (2001). A psychometric investigation of the Academic Motivation Scale using a United States sample. *Measurement and Evaluation in Counseling and Development, 34,* 109–119.

Cokley, K. O., & Helm, K. (2001). Testing the construct validity of scores on the Multidimensional Inventory of Black Identity. *Measurement and Evaluation in Counseling and Development, 34,* 80–95.

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.

Denzine, G. M., & Kowaliski, G. J. (2002). Confirmatory factor analysis of the Assessment for Living and Learning Scale: A cross-validation investigation. *Measurement and Evaluation in Counseling and Development, 35,* 14–26.

Erford, B. T., Peyrot, M., & Siska, L. (1998). Analysis of teacher responses to the Conners Abbreviated Symptoms Questionnaire (ASQ). *Measurement and Evaluation in Counseling and Development, 31,* 2–14.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indices. *Structural Equation Modeling, 6,* 56–83.

Fuertes, J. N., Miville, M. L., Mohr, J. J., Sedlacek, W. E., & Gretchen, D. (2000). Factor structure and short form of the Miville-Guzman Universality-Diversity Scale. *Measurement and Evaluation in Counseling and Development, 33,* 157–169.

Good, G. E., Robertson, J. M., O'Neil, J. M., Fitzgerald, L. F., Stevens, M., DeBord, K. A., et al. (1995). Male gender role conflict: Psychometric issues and relations to psychological distress. *Journal of Counseling Psychology, 42,* 3–10.

Hong, S.-M., & Faedda, S. (1996). Refinement of the HPRS. *Educational and Psychological Measurement, 56,* 173–182.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hyle (Ed.), *Structure equation modeling: Concept, issues, and applications* (pp. 76–97). Thousand Oaks, CA: Sage.

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

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

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

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

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kirisci, L., & Moss, H. B. (1997). Reliability and validity of the Situational Confidence Questionnaire in an adolescent sample: Confirmatory factor analysis and item response theory. *Measurement and Evaluation in Counseling and Development, 30,* 146–155.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Loo, R. (2001). Motivational orientations toward work: An evaluation of the Work Preference Inventory (student form). *Measurement and Evaluation in Counseling and Development, 33,* 222–233.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

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

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

McCoach, D. B. (2002) A validation study of the School Attitude Assessment Survey. *Measurement and Evaluation in Counseling and Development, 35,* 66–77.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

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

O'Rourke, N., & Cappeliez, P. (2001). Marital satisfaction and marital aggrandizement among older adults: Analysis of gender invariance. *Measurement and Evaluation in Counseling and Development, 34,* 66–79.

Rogers, J. R., Abbey-Hines, J., & Rando, R. A. (1997). Confirmatory factor analysis of the Gender Role Conflict Scale: A cross-validation of Good et al., 1995. *Measurement and Evaluation in Counseling and Development, 30,* 137–145.

Rogers, J. R., & Hanlon, P. J. (1996). Psychometric analysis of the College Student Reasons for Living Inventory. *Measurement and Evaluation in Counseling and Development, 29,* 13–24.

Sellers, P. M., Rowley, S. A., Chavous, T. M., Shelton, J. N., & Smith, M. A. (1997). Multidimensional Inventory of Black Identity: A preliminary investigation of reliability and construct validity. *Journal of Personality and Social Psychology, 73,* 805–815.

Slaney, R. B., Rice, K. G., Mobley, M., Trippi, J., & Ashby, J. S. (2001). The revised Almost Perfect Scale. *Measurement and Evaluation in Counseling and Development, 34,* 130–145.

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.

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

Thomas, A., Donnell, A. J., & Buboltz, W. C. (2001). The Hong Psychological Reactance Scale: A confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 34,* 2–13.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–284). Washington, DC: American Psychological Association.

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

Utsey, S. O. (1999). Development and validation of a short form of the Index of Race-Related Stress (IRRS) Brief version. *Measurement and Evaluation in Counseling and Development, 32,* 149–167.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement, 52,* 1003–1017.

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.

Williams, L. J., & Holahan, P. J. (1994). Parsimony-based fit indices for multiple indicator models: Do they work? *Structural Equation Modeling, 1,* 161–187.