# Making Sense of Fit Indices in Structural Equation Modeling (SEM)

A. J. Guarino
David M. Shannon
Margaret E. Ross

Auburn University

There are many excellent resources on the basic conceptual understanding of structural equation modeling (SEM): Bryant and Yarnold (1995); Byrne (1994 & 2001); Hair, Anderson, Tatham, and Black (1998); Klem (1995 & 2000); Kline (1998) Maruyama (1998); Mueller (1997); Pedhazur & Schmelkin (1991); Stevens (1996); Tabachnick and Fidell (2001) and Thompson (2000) to name a few. Our review of this *SEM* literature detected no consensus in the organization of the myriad of fit indices assessing the adequacy of a hypothesized model. The purpose of this paper is to propose a nomenclature of the numerous fit indices that are used to assess the validity of structural models.

*SEM* is a statistical technique that evaluates the plausibility of a hypothesized model. The full structural model can be decomposed into the structural model and the measurement model. The structural model assesses the relationships among the latent construct variables. These latent constructs are usually defined by three to five measured variables (e.g., survey items, test scores, attitude scores). These multiple measures (the measurement model) allow the researcher to more effectively control for the inevitable measurement errors of any construct. By controlling for measurement error, unbiased estimates of the relationships among the latent constructs are possible. Once a model is proposed (i.e., relationships among the variables have been hypothesized) a correlation/covariance matrix is created. The estimates of the relationships among the variables in the model are calculated utilizing the maximum likelihood estimation (*MLE*) procedure. *MLE* attempts to estimate the values of the parameters that would result in the highest likelihood of the actual data to the proposed model. These methods often require iterative solutions. With small samples, *MLE* may not be very accurate.

The model is then compared to the relationships (the correlation/covariances matrix) of the actual-observed data. SEM assesses how well the predicted interrelationships among the variables match the interrelationships among the actual or observed interrelationships. *SEM* assesses the measurement model (how well the measured variables define their respective construct) and the structural model (how well the latent constructs relate to each other) simultaneously. If the two matrices (the proposed and the actual-data) are consistent with one another, then the structural equation model can be considered a credible explanation for the hypothesized relationships.

## Assessing Fit of Hypothesized Models

The methodologies in assessing the fit between the proposed model and the data are rapidly developing in *SEM*. Rigdon (1998) observes, "This rapid change is a source of excitement for some researchers and a source of frustration for others" (p. 91). One area that may cause frustration and confusion for researchers is the myriad of fit indices assessing the adequacy of the proposed model to the observed data. Over the past 20 years, at least 24 fit indices have been proposed[1] (Klem, 2000).

All these *SEM* fit indices were developed to diminish the Type II error (i.e., concluding that the data does not support the proposed model when in fact it does). For the 24 *SEM* fit measures available through statistical software programs there is presently no general agreement on which measures are preferred. As Hair (1998) states, "*SEM* has no single statistical test that best describes the 'strength' of the model's predictions" (p. 653). None of the measures has a related statistical test, except for the chi-square test (Hair, 1998).

The confusing consequence of these competing fit indices is that different research studies report different fit indices. Complicating this matter of competing fit indices is the lack of consensus among *SEM* writers as to the organization of fit indices. Nor is there agreement regarding the particular classification of the individual fit measures, further complicating the decision of what fit index to report. Therefore, researchers have proposed various methods of classifying and organizing such indices. A discussion of these classification schemes follows.

## Classification of Fit Indices

According to Maruyama (1998), "The different fit indexes differ with respect to dimensions such as susceptibility to sample size differences, variability in the range of fit possible for any particular data set, and valuing simplicity of model specification needed to attain an improved fit" (p. 239). The different fit indices have been organized in various ways.

Many *SEM* writers present a few fit measures as alternatives to the chi-square test (e.g., Klem, 1995 & 2000; Kline, 1998; Pedhazur & Schmelkin, 1991; Stevens, 1996 and Thompson, 2000) while others have developed classification schemes. Maruyama (1998) proposes a two-classification scheme (absolute and relative with four subtypes). Other authors (Chin, 1995; Hair, Anderson, Tatham, & Black, 1998; and Jaccard & Wan, 1996) promote a three-classification scheme (absolute, relative, and parsimonious). Tabachnick & Fidell (2001) suggest a five-classification scheme (comparative, absolute, proportion of variance, parsimony, and residual-based). Arbuckle (1999) devised an eight-classification scheme (parsimony, sample discrepancy, population discrepancy, information-theoretic, baseline model, parsimony adjusted, goodness of fit, and miscellaneous). To assess the adequacy of a full structural

model, we believe that there are two basic classification schemes (absolute and relative) and that both of these can be subdivided into two more dimensions (adjusted and unadjusted).

Absolute fit measures judge how well the proposed interrelationships among the variables match the interrelationships among the actual or observed interrelationships. This means how well the correlation/covariance of the hypothesized model fits the correlation/covariance of the actual or observed data.

Relative fit measures are also known as comparisons to baseline measures or incremental fit measures. These are measures of fit relative to the independence model, which assumes that there are no relationships in the data (thus a poor fit) and the saturated model, which assumes a perfect fit. The incremental fit measures indicate the relative position on this continuum between worst fit to perfect fit with values greater than .90 suggesting an acceptable fit between the model and the data.

Both the absolute and the relative measures can be subdivided into two subcategories, adjusted and unadjusted. Adjusted is often referred to as parsimonious fit measures. These measures "adjust" the measures to provide a comparison between models with different numbers of estimated parameters to determine the impact of adding additional parameters to the model. These fit statistics are similar to the adjusted $R^2$ in multiple regression analysis: the parsimony fit statistics penalize larger models with more estimated parameters. Recall that *MLE* maximizes the likelihood that the data will support the proposed model. The more paths a researcher is estimating the more likely the fit will be acceptable no matter how nonsensical the model may be. Researchers could "stack the deck" in their favor by increasing the complexity of their model. However, one goal of superior research is developing parsimonious models. Typically, parsimony-based measures have lower acceptable values (e.g., .50 or greater is deemed acceptable; Mulaik et al.; 1989).

Table 1 presents the names of the individual fit indices and their proposed classification. Selected tests are highlighted below because we believe these are the fit measures most often reported to assess the adequacy of a full structural model.

**Absolute Fit Measures**

The four most common unadjusted, or non-parsimonious, absolute fit measures are the chi-square, the goodness-of-fit (GFI), the root mean square residual (RMSR) and the root mean square error of approximation (RMSEA). The adjusted goodness-of-fit (AGFI) and the parsimonious goodness-of-fit (PGFI) are common adjusted (parsimonious) absolute fit measures.

The chi-square statistic is utilized to test the difference between the predicted and the observed relationships (correlations/covariances). Because the researcher is predicting a close fit, a non-significant chi-square is preferred. The chi-square test, however, is too powerful. As sample size increases, power increases. Therefore, the chi-square test can detect small differences between the observed and predicted covariances suggesting that the model does not fit the data when in fact it does. Because of these limitations, other fit indices were developed as alternatives to the chi-square.

One alternative is the Goodness of Fit Index (GFI). GFI is conceptually similar to the R-Square in multiple regression (Kline, 1998). It is the proportion of variance in the sample correlation/covariance accounted for by the predicted model with values ranging from 0 (no fit) to 1 (a perfect fit). GFI varies from 0 to 1 but theoretically can yield meaningless negative values. By convention, GFI should be equal to or greater than .90 as indicative of an acceptable model.

The Root Mean Square Residual (RMSR) is a measure of the average size of the residuals between actual covariance and the proposed model covariance. The smaller the RMSR, the better the fit (e.g., < .05).

The Root Mean Square Error of Approximation (RMSEA) is the average of the residuals between the observed correlation/covariance from the sample and the expected model estimated for the population. Byrne (1998) states, "(RMSEA) has only recently been recognized as one of the most informative criteria in covariance structure modeling" (p. 112). Values less than .08 are deemed acceptable while values greater than .10 are generally unacceptable.

The adjusted, or parsimonious, absolute fit measures are the relative chi-square, adjusted goodness-of-fit index (AGFI), and the parsimonious goodness-of-fit index (PGFI). Proposed by Joreskog (1970), the relative chi-square is the chi-square divided by the degrees of freedom. Some researchers suggest values as high as five for an acceptable fit while others maintain relative chi-square be two or less.

The AGFI adjusts the GFI for degrees of freedom. Values greater than .90 indicate an acceptable model and this measure is useful when comparing models. It, too, varies from 0 to 1, but theoretically can yield meaningless negative values. AGFI should also be at least .90.

PGFI adjusts for degrees of freedom in the baseline model. It is a variant of GFI that penalizes GFI by multiplying it times the ratio formed by the degrees of freedom in your model and degrees of freedom in the independence model.

**Incremental Fit Measures**

Incremental fit measures are also referred to as comparisons to baseline measures or relative fit measures. Byrne (1998) suggests that the Comparative Fit Index (CFI) should be the fit statistic of choice in *SEM* research. Knight et al. (1994) have suggested the following CFI fit indices: good fit > .90; adequate but marginal fit = .80 to .89; poor fit = .60 to .79; very poor fit < .60. Hu and Bentler (1999) revised the minimum value to .95. Other common incremental fit measures are the Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), Incremental Fit Index (IFI) and the Relative Fit

Table 1
*Classification of Fit Measures with Acceptable Values*

| | Absolute Fit Indices | | Relative Fit Indices | |
|---|---|---|---|---|
| | Measures | Acceptable Fit Value | Measures | Acceptable Fit Value |
| Unadjusted Models | $\chi^2$ | $p > .05$ | CFI | ≥.95 |
| | GFI | ≥.09 | NFI | ≥.95 |
| | RMSR | ≤.05 | NNFI (TLI) | ≥.95 |
| | RMSEA | ≤.08 | IFI | ≥.95 |
| | | | RFI | ≥.95 |
| Adjusted Models | $\chi^2/df$ | <3.0 | PNFI | ≥.50 |
| | PGFI | ≥.50 | PCFI | ≥.50 |
| | AGFI | ≥.90 | | |

$\chi^2$ = Chi-square test; GFI = Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation; $\chi^2/df$ = Chi-square divided by degrees of freedom test; PGFI = Parsimony Goodness of Fit Index, AGFI = Adjusted Goodness of Fit Index; CFI = Comparative Fit Index; NFI = Normed Fit Index; NNFI = Non-Normed Fit Index (formerly known as the TLI = Tucker-Lewis Index; IFI = Incremental Fit Index; RFI = Relative Fit Index; PNFI = Parsimony Normed Fit Index; PCFI = Parsimony Comparative Fit Index.

Index (RFI). A summary of acceptable fit values is found in Table 1. Two parsimonious measures in this classification are the parsimony adjusted CFI (PCFI) and the parsimony adjusted GFI (PGFI). These measures "adjust" the measures to provide a comparison between models with different numbers of estimated parameters to determine the impact of adding additional parameters to the model. Although no statistical test is available for these measures, values of .50 or greater are deemed acceptable (Mulaik et al., 1989).

**What to Report?**

Not surprisingly there is no agreement on which fit indexes to report. Byrne (1994) states "assessment of model adequacy must be based on multiple criteria that take into account theoretical, statistical, and practical consideration" (p.119). In other words, Jaccard and Wan (1996) recommend reporting at least three fit tests, one absolute, one relative, and one parsimonious to reflect diverse criteria. Kline (1998) recommends at least four tests, such as chi-square; GFI, NFI, or CFI; NNFI; and RMSR. Garrison (2000) warns that one should avoid the shotgun approach of reporting all of them, which seems to imply the researcher is on a fishing expedition. We recommend reporting at least one fit measure from each of the four categories summarized in Table 1.

**Additional Assessment of the Model**

Stevens (1996) divides the assessment of a model into two categories "those that measure the overall fit of the model and those that are concerned with individual model parameters..." (p. 402-403). All of the *SEM* fit measures are analogous to the omnibus test in *ANOVA*s in that they provide an overall assessment of the model. As in the *ANOVA*, post-hocs need to be conducted to provide further interpretation of the analysis. Similarly, in *SEM* the overall fit of a model to the data may appear acceptable, yet some relations in the model may not be supported by the data. For example, an acceptable fit index could be achieved because of the strong measurement model though the structural model is fairly weak. Alternatively, the structural model may be impressive, but the measurement model may be quite weak making the interpretation meaningless. This means that fit indices are NOT related to the accuracy of prediction in the structural equations. Accuracy of the predictions are assessed by comparing the actual values of the dependent variables with their predicted values, usually in terms of root mean squared error or proportion of variance accounted for ($R^2$). Closely related to this concept of adequate variance explained, is the direction and magnitude of the individual parameters. Are the parameter values in the direction hypothesized? Most importantly, does the model make

sense?

## Conclusion

This review of the nomenclature of fit indices in structural equation modeling (*SEM*) was limited to those fit indices that assessed the adequacy of a full structural model. There is a class of fit measures known as the information-theoretic measures. These measures include the Akaike Information Criterion (AIC), the Browne-Cudeck Criterion (BCC), and the Bayes Information Criterion (BIC). Because these measures are used for comparing models and are not used to evaluate a single model, they were excluded from the current review. Our review reports a lack of consensus among *SEM* researchers regarding the organization of fit indices. We propose there are two basic classification schemes (absolute and relative) and that both of these can be subdivided into two more dimensions (adjusted and unadjusted). Additionally, there is little agreement about the particular classificatioin of the individual fit measures. Although we classified the adjusted goodness of fit index (AGFI) as an absolute fit measure, Hair, Anderson, Tatham, and Black (1998) described it as a relative fit measure. We also believe that the relative chi-square should be classified as an absolute-parsimonious measure.

In conclusion, Thompson (2000) states, "Because some of the fit indices evaluate different aspects of fit, it is important to evaluate fit based on multiple fit statistics..." (p. 271). Many others also recommend reporting different types to reflect the diverse criteria of fit measures (Jaccard and Wan, 1996; Klem, 1995 & 2000; Kline, 1998; Maruyama, 1998; Tabachnick and Fidell, 1996 & 2001; and Thompson, 2000). However, it is difficult to provide different fit measures when there is no consensus regarding their classification. We hope this brief review was helpful in providing one simple framework for organizing the many measures of fit used and reported with SEM research studies.

## References


Arbuckle, J.L. (1999). *Amos 4.0 User's Guide*. Chicago: SmallWaters Corporation.

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

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

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

Browne, M.W. and Cudeck, R. (1993). *Alternative ways of assessing model fit*. In Bollen, K.A. and Long, J.S. [Eds.] Testing structural equation models. Newbury Park, California: Sage, 136-162.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Thousand Oaks, CA: Sage Publications.

Bentler, P. M. and C. P. Chou (1987). Practical issues in structural modeling. *Sociological Methods and Research. 16*(1): 78-117.

Bollen, K. A. (1989). *Structural equations with latent variables*. NY: Wiley.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Hillsdale, NJ: Lawrence Erlbaum..

Carmines, E.G. and. McIver, J. P (1981). Analyzing models with unobserved variables: Analysis of covariance structures. Pp. 65-115 in George W. Bohmstedt and Edward F. Borgatta, eds., *Social Measurement*. Thousand Oaks, CA: Sage Publications.

Garson, G.D. (2000). PA 765 Statnotes: An Online Textbook http://www2.chass.ncsu.edu/garson/pa765/structur.htm

Hair, J.F., Anderson, R.E., Tathan, R.L., and Black, W. C. (1999). *Multivariate data analysis* (5th ed.). Upper Saddle River, New Jersey: Prentice Hall.

Hoyle, R.H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

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

Jaccard, J. and C.K. Wan (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage Publications.

Joreskog, K.G. and Sorbom, D. (1989). *LISREL-7 user's reference guide*. Mooreville, IN: Scientific Software.

Klem, L. (1995). Path analysis. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 65-98). Washington, DC: American Psychological Association.

Klem, L. (2000). Structural equation modeling. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 227-260). Washington, DC: American Psychological Association.

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

Knight, G. P., Virdin, L. M., Ocampo, K. A., & Roosa, M. (1994). An examination of the cross-ethnic equivalence of measures of negative life events and the mental health among Hispanic and Anglo American children. *American Journal of Community Psychology, 22*, 767-783.

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

Maruyama, G. M. (1997). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.

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

Mueller, R. O. (1996). *Basic Principles of structural equation modeling: An introduction to LISREL and EQS.*

Secaucus, NJ: Springer.
Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.
Rigdon, E. E. (1998). The Equal Correlation Baseline Model for Comparative Fit Assessment in Structural Equation Modeling. *Structural Equation Modeling. 5*(1) 63-77.
Schumaker, R.E. and R. G. Lomax, R.G. (1996). *A beginners guide to structural equation modeling.* Hillsdale, NJ: Erlbaum.
Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.
Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate Statistics* (3rd ed.). New York: HarperCollins.
Thompson, B. (2000). Ten commandments of structural equation modeling. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-283). Washington, DC: American Psychological Association.

## Footnotes

[1] LISREL provides 18 fit indices; EQS, 5 indices and AMOS 24 indices.

*A. J. Guarino teaches in the Department of Educational Foundations, Leadership, and Technology. His primary areas of expertise are structural equation modeling and college student persistence.*

*David Shannon is a Professor in the Department of Educational Foundations, Leadership, and Technology. His areas of research include student and teacher assessment, teacher effectiveness, and research methodology issues.*

*Margaret E. Ross is currently employed in the Educational Foundations, Leadership and Technology Department at Auburn University where she teaches assessment and statistics courses at the graduate and undergraduate level. Her research focuses on assessment issues, especially in relation to contextual variables in the classroom and student learning.*