Assessing overall model fit is an important problem in general structural equation models. One of the most widely used fit measures is Bentler and Bonett's (1980) normed index. This article has three purposes: (1) to propose a new incremental fit measure that provides an adjustment to the normed index for sample size and degrees of freedom, (2) to explain the relation between this new fit measure and the other ones, and (3) to illustrate its properties with an empirical example and a Monte Carlo simulation. The simulation suggests that the mean of the sampling distribution of the new fit measure stays at about one for different sample sizes whereas that for the normed fit index increases with N. In addition, the standard deviation of the new measure is relatively low compared to some other measures (e.g., Tucker and Lewis's (1973) and Bentler and Bonett's (1980) nonnormed index). The empirical example suggests that the new fit measure is relatively stable for the same model in different samples. In sum, it appears that the new incremental measure is a useful complement to the existing fit measures.

A New Incremental Fit Index for General Structural Equation Models

KENNETH A. BOLLEN
University of North Carolina at Chapel Hill

A highly controversial area in covariance structure analysis is the assessment of overall model fit. A chi-square estimate that tests the overidentifying restrictions of a model is available. However, as Jöreskog and Sörbom (1986: I-38) and others note, the assumptions that underlie the chi-square test are seldom satisfied in practice. To supplement the chi-square estimate, several other overall fit measures have been proposed including the Goodness of Fit Index (GFI) and adjusted GFI (Jöreskog and Sörbom, 1986), normed and nonnormed fit

AUTHOR'S NOTE: *An earlier version of this article was presented at the March 1988 Southern Sociological Society in Nashville, Tennessee. I wish to thank the referees and Scott Long for valuable comments.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 17 No. 3, February 1989 303-316

indices (Tucker and Lewis, 1973; Bentler and Bonett, 1980; Bollen, 1986), and Hoelter's (1983) Critical N (CN).

One of the most popular of these is Bentler and Bonett's (1980) normed fit index, Δ_1. In practice, two problems with Δ_1 have been largely ignored: (1) the mean of the sampling distribution of Δ_1 is positively related to sample size (Bearden, Sharma, and Teel, 1982) and (2) Δ_1 has no adjustment for degrees of freedom. The first problem implies that Δ_1 tends to give an overly pessimistic image of fit for small samples compared to large ones even if the identical model underlies all samples. The second problem could lead to a systematic bias against more parsimonious models, since an overparameterized model nearly always will have a higher Δ_1 than models that are nested within it.

The purposes of this article are (1) to propose a new incremental fit measure that provides an adjustment to Δ_1 for sample size and degrees of freedom, (2) to explain the relation of this new measure to Δ_1 and other measures, and (3) to report simulation results and an empirical example which compare the new measure to Δ_1 and several other fit indices.

THE NEW MEASURE

Consider Bentler and Bonett's (1980) normed fit index, Δ_1:

$$\Delta_1 = \frac{\chi_b^2 - \chi_m^2}{\chi_b^2} \qquad [1]$$

$$= \frac{F_b - F_m}{F_b} \qquad [2]$$

In equation 1 χ_b^2 is the value of the chi-square estimator for the "baseline" model, χ_m^2 is the value of the chi-square estimator for the maintained or hypothesized model, and in equation 2 F_b and F_m are analogously defined values of the fitting functions. Both equations are equivalent but equation 1 is often the most computationally convenient.[1] The baseline model is the most restrictive, theoretically defensible model (Bentler and Bonett, 1980). It provides a standard for a "bad" fit to which the maintained model is compared. The Δ_1 measures the proportionate improvement in fit by going from the baseline model to the maintained model. It varies from zero to one and the closer Δ_1 is to one, the better is the overall fit. The maximum value of one occurs only when the χ^2_m (or F_m) is zero.

A zero value for χ^2_m is not realistic to expect even when the model is perfectly correct. For a valid model with observed variables that obey well-known distributional assumptions (e.g., no excess kurtosis, see Browne, 1982), the chi-square estimator of χ^2_m follows an asymptotic chi-square distribution. The mean of a chi-square is its degrees of freedom (df), so for large samples the mean of χ^2_m is approximately df_m. For a correct maintained model, the numerator, $(\chi^2_b - \chi^2_m)$, is *on average* $(\overline{\chi}^2_b - df_m)$ where $\overline{\chi}^2_b$ is the "average" value of the baseline chi-square. In general the χ^2_b *estimator* does not follow a chi-square distribution since we know that the baseline model is too restrictive. Under some conditions it may follow a noncentral chi-square distribution (Steiger et al., 1985) but this is not essential to my argument.

If on average $(\overline{\chi}^2_b - df_m)$ is what we expect for a correct model, this provides an alternative standard to which to compare $(\chi^2_b - \chi^2_m)$. A correct model should have both of these equal on average. The Δ_2 measure is based on this rationale:

$$\Delta_2 = \frac{\chi^2_b - \chi^2_m}{\chi^2_b - df_m} \qquad [3]$$

The χ^2_b replaces the $\overline{\chi}^2_b$ in the denominator since it is the only estimator of $\overline{\chi}^2_b$ available.

Equation 3 shows that Δ_2 adjusts for the degrees of freedom of the maintained model such that models with fewer parameters have higher Δ_2 values for given values of χ^2_b and χ^2_m. The correction for sample size influences is made clearer by rewriting equation 3 in terms of the values of the fitting functions:

$$\Delta_2 = \frac{F_b - F_m}{F_b - (df_m/(N-1))} \qquad [4]$$

Other things equal, the denominator of equation 4 is smaller and Δ_2 is bigger, the smaller is N. If Δ_2 works as predicted, its mean should be about one for valid models in sufficiently large sample sizes. It is difficult to know its behavior with small to moderate N's. I present some simulation evidence on this issue below.

When the numerator and denominator of Δ_2 are positive, Δ_2 is greater than Δ_1. The Δ_2 is a "nonnormed" measure since it is possible for it to fall outside the zero to one bounds. The maximum value of Δ_2 is $\chi^2_b/(\chi^2_b - df_m)$ *provided that* $\chi^2_b > df_m$. When the baseline model is nested in the maintained model, the minimum of Δ_2 is zero. Given the known maximum of Δ_2, it is tempting to "norm" Δ_2 to have a ceiling of one by dividing Δ_2 by its maximum (that is, $\Delta_2/(\chi^2_b/(\chi^2_b - df_m))$. Simple algebra shows that this leads us back to Δ_1 and the associated properties of Δ_1. Values of Δ_2 much lower than one indicate a poor fit while those much greater than one may indicate overfitting. However, it is possible to overfit a model even if Δ_2 does not exceed one. For example, due to sampling fluctuations a valid model in a given sample could have a Δ_2 somewhat less than one. A researcher might add unnecessary free parameters in an attempt to improve Δ_2 and hence may overfit the data.

The rationale for Δ_2 assumes that $(N-1)F_m$ approximates a χ^2 variate. This is not true for the Unweighted Least Squares (ULS) fitting function. So, the justification for Δ_2 using the F_b and F_m from ULS in equation 4 is undermined. However, following Browne (1982), chi-square estimators for ULS solutions are available and these could be substituted into the chi-square version of Δ_2 (that is, equation 3).[2]

A more general formula for Δ_2 than equation 3 is:

$$\Delta_2 = \frac{\chi_r^2 - \chi_m^2}{\chi_b^2 - df_m} \qquad [5]$$

where χ_r^2 is the chi-square estimator for a model less restrictive than the baseline one but more restrictive than the maintained model. Bentler and Bonett's (1980) Δ_1 has an analogous general form, though in practice χ_r^2 is typically set to χ_b^2.

An index similar to Δ_1 is ρ_1 (Bollen, 1986):

$$\rho_1 = \frac{\left(\frac{\chi_b^2}{df_b}\right) - \left(\frac{\chi_m^2}{df_m}\right)}{\left(\frac{\chi_b^2}{df_b}\right)} \qquad [6]$$

$$= \frac{\left(\frac{F_b}{df_b}\right) - \left(\frac{F_m}{df_m}\right)}{\left(\frac{F_b}{df_b}\right)} \qquad [7]$$

ρ_1 is similar to the normed fit index (Δ_1), in that ρ_1 has a maximum of one and the mean of ρ_1's sampling distribution is larger for bigger samples than for smaller ones. Like the new Δ_2, ρ_1 adjusts for degrees of freedom. Conceptually, however, ρ_1 differs from Δ_2 in its definition of fit. The ρ_1 compares the ratios of the chi-square (or F) to degrees of freedom for the baseline and maintained models. The Δ_2 compares the chi-squares (or F values) for the same models.

Another incremental fit measure is that of Tucker and Lewis (1973). Originally their fit measure was for maximum likelihood factor analysis but Bentler and Bonett (1980) extended it to general structural equation models. This nonnormed fit index is

$$\rho_2 = \frac{\left(\dfrac{\chi_b^2}{df_b}\right) - \left(\dfrac{\chi_m^2}{df_m}\right)}{\left(\dfrac{\chi_b^2}{df_b}\right) - 1} \qquad [8]$$

$$= \frac{\left(\dfrac{F_b}{df_b}\right) - \left(\dfrac{F_m}{df_m}\right)}{\left(\dfrac{F_b}{df_b}\right) - (1/(N-1))} \qquad [9]$$

The ρ_2 is similar to ρ_1 in its adjustment for degrees of freedom and its use of chi-square (or F) to degrees of freedom to gauge fit. As with Δ_2 but unlike ρ_1 or Δ_1, it also adjusts for sample size so that for a correct model, the mean of its sampling distribution should be about one regardless of N. The ρ_2 has a similar relation to ρ_1 as the relation of Δ_2 to Δ_1.

There are two types of possible sample size influences on these fit measures. One is whether N directly enters the *calculation* of the fit index. This occurs if for fixed values of F_b, F_m, df_b, and df_m, altering the values of N alters the value of the fit measure. Bollen (1986) shows that ρ_2 is subject to this sample size effect while ρ_1 is not. Similarly, an examination of equations 2 and 4 reveals this sample size influence for Δ_2 but not for Δ_1.

The second sample size effect is whether the *mean of the sampling distribution* of a fit index is related to N. To understand this effect, suppose we have a valid model and for each sample size N, we draw a large number of independent, random samples of values for the variables in the model. Then for each of these samples, we estimate the model, calculate the fit index, and form the sampling distribution of the fit indices. Repeating these steps for many different sample sizes, we can analyze how the sampling distributions of the fit index changes with respect to N. If the mean depends on N, we have the second sample size effect. The means of the sampling distributions of Δ_1 and ρ_1 seem to be positively related to N while this dependence should be near zero for Δ_2 and ρ_2.

Finally, I consider the probability limits (*plim*) of these measures. Provided that $\text{plim}(F_b)$ and $\text{plim}(F_m)$ exist:

$$\text{plim}\,(\Delta_1) = c \quad [10]$$

$$\text{plim}\,(\Delta_2) = c \quad [11]$$

$$\text{plim}\,(\Delta_1 - \Delta_2) = 0 \quad [12]$$

where c is some constant and $0 \leq c \leq 1$. Equations 10 and 11 follow from equations 2 and 4 whereas equation 12 follows from equations 10 and 11. If the maintained model is valid, c equals one. An implication of equation 12 is that the difference between Δ_1 and Δ_2 decreases as the sample becomes larger.

Similarly we have:

$$\text{plim}\,(\rho_1) = d \quad [13]$$

$$\text{plim}\,(\rho_2) = d \quad [14]$$

$$\text{plim}\,(\rho_1 - \rho_2) = 0 \quad [15]$$

where d is a constant which equals one when the maintained model is correct.

Another large sample property of Δ_2 and ρ_2 is that the chances of these being greater than one decreases as N grows larger (see equations 4 and 9).

SIMULATION ILLUSTRATION

The exact sampling distributions of Δ_2 and the other incremental fit indices are unknown. A small Monte Carlo simulation provides some evidence on the means and standard deviations of the sampling distribution of the new Δ_2 compared to those of Δ_1, ρ_1, and ρ_2.

A Confirmatory Factor Analysis (CFA) model is the true model:

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

where $\mathbf{x}$ is a 5×1 vector of observed variables, $\boldsymbol{\xi}$ is a 2×1 vector of latent variables, Λ_x is a 5×2 matrix of factor loadings, and $\boldsymbol{\delta}$ is a 5×1 vector of errors of measurement. The model parameters are

$$\Lambda_x = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \\ 1 & 2 \end{bmatrix}$$

$$\operatorname{diag} \Theta_\delta = [1 \;\; 1 \;\; 1 \;\; 1 \;\; 1]$$

where $\boldsymbol{\Phi}$ is the covariance matrix of $\boldsymbol{\xi}$, and $\boldsymbol{\Phi}_\delta$ is the covariance matrix of $\boldsymbol{\delta}$.

All the x's are normally distributed. The simulation includes three different sample sizes (75, 150, 300) and fourteen replications for each sample size.[3] Table 1 reports the means and standard deviations (s.d.) for each fit index using the maximum likelihood estimator (F_{ML}).

The positive association between the means of the sampling distributions of Δ_1 and ρ_1 with N is evident in this table. The averages of Δ_2 and ρ_2 stay near one for all sample sizes. It is unclear whether the decline in the third decimal place of Δ_2 and ρ_2 is

TABLE 1
Simulation Results for CFA Model for Three Sample Sizes (75, 150, 300)

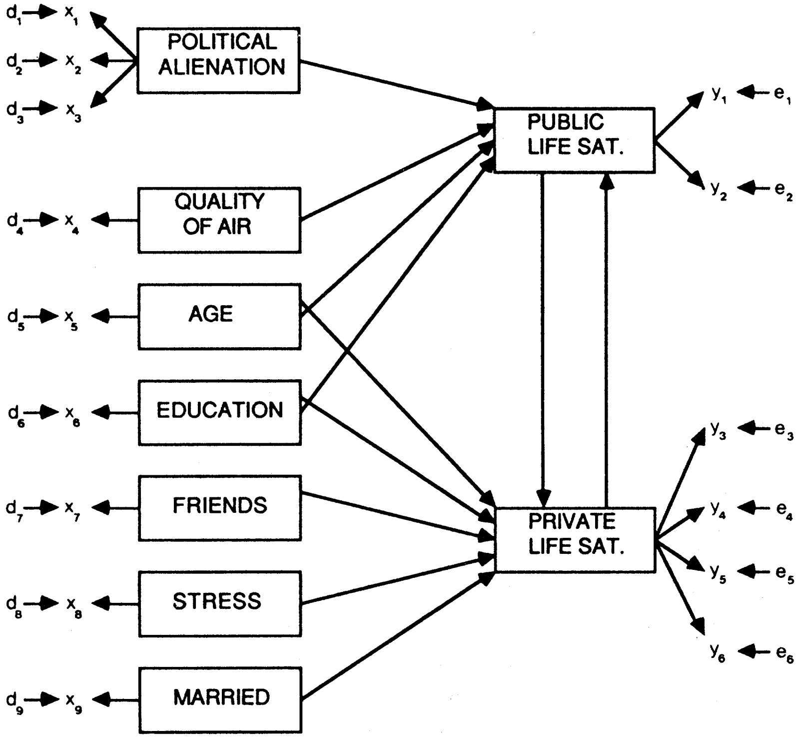
Incremental Fit Index	Mean Value* (standard deviation)		
	sample size		
	75	150	300
Δ_2	1.003 (.019)	1.002 (.012)	1.000 (.008)
Δ_1	.967 (.018)	.985 (.012)	.992 (.008)
ρ_2	1.008 (.052)	1.005 (.030)	1.001 (.021)
ρ_1	.918 (.046)	.963 (.030)	.980 (.020)

*Fourteen replications for each sample size.

meaningful. The s.d.'s of the Δ's are less than half those of the ρ's. This is important since at least in the case of ρ_2, Anderson and Gerbing (1984) report a propensity of ρ_2 to have excessive variability and a tendency to produce outliers. Finally, we see that the differences ($\Delta_2 - \Delta_1$) and ($\rho_2 - \rho_1$) grow smaller as N increases.

EMPIRICAL EXAMPLE

The simulation examined the fit indices for a simple model in small to moderate sized samples. The empirical example treats more complex models in larger samples. In addition, in the simulation we knew that the model was valid while in the empirical example the models are subject to unknown errors. The example is Wheaton's (1987) model of the reciprocal relation between the latent variables of public and private life satisfaction. Figure 1 reproduces Wheaton's path diagram for this model. Public life satisfaction is measured with two indicators while private satisfaction has four indicators. Seven exogenous variables each have



NOTE: x_1 = no control over Parliament; x_2 = elected officials don't care; x_3 = no say in what politicians do; x_4 = clean air in local area; x_5 = age in years; x_6 = education in years; x_7 = Number of close friends; x_8 = number of stressors this year; x_9 = married; y_1 = satisfaction with city/town; y_2 = satisfaction with life in Canada; y_3 = overall life satisfaction; y_4 = satisfaction with leisure; y_5 satisfaction with finances; y_6 = satisfaction with relationship.

Figure 1: An Initial Model for the Reciprocal Effects of Public and Private Life Satisfaction

one indicator except for political alienation, which has three. Political alienation and quality of the air in the local area are exclusive determinants of public life satisfaction. The number of close friends, the number of life stressors experienced in the last year, and being married are exclusive causes of private life satisfaction. The last two exogenous variables, age and education, influence both types of satisfaction.

TABLE 2
Wheaton's (1987) Models of Reciprocal Effects of Public and Private Life Satisfaction

Model (df)	N=355					N=2568				
	χ^2	Δ_1	Δ_2	ρ_1	ρ_2	χ^2	Δ_1	Δ_2	ρ_1	ρ_2
M_0(111)	691	--	--	--	--	5389	--	--	--	--
M_1(108)	419	.39	.47	.38	.45	2610	.52	.53	.50	.51
M_2(102)	300	.57	.66	.53	.63	1904	.65	.66	.62	.63
M_3(95)	239	.66	.76	.60	.71	1221	.77	.79	.74	.75
M_4(84)	142	.79	.90	.73	.87	673	.88	.89	.84	.85
M_5(83)	103	.85	.97	.80	.95	441	.92	.93	.89	.91

where:

M_0 = "complete null"
M_1 = "restrictive informed null"
M_2 = "less restrictive informed null"
M_3 = "causal null only"
M_4 = "basic causal null"
M_5 = M_4 plus "marriage effect"

Wheaton (1987) estimates a series of nested models ranging from extremely restrictive null ones to less restrictive elaborations of the original model. The original sample of 2,568 individuals and a random subsample of 355 cases are fitted for each model. Table 2 contains the chi-square estimates, degrees of freedom, Δ_1, Δ_2, ρ_1, and ρ_2 for the six models which Wheaton (1987) estimated for both samples. The labels he uses to identify these models are at the base of Table 2. The M_o is the baseline model for all the Δ's and ρ's. Wheaton provides an extensive discussion of the behavior of Δ_1, ρ_2 and other fit indices with other baselines so I will not repeat his description here. Rather I focus on Δ_2 and ρ_1, not reported by him, and their relations to Δ_1 and ρ_2.

One result is that for the same model, Δ_1 always is larger in the big sample than in the small one. For instance, model M_5 has a $\Delta_1 = .92$ for an N of 2,568 but $\Delta_1 = .85$ for an N of 355. Similarly ρ_1 is consistently larger for the big sample when comparing the same model. In contrast the Δ_2's tend to be closer for both sample sizes with their biggest differences for M_1 and M_5. We see the same pattern for the ρ_2's. As expected, the $(\Delta_2 - \Delta_1)$ and $(\rho_2 - \rho_1)$ differences for the same model are smaller for N = 2,568 than for N = 355.

We also see that ρ_1 is consistently the most conservative of the incremental fit indices. Overall Δ_2 behaves much like ρ_2. The simulation evidence from the prior section suggests that Δ_2 would have a smaller standard deviation than ρ_2 or ρ_1.

SUMMARY

I have proposed a new incremental fit index, Δ_2, which adjusts the normed fit index for sample size and for the degrees of freedom of the maintained model. It is designed so that on average it is about one for correct models. This contrasts with Δ_1 and ρ_1 which tend to increase with sample size. I recommend that researchers calculate Δ_2 with Δ_1 and ρ_2 with ρ_1.

The simulation results revealed that the mean of the sampling distribution of Δ_2 (and ρ_2) stay near one for different sample sizes with Δ_2 and Δ_1 having standard deviations nearly half those of ρ_2 or ρ_1. These results parallel simulation work by Bollen and Stine (1988) suggesting that the findings generalize beyond the particular model chosen. The empirical example further demonstrated the properties of these fit measures. Of course, far more analytic and simulation work is required to understand the properties of the incremental fit indices but these preliminary results suggest that Δ_2 has some desirable features.

Finally, let me emphasize that Δ_2 or any other fit measure alone cannot determine whether a model is valid. It is just one indicator that goes into assessing model fit. Factors such as the standards of fit set by other analyses of similar data, the distance of the fit measure from the ideal value, sampling variability, the choice of the baseline model, principles of parsimony, and the reasonableness of parameter estimates must be considered in evaluating a model's fit.

NOTES

1. I assume that (N - 1) times the value of the fitting function forms a chi-square estimate such as is true for the most common estimators (e.g., Maximum Likelihood, Generalized Least Squares).

2. The same situation occurs for ρ_2 that I present later.

3. The simulation used the random number generator function NORMAL from SAS to create the variables with the specific structure. My thanks to the participants in the seminar "Structural Equations with Latent Variables" who helped to create the simulation data.

REFERENCES

ANDERSON, J. and D. W. GERBING (1984) "The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis." Psychometrika 49: 155-173.

BEARDEN, W. O., S. SHARMA, and J. E. TEEL (1982) "Sample size effects on chi-square and other statistics used in evaluating casual models." J. of Marketing Research 19: 425-430.

BENTLER, P. M. and D. G. BONETT (1980) "Significance tests and goodness-of-fit in the analysis of covariance structures." Psych. Bull. 88: 588-600.

BOLLEN, K. A. (1986) "Sample size and Bentler and Bonett's nonnormed fit index." Psychometrika 51: 375-377.

BOLLEN, K. A. and R. STINE (1988) "Bootstrapping structural equation models: variability of indirect effects and goodness of fit measures." Presented at the 1988 Convention of the American Sociological Association, Atlanta, GA.

BROWNE, M. (1982) "Covariance structures," pp. 72-141 in D. M. Hawkins (ed.) Topics in Multivariate Analysis. New York: Cambridge.

HOELTER, J. (1983) "The analysis of covariance structures: goodness-of-fit indices." Soc. Methods & Research 11: 325-344.

JÖRESKOG, K. G. and D. SÖRBOM (1986) LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods. Mooresville, IN: Scientific Software.

STEIGER, J. H., A. SHAPIRO, and M. W. BROWNE (1985) "On the multivariate asymptotic distribution of sequential chi-square statistics." Psychometrika 50: 253-264.

TUCKER, L. R. and C. LEWIS (1973) "A reliability coefficient for maximum likelihood factor analysis." Psychometrika 38: 1-10.

WHEATON, B. (1987) "Assessment of fit in overidentified models with latent variables." Soc. Methods & Research 16: 118-154.

Kenneth A. Bollen is an Associate Professor of Sociology at the University of North Carolina at Chapel Hill. His major research interests are in international development and statistics. His book titled Structural Equations with Latent Variables *will be published in John Wiley's Applied Probability and Statistics series.*