DOCUMENT RESUME

ED 267 091 TM 860 167

AUTHOR Marsh, Herbert W.; Balla, John R.
TITLE Goodness-of-Fit Indices in Confirmatory Factor Analysis: The Effect of Sample Size.
PUB DATE 26 Feb 86
NOTE 25p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Elementary Secondary Education; *Factor Analysis; *Goodness of Fit; Hypothesis Testing; *Mathematical Models; Questionnaires; *Sample Size; Self Concept Measures; Statistical Studies; Student Evaluation of Teacher Performance
IDENTIFIERS *Confirmatory Factory Analysis; *LISREL Computer Program; Self Description Questionnaire; Students Evaluation of Educational Quality

ABSTRACT

This investigation examined the influence of sample size on different goodness-of-fit indices used in confirmatory factor analysis (CFA). The first two data sets were derived from large normative samples of responses to a multidimensional self-concept instrument and to a multidimensional instrument used to assess students' evaluations of teaching effectiveness. In the third set, data were simulated and generated according to the model to be tested. In the fourth, data were simulated and generated according to a three-factor model that did not have a simple structure. Twelve fit indicators were used to assess goodness-of-fit in all CFAs. All analyses were conducted with the LISREL V package. One-way ANOVAs and a visual inspection of graphs were used to assess the sample size effect on each index for the four data sets. Despite the inconsistency of the findings with previous claims, the results are consistent with the observation that the amount of random, unexplained variance varies inversely with sample size. Appendices include a set of computed statements, an explanation and listing of the 12 goodness-of-fit indicators, a bibliography, a table of results, and figures showing sample size effect. (Author/LMO)

Goodness-of-fit Indices In Confirmatory Factor Analysis: The Effect of Sample Size

Herbert W. Marsh and John R. Balla

University of Sydney, Australia

26 February, 1986

Running Head: Goodness of Fit

Goodness-of-fit Indices In Confirmatory Factor Analysis: The Effect of Sample Size

ABSTRACT

The present investigation examines the influence of sample size on different goodness-of-fit indices used in confirmatory factor analysis (CFA). Contrary to Bentler and Bonett (1980), their incremental fit index was substantially affected by sample size. Contrary to Joreskog and Sorbom (1981), their goodness of fit indices provided by LISREL were substantially affected by sample size. Contrary to Hoelter (1983), his critical N index was substantially affected by sample size. Of the 12 indices considered, only the Tucker-Lewis index and a conceptually similar new index were relatively independent of sample size, and these results were consistent across two real and two simulated sets of data. Despite the inconsistency of these findings with previous claims, the results are consistent with the observation that the amount of random, unexplained variance varies inversely with sample size.

Goodness-of-fit Indices In Confirmatory Factor Analysis: The Effect of Sample Size

The purpose of the present investigation is to examine the influence of sample size on different goodness-of-fit indicators used in confirmatory factor analysis (CFA). While the present investigation is limited to CFA, the problems, issues and results also generalize to the analysis of covariance structures. The advantages of the use of CFA are well known and numerous introductions to the LISREL approach used in the present investigation are available elsewhere (e.g., Bagozzi, 1980; Joreskog & Sorbom, 1981; Long, 1983; Marsh & Hocevar, 1985; Pedhauzur, 1982). Briefly, in CFA, responses to p observed variables by N subjects are summarized by a (p x p) sample covariance matrix and it is hypothesized that the corresponding population covariance matrix can be summarized by q true but unknown parameters (Bentler & Bonett, 1980). The q parameters in the present investigation are the factor loadings, the factor variances and covariances, and the error/uniquenesses. To the extent that the inferred population covariance E derived from these parameters corresponds to the observed sample covariance matrix S, the model is supported. The problem of goodness of fit is how to decide whether E is sufficiently similar to S to justify the conclusion that a specific model adequately fits a particular set of data (Hoelter, 1983).

In the maximum likelihood approach to CFA a discrepancy or loss function is minimized with respect to the q parameters such that its value approaches zero as the S and E become identical (Bentler & Bonett, 1980; Joreskog & Sorbom, 1981). If all the observed variables have a multivariate normal distribution, if the sample size is large, and if the model is correct, then N - 1 times the minimum value of the loss function can be interpreted as a X^2 test statistic with degrees of freedom (df) equal to .5 x p x (p + 1) - q. As typically used the model is rejected if the X^2 is large relative to the df, and accepted if the X^2 is nonsignificant or small. However, Bentler and Bonett (1980) warn that the probability of detecting a false model increases with N even when the model is minimally false (i.e., differences between E and S are trivial) so that for very large sample sizes nearly all models are rejected. (It is important to note that for a true model the expected value of X^2 is equal to the df and does not vary with sample size.) Because of the apparently restrictive assumptions underlying the X^2 test statistic and because of the power of the test for large sample sizes, Marsh and Hocevar (1985) concluded that "most applications of confirmatory factor analysis require a subjective evaluation of whether or not a statistically significant chi-square is small enough to constitute an adequate fit" (p. 567), that this subjectivity undermines some of the rigor that is possible with CFA, and that "this issue will continue to be an important one in the future development of this statistical procedure" (p.568).

Bentler and Bonett (1980) note the dubious logic of inferring support for a model from a nonsignificant X^2 (i.e., attempting to prove the null hypothesis) since the X^2 can be made small simply by reducing the sample size. Because of this influence of sample size, a poor fit based on a small sample size may result in nonsignificant X^2, whereas a good fit based on a large sample size may result in statistically significant X^2. The substantial influence of sample size on X^2 for a false model may lead to a counter-productive practice in CFA. In order to obtain a nonsignificant X^2, or at least one that is acceptably low, researchers may be tempted to limit analyses to a small number of cases, or if their sample size is large to analyze only a subsample of their data. For example, Bentler and Bonett (1980, p. 591) assert that "one's favorite model will stand the best chance of being accepted when tested against the data of small samples" and Fornell (1983, p. 446) suggests that one could "make sure the sample size is not large enough to reveal a difference between the model and the data". As emphasized by these authors, and by common sense, such a practice is counter-productive and Hoelter (1983, p. 328) stated that:

> "sacrificing the power of a test by utilizing small sample sizes simply blinds the researcher to significant differences between a model and the data" and
>
> "testing models with large samples is always desirable, and the question that needs to be addressed deals with how well a model <u>approximates</u> the observed data rather than whether of not the model <u>fits</u> the data."

Because of the substantial influence of sample size on X^2, researchers have developed a variety of goodness-of-fit indices that they claim to be unaffected by sample size, and these are among the fit indices considered in the present investigation.

<u>Goodness of Fit Indicators.</u>

<u>X^2 and X^2/df Ratio.</u> These two indices continue to be the most frequently used goodness-of-fit indicators. As noted above, the X^2 for a false model varies directly with sample size, but the X^2 for a true model does not. In CFA the df does not vary with the sample size, so that the effect of sample size on the X^2/df must necessarily be the same as for the X^2. Hence, while these indicators do not vary with sample size for a true model, they are substantially influenced by sample size when the model is false and this dependence on sample size is larger for poorer models.

Incremental Fit Indices. Bentler and Bonett (1980) proposed that valuable information could be obtained by comparing the ability of nested models to fit the same data. In particular, for CFA it is useful to compare the fit of a null model in which all the p variables are assumed to be uncorrelated with the fit of the proposed model. If the fit of a null model is reasonable, because the sample size is small or because the measured variables are relatively uncorrelated, then the fit of target model will automatically be reasonable. However, if the increment in the goodness-of-fit is small, then there may be no basis of support for the model. Bentler and Bonett (1980) presented two incremental fit indices. First they described the Tucker-Lewis Index (TLI) and a more general version of this index. Second, they proposed an alternative index, called the Bentler-Bonett Index (BBI) for purposes of the present investigation. They specifically noted that these indices are useful for comparing the fit of a particular model across samples that have unequal sizes. They cautioned that the absolute value of these indices may be difficult to interpret, but that values of less than .9 usually mean that the model can be improved substantially. Much of the value of these indices is based on the assumption that they are independent of sample size, and this assumption will be tested in the present investigation.

Other fit indices computed by LISREL. Joreskog and Sorbom (1981) describe three other indices that are computed by LISREL: the goodness-of-fit index (GFI), the adjusted GFI (AGFI), and the root mean square residual (RMS). They describe GFI as "a measure of the relative amount of variances and covariances jointly accounted for by the model" and assert that "unlike X^2, GFI is independent of the sample size" while the adjusted GFI (AGFI) "corresponds to using mean squares instead of total sums of squares" (Joreskog & Sorbom, 1981, p. I. 40-41). The RMS, based on the differences between the elements in the observed and inferred covariance matrices, is easily interpreted when correlation matrices are analyzed. For the examination of covariance matrices RMS still has a lower-bound of zero but its upper bound is indeterminant. Thus RMS must be interpreted in relation to the size of the variances and covariances of the measured variables, and cannot be compared across applications based on different variables. Joreskog and Sorbom suggest that GFI and AGFI will generally fall between 0 and 1, but that it is possible for them to be negative. Again, much of the value of these indicators is based on the assumption that they are independent of sample size, and this assumption will be tested in the present investigation.

Critical N. Hoelter (1983, p. 528) argued that "rather than ignoring or completely neutralizing sample size we can estimate the size that a sample must reach in order to accept the fit of a given model on a statistical basis. This estimate, referred to here as 'critical N' (CN), allows one to assess the fit of a model relative to identical hypothetical models estimated with different sample sizes." Hoelter cautioned that no firm basis could be offered as to what constituted an adequate fit, but he suggested that a value of 200 was a reasonable starting point for suggesting that differences between the model and data may be unimportant. The usefulness of CN also rests on the assumption that its value is independent of sample size, and this assumption will be tested in the present investigation.

<u>The Sobel and Bohrnstedt approach.</u> For purposes of the present investigation only two models, a null model and a target model, are fit to each covariance matrix. However, as emphasized by Sobel and Bohrnstedt (1985), by Bentler and Bonett (1980), and by Marsh and Hocevar (1984; 1985) in most substantive applications a variety of nontrivial models can be generated on the basis of previous research or the logic of the data. Often these nontrivial models will differ from each other according to strict statistical criteria, but differences may not be of practical importance. Goodness-of-fit indicators in which each of the competing nontrivial models is compared to the null model may, perhaps, provide an externally meaningful, well defined, absolute scale for determining whether the statistically significant differences are of practical importance (e.g., Marsh & Hocevar, 1984; 1985).

Sobel and Bohrnstedt (1985) contend that the use of the null model to determine one end-point for incremental fit indices is only appropriate for exploratory studies in which nontrivial alternative models are unavailable. Instead, they argue that a more parsimonious nontrivial model should be used to define the lower-end of the goodness-of-fit index. However, there are two important problems with their approach. First, the use of the null model to define one end-point of goodness-of-fit indices does not preclude the comparison of alternative nontrivial models and need not be limited to exploratory studies (e.g., Marsh, 1985; Marsh & Hocevar, 1984; 1985). Second, and more importantly, if their recommendation were followed then the scale of their goodness-of-fit indices would be arbitrarily determined by the idiosyncratic choice of nontrivial models in a particular study, thus precluding inferences of practical significance of differences between competing nontrivial models on a nonarbitrary scale. Furthermore, for the incremental fit indices they considered and considered here, the value obtained from the Sobel/Bohrnstedt approach is just a linear transformation of the one based on the null model. Whether the promise of an externally meaningful, well-defined, absolute scale for goodness-of-fit indicators can be fulfilled by existing fit indices is an empirical question that is addressed in part by the present study, but the promise cannot be fulfilled if the Sobel/Bohrnstedt approach is adopted. It should also be noted that if the sample size affects goodness of fit as used in the present approach based on the null model, it must also affect goodness of fit as defined with the Sobel/Bohrnstadt approach.

The Present Investigation.

For purposes of the present investigation, it is proposed that an ideal goodness-of-fit index should:

1) be relatively independent of sample size;

2) be valid, that is accurately and consistently reflect differences in goodness of fit for competing models of the same data (e.g., nested models) and for the same model applied to different data;

3) vary along an externally meaningful, well-defined, absolute continuum so that its value can be easily interpreted.

The present investigation emphasizes the examination of the first criterion, but its violation would also undermine the second and particularly the third criteria. In order to examine the first criterion the same three-factor, simple-structure model was fit to data from four data sets. The first two data sets, based on real data, were derived from large normative samples of responses to a multidimensional self-concept (SC) instrument and to a multidimensional instrument used to assess students' evaluations of teaching effectiveness (SET). Based on the a priori design of each instrument, the selection of variables, and the results of previous research, a three-factor model is reasonable. However, only the SC data approximates a simple structure so that it should be better fit by the model than should the SET data.

The third and fourth data sets are based on simulated data. In the third, data were generated according to the model to be tested, so that E and S differ only due to random chance. In the fourth, data were generated according a three-factor model that did not have a simple structure. Hence, goodness of fit indices should be better for the third data set than for the fourth, and also better for the third than either of the first two.

For each of the four data sets, ten random samples were generated with sample sizes of 25, 50, 100, 200, 400, 800, and 1600, and the same three-factor model was fit to 280 covariance matrices derived from the four data sets. To the extent that the values of a particular index of fit are similar across the seven sample sizes, then there is support for its independence of sample size.

Method

The CFA Model and Analyses.

All analyses were conducted with the commercially available LISREL V package (Joreskog & Sorbom, 1981). The target model in each of the analyses contained 21 estimated parameters: six factor loadings (in LAMBDA Y), three factor variances (in PSI), three factor covariances (in PSI), and nine error/uniquenesses (in THETA EPSILON). The first measured variable for each factor was designated to be a reference indicator and given a factor loading of 1.0, while loadings for the other two variables were estimated. Hence the df (.5 x 9 x 10 - 21 = 24) was constant for all the analyses. In addition to the target model, a null model was tested for each covariance matrix such that the reproduced covariance matrix was a diagonal matrix of variances and the nine measured variables were posited to be uncorrelated. The df for the null model (.5 x 9 x 10 - 9 = 36) was also constant for all the analyses. The same null and hypothesized models were tested for each of 280 covariance matrices described below. (Also, in order to test the generality of findings to be described latter, one additional set of analyses was conducted on a 14-variable, 4-factor model described latter).

The Data.

The Sample Sizes. The seven sample sizes to be considered in the present investigation, 25, 50, 100, 200, 400, 800 and 1600, were selected to span the range of sample sizes typically considered in CFA. Hoelter (1983) and Bentler and Bonett (1980) each considered a sample size of 23 for a similar purpose, and so 25 appeared to be a reasonable lower limit. The upper limit of 1600 was arbitrary, but it is apparently larger than the sample sizes typically used in CFA. (Also, in order to test the generality of findings to be described latter, one additional set of analyses was conducted with a sample of 32,000).

Self-concept (SC) data. The SC data came from the normative archive of over 4,000 sets of responses to the Self Description Questionnaire (Marsh, 1986). For purposes of the present investigation the first three indicators were selected from the Physical Ability, Reading, and Math scales. Previous research (e.g., Marsh & Hocevar, 1985; Marsh, Smith & Barnes, 1985) suggests that the simple structure model should provide a reasonable fit for this data. A random sample of 3175 cases was divided into sets of 25, 50, 100, 200, 400, 800, and 1600, and this process was repeated 10 times so that a total of 70 covariance matrices were created. The hypothesized and null models were fit to the a 9 x 9 covariance matrix derived from this data set.

Students' evaluations of teaching (SET). This data came from the set of 4,471 sets of class-average responses to the Students' Evaluation of Educational

Quality instrument that was factor analyzed by Marsh (1983; 1984). For purposes of the present investigation the first three items were selected from the Leaning/Value, Organization, and Workload/Difficulty scales. Previous factor analyses suggest that while the three factors are well defined, several of the items load on more than one factor so that solution is not "simple." As with the SC data, the null and hypothesized model were fit to 70 covariance matrices.

Simple structure simulated data (SSSD). The nine measured variables were defined with the random number generator from the commercially available SPSS package (Hull & Nie, 1981). Each variable was defined to reflect only one factor and a normally distributed random error component, and the three factors were defined to be correlated (see Appendix I). A total a 31,750 cases were generated and divided into 70 sets of data such that each sample size was represented by 10 covariance matrices, and the null and hypothesized models were fit to the 70 covariance matrices.

Complex Structure Simulated Data (CSSD). The nine measured variables were defined as with the SSSD except that six of the nine measured variables -- two for each factor -- were defined such that each should have had a small loading on one factor in addition to the one it was designated to reflect (see Appendix I). Again a total of 31,750 cases were generated and divided into 70 sets of data such that each sample size was represented by 10 covariance matrices, and the null and hypothesized models were fit to the 70 covariance matrices.

Goodness of Fit Indicators.

Six goodness-of-fit indicators -- X^2, X^2/df, GFI, AGFI, RMS, and CN -- were examined that did not require results from a corresponding null model. Two forms of an incremental fit index described by Bentler and Bonett (1980) were used to reflect a scaled difference between the goodness of fits for the null and hypothesized models. One form is illustrated by the BBI; the difference in X^2 for the Null and tested model divided by the X^2 for the null model. Using this approach, new incremental fit indices (see Appendix II) were derived from X^2/df (called MB1), RMS (called incremental RMS), and CN (called incremental CN). The range of values for these indices is strictly bounded by 0 and 1. A second form is illustrated by the TLI; the difference in X^2/df for the Null and posited models is divided by the difference between the X^2/df for the Null model and some ideal expected value -- in this case 1.0. Using this approach a new incremental index was derived from the X^2 (called MB2). For these indices, the range of expected values is 0 to 1.0. Technically it is possible for these indices to be slightly negative (if the X^2/df for the null model is less than 1.0) or slightly larger than 1.0 (if the X^2/df for the target model is less than 1.0) but such occurrences will be very rare for real data tested with a priori models. The set of 12 fit indicators, as defined in Appendix II, was used to assess the goodness of fit in all 280 CFAs.

Results & Discussion

One-way ANOVAs (Table 1) and a visual inspection of graphs (Figure 1) were used to assess the sample size effect on each index for the four data sets. The effect of sample size varies dramatically depending on the index. Of the 12 indices considered, only the TLI and the conceptually similar MB2 are relatively independent of sample size. For the other 10 indices, the sample size effect varies with both the sample size and the data set. For 7 of these 10 indices -- all but X^2, X^2/df, and CN -- most of the sample size effect occurs for the smallest sample sizes (25, 50, 100 & 200). The sample size effect for all but TLI and MB2 also appears to vary with the data being considered. In particular, the sample size effect for all the indices except X^2, X^2/df, TLI and MB2 are substantially larger for the SSSD data that was simulated to be best fit by the model than for the SC and/or SET data.

Insert Table 1 and Figure 1 About Here

Though not the primary focus of the present investigation, the four data sets were constructed so as to vary in their ability to be fit by the target model. In particular, the SSSD data were simulated to be best fit by the model, and the SC data were selected so as to be better fit by the model than were the SET data. This ordering is accurately reflected by X^2, X^2/df, GFI, AGFI, CN, TLI and MB2, but not by BBI, MB1, incremental RMS, incremental CN -- all of which are incremental indices. (Note that the RMS was not considered since it is dependent on the size of variances and covariances in the particular application and so is clearly not comparable across applications unless analyses are based on the correlation matrices.) It was also assumed that the fit for the target model should always be better than for the corresponding null model. However, AGFI and occasionally GFI had negative values for the target model whereas the corresponding null models always had positive values. Thus, at least for these cases, GFI and AGFI did not accurately reflect differences in the ability of the two models to fit the same data. These findings suggest that the five incremental indices other than TLI and MB2, and perhaps GFI and AGFI, in addition to being substantially influenced by sample size, may not validly reflect real differences in goodness of fit.

Contrary to claims by Bentler and Bonett (1981), their BBI is substantially related to sample size in all four data sets. Contrary to claims by Joreskog and Sorbom (1981), their GFI and AGFI are both substantially affected by sample size in all four data sets (as well as the RMS that is also provided by LISREL). Contrary to claims by Hoelter (1983), his CN is substantially related to sample size in all four data sets. Furthermore, all the other indices except for TLI and MB2 were also substantially affected by sample sizes. While the empirical findings are clear, two important questions remain: Why are BBI, GFI, AGFI, RMS, and CN so substantially affected by sample size? Why are the TLI and MB2 relatively unaffected by sample size?

The substantial effect of sample size on BBI, GFI, AGFI, RMS, and CN is consistent with the observation that the standard errors of the observed variances and covariances becomes smaller as sample sizes become larger. Data to be fit contains variance that can be explained by the target model (explained variance), systematic variance that cannot be explained by the model (uniqueness), and random variance that cannot be fit by the model (error). The proportion of random error decreases systematically with sample size (i.e., standard errors become smaller), so that the proportion of variance that can be explained and the uniqueness must increase with sample size. The relation between explained and random variance is most clear for the SSSD data since it was created to have no uniqueness. Thus this explanation of the sample size effect can be examined most easily for this data.

For the SSSD data, the RMS reflects only random variance and the amount of random variance is inversely related to sample size. Thus RMS also varies inversely with sample size. The BBI is based on the ratio of the X^2s for the null and target model, but for the SSSD data the X^2 for the null model varies substantially with sample size while the X^2 for the target model does not. Thus BBI must also vary with sample size. CN is based on the unstated assumption that the proportion of random variance will not vary with sample size, but this assumption is false. For the SSSD data with n=25 CN is small (41) because the proportion of random variance is relatively large, whereas for n=1600 CN is much larger (3114) because the proportion of random variance is relatively small. If the SSSD were tested with a sample size of 41 and the proportion of random variance was the same as it was in a sample size of 25, then Holter's claim of the independence of CN and sample size would be satisfied. However, as empirically demonstrated in the present investigation, the claim is false, and this is because the proportion of random variance will not be the same in samples of 25 and 41. [Note also that if a covariance matrix based on 1600 cases is used but LISREL is told that it is based on only 100 cases, a much smaller X^2 will be obtained than if a covariance based on a random sample of 100 of the 1600 cases is used.] GFI and AGFI, as measures of the relative variance that is accounted by the target model, will also vary with sample size since the amount of random variance that is unexplained by the model varies with sample size. Hence, RMS, GFI, AGFI, CN, and BBI all vary with sample size because the proportion of random variance varies with sample size. For real data the situation is complicated by the existence of uniqueness in addition to random and explained variance. However, even though it may be difficult to separate random error and uniqueness, the proportion of random variance still depends on sample size so that the logic of the present explanation generalizes to these data as well.

A comparison of the BBI and MB1 with TLI and MB2 suggests why the first two are so much more affected by sample size. Both sets of indices have the same numerators (see Appendix II), but the denominator of BBI and MB1 is based on the fit of the null model whereas the denominator of the TLI and MB2 is the difference between the null model and the expected value of the X^2 or X^2/df. In the present study these expected values -- df for the X^2 or 1.0 for the X^2/df -- are constant, but the size of this constant in relation to the value for the null model varies with sample size. For small sample sizes this constant is relatively larger compared to the value for the null model, and this provides some control for the sample size effect. Additional research is clearly warranted to further examine this suggested explanation and to test the generality of the findings in different applications.

<u>The Generality of the Sample Size Effect.</u>

Based on findings of this study two features of the sample size effect are clearly illustrated. First, the sample size effect is small or negligible for TLI and MB2, but is substantial for the other 10 fit indices. Second, for 7 of these 10 indices -- all expect X^2, X^2/df and CN -- the sample size effect is weaker, though still statistically significant, for the larger sample sizes (400, 800 & 1600). However, the findings are limited to seven sample sizes and to a single target model, and these limitations to the generality of the findings are worrisome. Consequently, two additional sets of analyses were conducted to test the generality of these findings. First, one additional sample based on 32,000 cases from the CSSD data -- 20 times the largest number of cases considered in the present investigation -- was fit with the same target model[1]. Second, a new data set like the CSSD was constructed for 14 measured variables designed to reflect four factors, and the sample size effect on the different fit indices was determined for this new data set[2].

The purpose of the 32,000-case analysis was to determine the behavior of the fit indices when the sample size was extremely large. Although the interpretation of the values is difficult for many of the indices that are approaching their optimum values for n=1600, only the TLI and MB2 appeared be unaffected by this large jump in sample size. The X^2 and X^2/df for the 32,000 cases were substantially larger (poorer) compared to values for 800 and 1600 cases, CN was substantially larger (improved), and the other indices -- except TLI and MB2 -- were marginally improved (see footnote 1). These findings support the generality of the findings summarized earlier, but do not indicate whether or not the indices will continue to change as sample size approaches infinity.

The purpose of the new data set with 14 variables was to explore the generality of the sample size effect found with the 9 variable model. While the sample size effect was statistically significant for all fit indices, the size of the effect was substantially smaller for the TLI and MB2 than for any other fit indices (see footnote 2). Furthermore, the group based on samples sizes of 25 was the only group to differ significantly from any other group for TLI and MB2, whereas nearly all possible pair-wise comparisons between different samples sizes were statistically significant for the other indices. These results also support that the generality of the findings summarized earlier.

Summary

The promise of an externally meaningful, well-defined, absolute scale does not appear to be fulfilled by most of the goodness-of-fit indices considered in the present investigation. First, when the variables to be fit and the model to be tested are held constant, values for 10 of the 12 fit indices are substantially affected by sample size. Furthermore, this sample size effect cannot be easily characterized since it varies depending on the particular index, the data set that is bei considered, and the range of sample sizes being considered. Second, when the model to be tested and the sample size are held constant, the fit indices may not be comparable across different data sets. In particular, the SSSD data was constructed so that it would be best fit by the target model, but of the incremental fit indices, only the TLI and MB2 accurately reflected this difference. Since only the TLI and MB2 performed satisfactorily in this investigation, no absolute criteria of what value constitutes an acceptable fit seems justified for any of the other indices (e.g., the .90 suggested by Bentler & Bonett for BBI, the 200 suggested by Hoelter for CN, or the value of 2.0 sometimes suggested for X^2/df). On the basis of this investigation it is recommended that: a) additional research is conducted with the TLI and MB2 to further test their characteristics; and, b) pending the outcome of further research, at least one of these two indices should be used, along with the X^2 test of statistical significance and the examination of parameter estimates in relation to substantive issues, to assess goodness of fit.

Footnotes

[1] -- A single 9 x 9 covariance matrix was derived from a new sample of 32,000 cases for the CSSD data (see Appendix I) and was fit with the same 3-factor model as the other 280 covariance matrices. The values for the 12 goodness-of-fit indices defined in Appendix II are as follows: X^2 (3155), X^2/df (31.5), GFI (.970), AGFI (.943), RMS (.074), CN (395.1), BBI (.927), MB1 (.890), incremental RMS (.774), incremental CN (.896), TLI (.891), MB2 (.928). Compared to indices for samples of 800 and 1600 with the same CSSD data (see Table 1), the values for the TLI and MB2 did not differ, values for GFI, AGFI, RMS, BBI, MB1, incremental RMS, and incremental CN were marginally improved, values for X^2 and X^2/df were substantially poorer, and the value for CN was substantially improved.

[2] -- Procedures similar to those used for the CSSD data were used to generate 14 random variables that reflected 4 correlated factors. The first nine variables were defined as with the CSSD data, and the fourth factor was defined by five additional variables that were constructed to reflect it. The data was fit with a simple structure model even though the simulated structure was complex. For just this analysis, 10 sets of 14 variables were generated with sample sizes of 25, 100, 400, and 1600. LISREL was used to derive the 12 goodness-of-fit indices for each of the 40 covariance matrices, and one-way ANOVAs and visual inspections were used to assess the sample size effect. F-ratios for the 12 fit indices were: X^2 (238), X^2/df (238), GFI (404), AGFI (403), RMS (153), CN (268), BBI (304), MB1 (304), incremental RMS (91), incremental CN (322), TLI (20), MB2 (16). A student-Newman-Kuels (Hull & Nie, 1981) test of pair-wise differences for the TLI and MB2 indicated that the mean value based on n=25 was significantly different from all other sample sizes, but that no other pair-wise differences were statistically significant. For each of the other 10 indices, nearly all of the possible pair-wise differences were statistically significant. These statistical analyses, and inspection of plots similar to Figure 1, indicate that TLI and MB2 are substantially less affected by sample size than any of the other 10 fit indices.

Appendix I

The following set of compute statements were used with SPSS (Hull & Nie, 1981) in order to create 31,750 sets of nine variables (x1 - x9) to represent the SSSD and CSSD data sets. Subsequent CFA analyses were based on 70 covariance matrices -- 10 for each of 7 sample sizes -- derived from these nine variables.

SSSD Data Set

```
compute  v = normal(1)
compute  f1 = .2 * v + normal (1)
compute  f2 = .4 * v + normal (1)
compute  f3 = .6 * v + normal (1)
compute  x1 = .6 * f1 + normal (1)
compute  x2 = .7 * f1 + normal (1)
compute  x3 = .8 * f1 + normal (1)
compute  x4 = .6 * f2 + normal (1)
compute  x5 = .7 * f2 + normal (1)
compute  x6 = .8 * f2 + normal (1)
compute  x7 = .6 * f3 + normal (1)
compute  x8 = .7 * f3 + normal (1)
compute  x9 = .8 * f3 + normal (1)
```

CSSD Data Set

```
compute  v = normal(1)
compute  f1 = .2 * v + normal (1)
compute  f2 = .4 * v + normal (1)
compute  f3 = .6 * v + normal (1)
compute  x1 = .2 * f2 + .6 * f1 + normal (1)
compute  x2 = .2 * f3 + .7 * f1 + normal (1)
compute  x3 = .8 * f1 + normal (1)
compute  x4 = .2 * f1 + .6 * f2 + normal (1)
compute  x5 = .2 * f3 + .7 * f2 + normal (1)
compute  x6 = .8 * f2 + normal (1)
compute  x7 = .2 * f1 + .6 * f3 + normal (1)
compute  x8 = .2 * f2 + .7 * f3 + normal (1)
compute  x9 = .8 * f3 + normal (1)
```

Appendix II

A total of 12 goodness-of-fit indicators were considered in the present investigation, and are described below. The first six indices require information from tests of only the hypothesized model, and are called stand-alone indices for purposes of this investigation. The rest are based on difference in goodness-of-fit for the hypothesized model and its corresponding null model, and are referred to as incremental indices. Incremental index were derived according to the form (x-y)/y where x refers to the goodness-of-fit for either the hypothesized model or the null model, whichever is expected to be largest, and y refers to the goodness of fit for the other. This set of indices is called form 1 incremental indices for purposes of the present investigation. A second set of incremental indices was derived according to the form + or - (x-y)/(y-I) where x and y refer to the goodness of fits for the target and null models respectively, I is an ideal or optimum value for x. These are called form 2 incremental indices for purposes of the present investigation.

<u>Stand Alone Indices.</u>

1) X^2 (see Joreskog & Sorbom, 1981)

2) X^2/df_T where $df_T = [\ .5 \times p \times (p+1)\ ] - q$

3) $GFI = 1 - [\ (t[\ (E^{-1} \times S - I)^2/(t[\ E\ S^{-1})^2$

(see Joreskog & Sorbom (1981, p. I.40).

4) $AGFI = 1 - [p \times (p+1)/2df_T] \times (1 - GFI)$

(see Joreskog & Sorbom (1981, p. I.40).

5) $RMS = [\ 2\ E\ E\ (s_{i[} - o_{i[})^2 / p \times (p+1)]^{1/2}$ where

$s_{i[}$ and $o_{i[}$ are elements in S and E

(see Joreskog & Sorbom (1981, p. I.41).

6) $CN = [[z_{cri[} + (2 \times df_T - 1)^{1/2}]^2/[2 \times X^2_T/(N-1]] +1$ where

$z_{cri[}$ = the critical value from a normal curve table for a given probability level -- 1.96 in the present investigation (see Hoelter, 1983, p. 31).

Incremental Form 1 Indices.

7) Incremental Form 1 index for X^2 = BBI = $(X^2_o - X^2_T)/(X^2_o)$ where the subscripts o and T refer to the null and target models (see Bentler & Bonett, 1980).

8)[1] Incremental Form 1 index for X^2/df (MB1) =

$[(X^2_o/df_o) - (X^2_T/df_T)]/[(X^2_o/df_o]$

9)[1] Incremental Form 1 index for RMS = $(RMS_o - RMS_T)/(RMS_o)$

10)[1] Incremental Form 1 index for CN = $(CN_T - CN_o)/(CN_T)$

Incremental Form 2 Indices.

11)[1] Incremental Form 2 index for X^2 (MB2) =

$(X^2_o - X^2_T)/(X^2_o - df_T)$ where df_T is the expected value of X^2 when the target model is true.

12) Incremental Form 2 index for X^2/df = TLI =

$[(X^2_o/df_o) - (X^2_T/df_T)]/[(X^2_o/df_o - 1]$

where 1 is the expected value of X^2/df when the target model is true.

---

[1] We know of no previous descriptions of these goodness of fit indicators.

REFERENCES

Bagozzi, R. P. (1980). Causal models in marketing. New York: Wiley.

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

Fornell, C. (1983). Issues in the application of covariance structure analysis. Journal of Consumer Research, 9, 443-48.

Hoelter, J. W. (1983) The analysis of covariance structures: Goodness-of-fit indices. Sociological Methods & Research, 11, 325-344.

Hull, C. H., & Nie, N. H. (1981). SPSS update 7-9. New York: McGraw-Hill.

Joreskog, K. G. (1981). Analysis of covariance structures. Scandinavian Journal of Statistics, 8, 65-92.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood. Chicago: International Educational Services.

Long, K. S. (1983). Confirmatory Factor Analysis. Beverly Hills, CA: Sage.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students form different academic departments and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. Multivariate Behavioral Research, 20, 427-449.

Marsh, H. W. (1986a). The Self Description Questionnaire (SDQ): A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph. Faculty of Education, University of Sydney, NSW Australia.

Marsh, H. W. & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. American Educational Research Journal, 21, 341-366.

Marsh, H. W. & Hocevar, D. (1985). The application of confirmatory factor analvsis to the study of self-concept: First and higher order factor structures and their invariance across age groups.

Psychological Bulletin, 97, 562-582.

Marsh, H. W., Smith, I. D. & Barnes, J. (1985). Multidimensional self-concepts: Relationships with sex and academic achievement. Journal of Educational Psychology, 77, 581-596.

Pedhauzur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Sobel, M. F. & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In Sociological Methodology 1985 (pp. 152-178). San Francisco: Jossey-Bass.

Table 1

Mean and Standard Error (SE) of 12 Goodness of Fit Indicators For Seven Sample Sizes in Four Data Sets

Indicator		n=25	n=50	n=100	n=200	n=400	n=800	n=1600	Total	F-ratio
1) X^2										
1	M	35.55	34.06	37.65	34.52	48.43	49.26	70.60	44.30	14.2**
	SE	4.46	2.71	3.18	3.19	1.75	3.82	4.60	1.95	
2	M	56.04	78.20	119.64	210.77	414.43	738.89	1439.03	435.53	983.2**
	SE	4.49	4.47	7.09	5.95	13.12	25.76	29.04	56.42	
3	M	24.84	25.33	24.27	23.34	24.79	23.64	23.38	24.23	.1
	SE	2.30	1.84	2.55	1.95	1.86	3.18	2.59	.86	
4	M	30.90	73.03	35.52	45.05	62.00	95.30	189.81	70.23	140.5**
	SE	2.27	3.22	2.37	2.60	5.12	5.19	9.11	6.63	
2) X^2/df										
1	M	1.48	1.42	1.57	1.44	2.02	2.05	2.94	1.85	14.2**
	SE	.19	.11	.13	.13	.07	.16	.19	.08	
2	M	2.34	2.26	4.90	8.78	17.27	30.79	59.96	18.20	983.2**
	SE	.19	.19	.30	.25	.55	1.07	1.21	2.38	
3	M	1.04	1.06	1.01	.9[illegible]	1.03	.99	.98	1.01	.1
	SE	.10	.08	.11	.08	.08	.13	.11	.04	
4	M	1.29	1.37	1.48	1.88	2.58	3.97	7.91	2.93	140.5**
	SE	.09	.13	.10	.11	.21	.22	.38	.28	
3) Goodness of Fit Index (GFI)										
1	M	.638	.735	.847	.911	.927	.961	.970	.856	60.1**
	SE	.030	.025	.014	.006	.005	.003	.002	.015	
2	M	.279	.398	.430	.4[illegible]9	.451	.474	.478	.422	21.6**
	SE	.020	.020	.020	.009	.010	.011	.006	.009	
3	M	.781	.880	.938	.967	.983	.991	.996	.934	231.0**
	SE	.011	.005	.006	.003	.001	.001	.001	.009	
4	M	.750	.849	.913	.940	.956	.964	.964	.906	91.5*
	SE	.016	.013	.006	.004	.004	.002	.002	.009	
4) Adjusted Goodness of Fit Index (AGFI)										
1	M	.322	.503	.714	.833	.865	.927	.944	.729	60.1**
	SE	.056	.047	.027	.011	.010	.006	.003	.028	
2	M	-.352	-.127	-.068	-.034	-.028	.013	.020	-.082	21.6**
	SE	.038	.037	.037	.018	.019	.020	.011	.018	
3	M	.589	.775	.884	.939	.968	.984	.992	.876	234.4*
	SE	.020	.009	.011	.006	.002	.002	.001	.017	
4	M	.513	.718	.837	.887	.919	.934	.934	.823	91.2*
	SE	.030	.024	.012	.007	.007	.004	.004	.018	

Table 1 continued

Indicator		n=25	n=50	n=100	n=200	n=400	n=800	n=1600	Total	F-ratio
5) Root Mean Square Residual (RMS)										
1	M	.437	.317	.235	.176	.164	.121	.108	.223	53.2**
	SE	.030	.022	.017	.010	.007	.005	.004	.014	
2	M	.031	.025	.024	.022	.021	.021	.020	.023	5.1*
	SE	.004	.002	.002	.001	.001	.001	.001	.001	
3	M	.189	.137	.096	.068	.049	.036	.024	.086	141.1*
	SE	.009	.004	.006	.004	.002	.002	.002	.007	
4	M	.246	.168	.122	.099	.086	.078	.079	.125	62.87
	SE	.014	.013	.005	.004	.005	.003	.002	.008	
6) Critical N										
1	M	31.2	61.6	109.6	241.1	324.9	680.6	619.4	337.8	80.9**
	SE	3.7	6.8	8.6	19.8	11.6	74.9	59.9	40.0	
2	M	18.6	26.3	34.2	38.0	38.8	43.5	44.3	34.8	42.2**
	SE	1.4	1.8	1.9	1.1	1.2	1.6	0.9	1.2	
3	M	41.0	79.9	195.3	357.7	667.5	1456.5	3114.1	844.6	30.2*
	SE	3.1	5.8	41.8	34.2	62.7	124.9	515.0	144.1	
4	M	33.0	65.5	113.8	178.0	268.2	335.7	334.8	189.9	78.9**
	SE	2.8	8.5	7.6	10.5	23.7	13.9	15.1	14.9	
7) X^2 Bentler-Bonett Index (BBI)										
1	M	.777	.883	.930	.966	.975	.987	.991	.930	82.1**
	SE	.020	.007	.006	.003	.001	.001	.001	.009	
2	M	.820	.856	.893	.900	.902	.912	.913	.885	20.5*
	SE	.016	.009	.005	.003	.003	.003	.002	.005	
3	M	.633	.745	.830	.907	.948	.974	.987	.861	67.3**
	SE	.028	.022	.022	.008	.004	.003	.001	.016	
4	M	.615	.710	.775	.846	.883	.915	.913	.808	81.6*
	SE	.011	.023	.014	.009	.011	.004	.004	.014	
8) X^2/df Incremental Fit Index 1 (MB1)										
1	M	.666	.824	.894	.949	.962	.981	.986	.895	82.1**
	SE	.030	.011	.009	.005	.002	.002	.001	.014	
2	M	.730	.784	.839	.850	.850	.867	.870	.827	20.5*
	SE	.024	.014	.008	.004	.005	.005	.003	.007	
3	M	.451	.617	.745	.861	.921	.961	.981	.791	67.3**
	SE	.042	.033	.032	.012	.006	.005	.002	.024	
4	M	.423	.564	.662	.769	.824	.873	.869	.712	81.6*
	SE	.016	.035	.021	.013	.016	.007	.006	.020	
9) Incremental RMS										
1	M	.731	.808	.862	.894	.897	.924	.932	.864	84.9**
	SE	.013	.010	.009	.006	.003	.003	.003	.009	
2	M	.809	.797	.834	.842	.848	.853	.852	.833	3.1*
	SE	.020	.023	.010	.006	.004	.003	.003	.005	
3	M	.517	.607	.681	.754	.823	.868	.912	.738	49.8**
	SE	.031	.031	.024	.017	.007	.008	.006	.018	
4	M	.467	.575	.628	.702	.734	.770	.758	.662	35.5**
	SE	.026	.031	.020	.011	.013	.006	.007	.014	

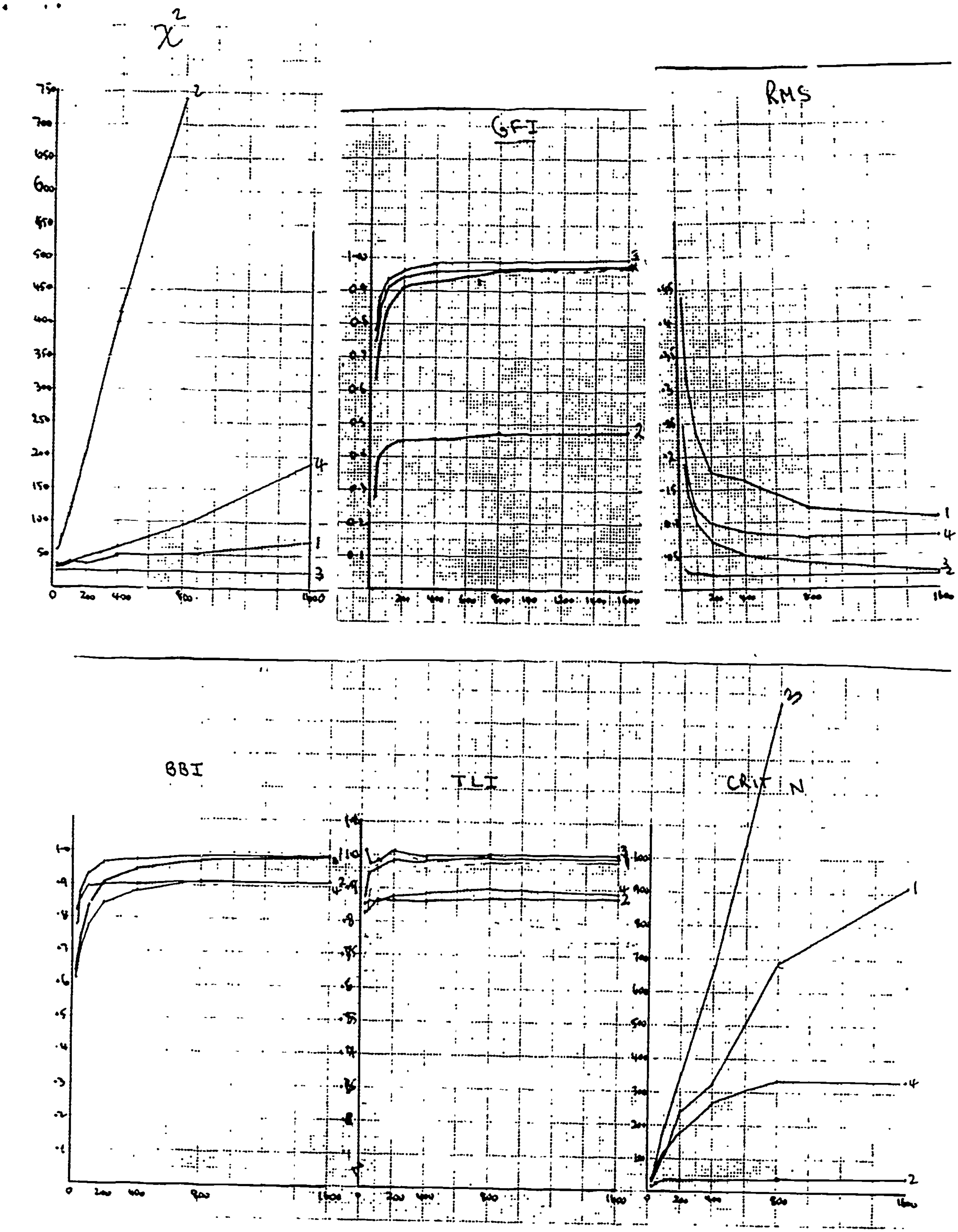
Table 1 continued

Indicator		n=25	n=50	n=100	n=200	n=400	n=800	n=1600	Total	F-ratio
		Sample Sizes								
10) Incremental Critical N										
1	M	.667	.822	.893	.949	.962	.981	.986	.894	84.6**
	SE	.030	.011	.009	.005	.001	.002	.001	.014	
2	M	.709	.769	.825	.838	.841	.857	.860	.814	23.6**
	SE	.024	.014	.008	.004	.005	.005	.003	.008	
3	M	.479	.637	.760	.869	.926	.964	.982	.802	70.6**
	SE	.038	.031	.030	.011	.005	.005	.002	.022	
4	M	.451	.588	.681	.781	.839	.880	.877	.728	89.0**
	SE	.015	.033	.020	.013	.012	.006	.006	.019	
11) Tucker-Lewis Index (TLI)										
1	M	.877	.946	.959	.984	.981	.990	.991	.961	4.2*
	SE	.048	.017	.009	.005	.001	.001	.001	.009	
2	M	.825	.839	.867	.864	.860	.871	.872	.856	2.2
	SE	.026	.014	.008	.004	.005	.005	.003	.005	
3	M	1.013	.975	.988	1.010	.998	1.001	1.001	.998	.1
	SE	.106	.041	.035	.017	.006	.005	.002	.017	
4	M	.852	.860	.862	.878	.884	.902	.884	.874	.1
	SE	.093	.075	.027	.013	.017	.007	.006	.017	
12) X^2 Incremental Fit Index 2 (MB2)										
1	M	.924	.966	.974	.989	.987	.993	.994	.975	4.2*
	SE	.029	.011	.006	.003	.001	.001	.001	.005	
2	M	.888	.895	.912	.910	.907	.914	.914	.906	1.7
	SE	.017	.009	.006	.003	.003	.003	.002	.003	
3	M	1.001	.986	.994	1.010	.999	1.001	1.000	.998	.1
	SE	.048	.023	.021	.011	.004	.003	.001	.008	
4	M	.910	.915	.916	.923	.924	.935	.923	.921	.1
	SE	.042	.040	.016	.008	.011	.004	.004	.009	
X^2 [a] Null										
1	M	156.3	288.6	538.4	1024.2	1942.3	3820.7	7593.5	2194.8	5341.**
	SE	8.3	15.5	20.3	23.5	48.2	38.7	66.7	301.3	
2	M	315.0	546.9	1111.4	2102.4	4213.1	8365.8	16565.1	4745.6	21341**
	SE	11.5	17.4	30.7	31.4	36.8	53.3	70.3	660.4	
3	M	68.1	102.6	147.3	255.8	472.6	915.5	1797.8	537.1	1915**
	SE	4.0	6.9	5.0	14.4	12.9	14.4	27.9	70.3	
4	M	80.3	114.2	158.9	296.7	543.4	1127.6	2179.2	642.9	1665.**
	SE	5.5	7.2	8.2	18.6	12.7	26.4	33.6	86.1	

[a] The X^2 for the null model is presented to illustrate the substantial effect that sample size has on its value.

* $p < .05$; ** $p < .01$.

Figure 1. The sample size effect for selected fit indices in each of the four data sets. Since the form of the sample size effect was similar for X^2 and X^2/df, for BBI and MB2, and for TLI and MB2, only the first of each of these pairs is presented. However, the values for the fit indices not included in Figure 1 are presented in Table 1.


χ²
GFI
RMS
BBI
TLI
CRIT N