DOCUMENT RESUME

ED 387 516 TM 023 750

AUTHOR Chau, Hung; Hocevar, Dennis

TITLE The Effects of Number of Measured Variables on Goodness-of-Fit in Confirmatory Factor Analysis.

PUB DATE Apr 95

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Chi Square; College Students; *Factor Structure; *Goodness of Fit; Higher Education; Measurement Techniques; *Sample Size; Scores; Statistical Analysis; *Statistical Bias

IDENTIFIERS *Confirmatory Factor Analysis; *Student Evaluation of Educational Quality

ABSTRACT

This study addressed which, if any, contemporary fit indices are least susceptible to the bias associated with confirmatory factor analysis (CFA) involving a large number of measured variables. Data were obtained from student responses from 1980 to 1990 on the Student Evaluations of Educational Quality (SEEQ) instrument of H. Marsh (1987). Factor analytic studies have validated the factor structure of the SEEQ. For this study, only student scores for 28 SEEQ items (7,407 classes) were included in a CFA model. Fit indices evaluated were: (1) the GFI (goodness of fit) index of K. G. Joreskog; (2) the Satorra-Bentler chi-square test; (3) Joreskog's adjusted goodness of fit index (AGFI); (4) the Bentler-Bonnett normed fit index (NFI); (5) the comparative fit index (CFI) of P. M. Bentler; and (6) the index of L. R. Tucker and C. Lewis (TLI). Bentler's CFI, the Bentler-Bonnett NFI, and the TLI were highly stable within seven factor models varying from 14 to 28 items. Because the CFI and TLI have the traditional advantage of protecting against the bias associated with large samples, results support their routine use as an adjunct to the chi-square test in CFA. Three tables present analysis and comparison results. (Contains 13 references.) (SLD)

THE EFFECTS OF NUMBER OF MEASURED VARIABLES ON GOODNESS-OF-FIT IN CONFIRMATORY FACTOR ANALYSIS

Hung Chau & Dennis Hocevar
Univ. of Southern California

Numerous investigations have been conducted on dozen of proposed indices of goodness-of-fit in confirmatory factor analysis. The focus of these investigations has been largely limited to sample size (see review and discussions by Bentler, 1990; Bollen, 1990; Gerbing el al., 1992, Marsh et al., 1988; Mulaik et al., 1989). It is well-known that a large sample size "biases" the chi-square test in a confirmatory analysis in favor of model rejection, and researchers have proposed a variety fit indices as solutions to this problem. Indeed, two fit indices, the Tucker-Lewis (1973) Index TLI (also called the non-normed fit index by Bentler and others), and Bentler's (1990) comparative fit index (CFI), appear to have at least partially solved this problem (Marsh et al., 1988; Bentler, 1990).

Although originally pointed out by Fornell (1983), much less attention has been paid to another type of "bias" that is inherent in a confirmatory factor analysis. Fornell (1983) points out that larger models (those with many items or indicators) are more likely to be rejected than smaller models. Stated another way, if we were to analyze a 12-item personality test with three 4-item dimensions using the 12x12 item covariance matrix as input, we would likely reject the model (at least our experience suggests that this is the case). In contrast, if we were to simplify the aforementioned measurement model by adding items from the same dimension and forming a 6x6 covariance of "doublets" (as is often done), we would likely not reject the model, or at least its fit would be considerably better than when analyzed a 12x12 matrix. Thus, the problem is that identical data can be arbitrarily configured in ways that either support or do not support the a priori measurement model.

The problem is even more evident when one considers confirmatory factor analysis in light of traditional reliability theory. As is well known, increasing the number of indicators (e.g., items) is a widely recommended method for assuring high reliability. It is almost certain that a longer measure will be better than a shorter measure. But in the context of confirmatory factor analysis, the opposite is true. Longer measures will typically be poorer than shorter measures, at least in terms of model fit. One can easily verify this fact by analyzing a full 20-item, uni-dimensional questionnaire, and then comparing its fit to that of its two randomly determined halves (as determined in two separate confirmatory factor analyses). In our experience, the fit of the complete questionnaire will be considerably better than the fit of the shorter questionnaire. Perhaps because this tendency is not well known, a cursory review of recent confirmatory factor analytic articles indicates that the typical researcher readily accepts two or three indicator confirmatory models without even examining the reliability of the constructs that the confirmatory factor analysis "supports".

Our anecdotal evidence and at least some empirical evidence (Hocevar et al., 1984) suggests that the traditional chi-square test is strongly biased against models with a large number of measured variables. It is reasonably to expect that some contemporary fit indices might control for this bias. In our estimation, over fifty such indices have been proposed to date. For practical reasons, we will limit the present analysis to those which are available in two well-known structural equation modeling computer programs - - - EQS Version 4.02 (Bentler, 1993) and LISREL 8 (Joreskog & Sorbom, 1993). The issue to be addressed in this study is which (if any) contemporary fit indices are least susceptible to the bias associated with confirmatory factor analysis that involves a large number of measured variables.

Method

Data were obtained from student responses between 1980 to 1990 to Marsh' (1987) Students' Evaluations of Educational Quality (SEEQ) instrument. The instrument has 41 items with clusters of these items designed to measure nine separate dimensions of instructor and course effectiveness. Factor analytic studies (e.g., Marsh & Hocevar, 1991) have validated the SEEQ factor structure underlying nine dimensions of teaching and course effectiveness. Each SEEQ item was rated on a Likert-scale from 1 to 5 with high score indicating rating effectiveness. For this study only student scores for 28 SEEQ items were included in a confirmatory factor analytic model. The CFA model specified *a priori* measurement model with seven factors, each having four item loadings as follows: Learning Value (item 1-4), Instructor Enthusiam (item 5-8), Organization/Clarity (item 9-12), Group Interaction (item 13-16), Individual Rapport (item 17-20), Breadth of Coverage (item 21-24), and Workload/Difficult (item 32-35).

The data were screened with listwise deletion by PRELIS 2 (Joreskog & Sorbom, 1993) which resulted in a final sample of 7,407 classes. Item responses in each class were averaged across students to create a data matrix with 28 x 7,407 continuous elements. Sample covariances were derived from this matrix for model estimation using LISREL 8 and EQS Version 4.02.

An initial CFA model with 28 items loading on their designated seven separate but intercorrelated factors was first estimated. In subsequent runs, the same CFA model was maintained but the number of items per factor was then reduced by random deletion to 3 and then to 2. Thus, three highly similar CFA models with 28 (4x7), 21 (3x7), and 14 (2x7) items were analyzed. All model parameters and goodness-of-fit indices were estimated by both LISREL 8 and EQS. Because the items exhibited high skewnesses ranging from -3.1074 to .4512 and high kurtoses ranging from 1.5576 to 15.6020, two methods of estimation were used: (a) maximum likelihood (ML) method by both LISREL 8 and EQS, and (b) robust ML by EQS and LISREL weighted least square (WLS) distribution-free method for non-normal data. A comparison of the two methods of estimation provided a test for the highly non-normal data of the influence of a violation of the normal theory assumption by subjecting the data under the normal ML and the robust ML provided by EQS's Satorra-Bentler scaled chi-square and the LISREL asymptotic distribution-free estimation.

Results

With the ML method, both LISREL 8 and EQS produced consistent results for model fits and parameter estimates. The chi-square values were inexplicably somewhat lower with LISREL 8 than with EQS but the differences were of no significant meaning. The estimates of standard errors by the normal ML were negatively downward biased by a range of -.002 to -.007 in comparison to the same estimates obtained with the robust ML method by EQS. This result confirmed existing research findings (e.g., Muthen & Kaplan, 1985) of the downward bias of the normal ML when used with severely non-normal data. Model fits were improved markedly with the Satorra-Bentler scaled statistic under EQS robust ML estimation. Thus, the findings discussed below are based on the Satorra-Bentler scaled test statistic when possible.

1. Joreskog's GFI index. Poorer fit for larger models was noted on Joreskog's goodness-of-fit index. GFI index values were .872, .813, and .740 for the 14, 21, and 28 item models, respectively (Table 1).

Insert Table 1 about here

2. Satorra-Bentler chi-square test. As predicted, the chi-square goodness-of-fit test was strongly "biased" against models that included a large number of measured variables. Specifically, chi-squares equaled 1,960, 3,918, and 9,042 for the 14, 21, and 28 item models, respectively (Table 2).

Insert Table 2 about here

3. Joreskog's AGFI index. Joreskog's adjusted goodness-of-fit index adjusts for degree of freedom. Thus, we expected that this index might not be susceptible to large model bias. This expectation was disconfirmed. The AGFI index had values of .760, .742, and .679 for the 14, 21, and 28 item models, respectively (Table 1).

4. Bentler-Bonett normed fit index (NFI). The NFI had a strong negative monotonic relationship with the number of items. For models with 28, 21, and 14 items, the NFI ranged from .975, .985, and .987 (Table 2) and from .969, .978, and .988 (Table 3). The strong stability of the NFI in models containing different numbers of measured variables supports the conclusion that the NFI is not biased against larger models.

Insert Table 3 about here

5. Bentler's comparative fit index (CFI). The CFI was proposed as a way of controlling for the well-known sample size bias inherent in the chi-square test. In our analysis, the CFI index had a negative monotonic relationship with the number of items, but its strong stability (Table 2) in models containing different numbers of measured variables supports the conclusion that the CFI is not biased against larger models.

6. Tucker-Lewis index (TLI) (also known as the non-normed fit index). The TLI, originally proposed in 1973 by Tucker and Lewis, has been more recently advocated by Marsh et al. (1988) as a way of controlling for sample size effects. In this study, the NNFI was the only index that did not have a monotonic relationship with the number of items, and similar to the CFI, it was very stable. Specifically, the TLI had values of .972, .981, and .978 (Table 2) and .964, .972, and .964 (Table 3) for analyses with 28, 21, and 14 items respectively.

Conclusion

As predicted at the onset of this study, models with a larger number of items had poorer fits when fit was assessed using the chi-square statistic. Neither Joreskog's GFI or his AGFI adequately controlled for the number of items. However, Bentler's CFI, Bentler-Bonett NFI, and the Tucker-Lewis TLI were highly stable within seven factor models varying from 14 to 28 items. Because the CFI and TLI have the traditional advantage of protecting against the bias associated with large samples, our results support their routine use as an adjunct to the chi-square test in confirmatory factor analysis.

References

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. Psychological Bulletin, 107, 238-246.

Bentler, P. M. EQS, A Structural Equation Program, BMDP Statistical Software, Inc, Version 4.02 (c), 1993

Bollen, K. A. (1990). Overall Fit in Covariance Structure Models: Two Types of Sample Size Effects. *Psychological Bulletin, 107*, 256-259.

Fornell, C. (1983). Issues in the Application of Covariance Structure Analysis: A Comment. *Journal of Consumer Research, 9*, 443-450.

Gerbing, D. W. & Anderson, J. C. (1992). Monte Carlo Evaluations of Goodness of Fit Indices for Structural Equation Models. *Sociological Methods & Research, 21*, 132-159.

Hocevar, D., Zimmer, J., & Strom, B. (1984). The Confirmatory Factor Analytic Approach to Scale Development and Evaluation. *Paper presented at the 1984 conference of the National Council on Measurement in Education.*

Joreskog, K. G., & Sorbom, D. *LISREL 8 User's Reference Guide*, Scientific Software International, Inc., 1993

Mulaik, S. A., James, L. R., Alstine, J. V., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of Goodness-of-Fit Indices for Structural Equation Models. *Psychological Bulletin*, *105*, 430-445.

Muthen, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171-189.

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size. Psychological Bulletin, 103, 391-410.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: the generality of factor structures across academic discipline, instructor level, and course level. Teaching & Teacher Education, 7, 9-18.

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

Table 1
Number of Items and Fit Indices by LISREL[a] and EQS[a] (n = 7,407)

Items	Items/ Factor		Null Model		Estimated Model		Fit Indices					
			df	Chi-square	df	Chi-square	NFI	NNFI	CFI	GFI	AGFI	ρ[b]
28	4/factor	LISREL	378	366,595.46	329	35,288.51	.904	.890	.905	.740	.679	.973
		EQS	378	366,595.46	329	35,294.07	.904	.890	.905	-	-	.973
21	3/factor	LISREL	210	262,495.39	168	19,289.71	.927	.909	.927	.813	.742	.964
		EQS	210	262,495.39	168	19,318.92	.926	.909	.927	-	-	.964
14	2/factor	LISREL	91	146,183.58	56	8,544.77	.942	.906	.942	.872	.760	.955
		EQS	91	146,183.58	56	8,548.02	.942	.906	.942	-	-	.955

Note: [a]maximum likelihood estimation for a seven-factor model. ρ[b]=reliability coefficient.

Table 2
Number of Items and Fit Indices by EQS (n = 7,407)
Estimation methods for normal data (ML) vs. non-normal data (robust ML)

Items	Items/ Factor	Null Model		ML		Robust ML		ML Fit Indices			Robust ML Fit Indices		
		df	χ^2	df	χ^2	df	χ^2_{S-B}	NFI	NNFI	CFI	NFI	NNFI	CFI
28	4/factor	378	366,595.46	329	35,294.07	329	9,041.90	.904	.890	.905	.975	.972	.976
21	3/factor	210	262,495.39	168	19,318.92	168	3,918.34	.926	.909	.927	.985	.981	.986
14	2/factor	91	146,183.58	56	8,548.02	56	1,960.16	.942	.906	.942	.987	.978	.987

Note: ML=maximum likelihood method for normal data. Robust ML for non-normal data with χ^2_{S-B} (Satorra-Bentler scaled statistics adjusting for non-normality of data).

Table 3
Number of Items and Fit Indices by LISREL 8 (n = 7,407)
Estimation methods for normal data (ML) vs. non-normal data (weighted least square (WLS))

Items	Items/ Factor	Null Model		ML		WLS		ML Fit Indices			WLS Fit Indices		
		df	χ^2	df	χ^2	df	χ^2	NFI	NNFI	CFI	NFI	NNFI	CFI
28	4/factor	378	366,595.46	329	35,288.51	329	11,538.49[a]	.904	.890	.905	.969	.964	.
21	3/factor	210	262,495.39	168	19,289.71	168	5,792.72[b]	.927	.909	.927	.978	.972	.
14	2/factor	91	146,183.58	56	8,544.77	56	1,772.19[b]	.942	.906	.942	.988	.980	.

Note: ML=maximum likelihood method for normal data. WLS=weighted least square distribution-free method for non-normal data. [a]=estimated with the correlation matrix due to the asymptotic covariance matrix being positive indefinite. [b]=estimated with asymptotic covariance matrices.