DOCUMENT RESUME

ED 377 217 TM 022 385

AUTHOR	Hutchinsor., Susan R.
TITLE	The Stability of Post Hoc Model Modifications in Covariance Structure Models.
PUB DATE	Apr 94
NOTE	36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Computer Simulation; Models; Monte Carlo Methods; *Reliability; *Sample Size; Statistical Analysis; Statistical Studies
IDENTIFIERS	Confirmatory Factor Analysis; *Covariance Structure Models; *Post Hoc Methods; Specification Error; Stopping Rules

ABSTRACT

The work of R. MacCallum et al. (1992) was extended by examining chance modifications through a Monte Carlo simulation. The stability of post hoc model modifications was examined under varying sample size, model complexity, and severity of misspecification using 2- and 4-factor oblique confirmatory factor analysis (CFA) models with four and eight secondary loadings respectively and four levels of misspecification. Sample sizes were 200, 400, 800, and 1,200. Results in terms of recovery of population models partially support findings by MacCallum that more severe levels of misspecification result in less successful searches. The study also supports the sample size effect found by MacCallum, in which larger samples tend to result in greater recovery of the population models. Some evidence was found of differences due to model complexity, with model modifications somewhat more stable in the 4-factor model. An alternative to the number of modifications or overall nonsignificance as stopping criteria might be inspection of the modification indexes for a distinctive drop in successive values of the maximum modification index. Three tables and two figures present analysis results. (Contains 30 references.) (SLD)

The Stability of Post Hoc Model Modifications in Covariance Structure Models

Susan R. Hutchinson

University of Denver

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, 1994

In applications of covariance structure modeling (CSM), researchers often conclude, whether on the basis of overall and/or component measures of fit, that their models fail to exhibit acceptable levels of fit. Model misfit may be due to a number of factors including assumption violations, model underidentification, poor measurement, inadequate substantive theory, or a combination of the above. If one can rule out the first three reasons for misfit through examination of component fit measures, as suggested by Jöreskog and Sörbom (1988, pp. 40-42), the lack of fit is likely due to inappropriate model specification.

In the presence of model misspecification, the researcher essentially has three options: report the model as is, test several plausible, alternative models, or attempt to locate the source of misfit in the original model. While some might argue for the first option (Steiger, 1990), examination of the applied literature provides little evidence that this is the choice of most applied researchers. Others promote the practice of testing multiple a priori models (Browne & Cudeck, 1989; Cudeck & Browne, 1983; MacCallum, Roznowski, & Necowitz, 1992), but this approach is seldom realized. The most popular practice is to conduct a "specification search" (Leamer, 1978) for the purpose of conducting post hoc model modifications, with the resulting modified models often used as the basis for subsequent reformulation of theory.

There are several strategies available for conducting specification searches. Probably the most commonly employed of which is use of the "modification index" (MI) produced by LISREL. For each fixed parameter, the MI indicates the approximate decrease in the overall likelihood-ratio χ^2 test that would occur if the corresponding parameter were freed (Sörbom, 1989). The MI is used to test the hypothesis that a single constrained parameter is zero in the population. Typically, the MI is used in what MacCallum et al. (1992) have called "sequential model modification." Using this approach, researchers free the parameter having the largest MI, reestimate the model, and continue this process until the model has reached an acceptable level of fit. Other strategies for conducting specification searches include Chou and Bentler's (1990) multivariate Lagrange Multiplier and Wald tests, which are generalizations of the MI and LISREL's "t" values, respectively, and the combined MI/expected parameter change statistic (Kaplan, 1989; Saris, Satorra, & Sörbom, 1987), which provides a test of the "practical" significance of a modification.

The application of model modification procedures is quite popular in practice. A survey of covariance structure modeling applications in the psychological literature conducted by Breckler (1990), found that of 72 CSM applications reviewed, 28 reported post hoc model modifications to improve overall model fit. MacCallum et al. (1992) examined an additional 28 papers, of which 9 acknowledged having modified initial models to improve fit.

Despite its popularity, the use of post hoc model modification is a data-driven procedure that is characterized by capitalization on chance. Cliff (1983) noted that when *ex post facto* analyses are conducted under the guise of confirmatory hypothesis testing, the relevant probability distributions are no longer valid. Although Cliff and others (Leamer, 1978; Steiger, 1990) have cautioned against the use of post hoc model modifications, little is actually known about the extent to which post hoc procedures in CSM do capitalize on chance.

In statistical analyses such as analysis of variance (ANOVA) or regression, methods have been developed to control Type I errors among post hoc tests. The most well-known of these methods is the Scheffé S test (Sheffé, 1953), which maintains the experimentwise error rate at any preselected value for all possible comparisons. There are no similar procedures in CSM, despite the sometimes immoderate application of post hoc modeling procedures. Furthermore, unlike tests in ANOVA or regression for which error rates related to post hoc analyses are known, error rates associated with tests of misspecified parameters in CSM are unknown.

A recent study by MacCallum et al. (1992) appears to be the only study to date that has examined the sampling stability of post hoc model modifications. They found that modifications were highly inconsistent across repeated samples even with sample size as high as 1,200, where only 6 of 10 searches resulted in the same final model. Stability of modifications decreased with sample size until there were no searches resulting in the same modified models when sample size was 250 or less. Overall, the authors concluded that they would have little confidence in the validity of modified models, except with extremely large sample size. The study by MacCallum et al. was limited, however, in that the authors used an empirical data set for which the underlying population model was unknown, they did not manipulate levels of misspecification, and they used only 10 replications per sample size. Furthermore, use of nonsimulated data precluded their control of pertinent model characteristics, such as size of factor loadings.

No other study of specification searches has considered the issue of stability of searches per se, although the search histories provided by MacCallum (1986) could be used to roughly assess stability. Examination of these search histories reveals, not surprisingly, that final models obtained through modifications were less stable with increasing initial misspecification. Other studies of specification searches have used single samples (i.e., Silvia & MacCallum, 1988), empirical data (i.e., Kaplan, 1989), or population data from which no samples were drawn (i.e., Hutchinson, 1993; Kaplan, 1988; Tippets, 1991), which has precluded any exploration of the sampling variability of model modifications. The Monte Carlo study by Chou and Bentler (1990) did use 100 samples per cell, but did not provide sufficient detail about modifications to permit assessment of modification consistency across the samples.

The present study extends the work of MacCallum et al. (1992) by examining the problem of chance model modifications within the framework of a Monte Carlo simulation. The stability of post hoc model modifications was examined under varying levels of sample size, model complexity, and severity of misspecification.

Method

Design and Procedures

Hypothetical population models created for this study were two- and four-factor oblique confirmatory factor analysis (CFA) models, with four and eight secondary loadings, respectively (see Figures 1 and 2). Both models have four primary indicators per latent variable. Although neither model is extremely complex, Model B is the more complex of the two because its additional latent variables require estimation of a greater number of parameters, i.e., for the correctly specified models, Model A requires estimation of 21 parameters while Model B requires estimation of 46 parameters. Degrees of freedom for Models A and B are 15 and 90, respectively.

Insert Figures 1 and 2 Here

The decision to limit this study to CFA type models rather than structural models was based on the difficulty of isolating the effects of misspecification of measurement parameters from the misspecification of structural parameters as noted by Anderson and Gerbing (1988) and Kaplan (1988). Therefore, because the issue of stability in post hoc model modification is not well understood, it was decided to avoid the potential problem of confounded measurement and structural misspecification.

Four levels of misspecification were imposed on Models A and B by incorrectly constraining complex factor loadings to zero. This type of error was chosen because it is thought to mimic what is seen in practice. In applications of CFA, researchers often attempt to force simple structure on their models because of the obvious interpretational simplicity. When these models fail to fit the data adequately, researchers engage in specification searches to locate and modify errors which will improve the fit of their models. In CFA applications, these modifications frequently involve the estimation of complex loadings or the relaxation of constraints on correlated residuals. Although the latter practice is seldom substantively justifiable (Fornell, 1983; Gerbing & Anderson, 1984), modification of correlated residuals was permitted in this study to demonstrate the potential problem associated with this practice.

The four levels of misspecification differed in terms of number and size of omitted loadings. Omitted loadings were comprised of either all secondary loadings or half secondary and half primary loadings. Misspecifications for Model A involved incorrectly fixing to zero 2 secondary, 1 primary and 1 secondary, 4 secondary, or 2 primary and 2 secondary loadings. Misspecifications for Model B included incorrectly omitting 4 secondary, 2 primary and 2 secondary, 8 secondary, or 4 primary and 4 secondary loadings. While all omitted secondary loadings had population values of .4, omitted primary loadings had population values of either .6 or .7.

It was decided that although the number of levels of misspecification would be the same for both models, the disparity in model size necessitated different numbers of errors per level for the two models. For example, four errors of omission in Model A would be considerably more serious than the four errors in Model B, which has more parameters. Consequently, levels of misspecification were nested within Models A and B, although proportions of misspecifications were roughly equivalent. The proportions of errors per parameters estimated were approximately .11 (2 errors in Levels 1 and 2) and .24 (4 errors in Levels 3 and 4) for Model A, and .10 (4 errors in Levels 1 and 2) and .21 (8 errors in Levels 3 and 4) for Model B.

Sample sizes for this study were 200, 400, 800, and 1,200. Two hundred was chosen as the lowest size based on studies by Boomsma (1982; 1987) which demonstrated that parameter estimates based on maximum likelihood estimation were robust against sample sizes as small as 200. The inclusion of 1,200 as the largest sample size stemmed from the conclusion by MacCallum et al.

(1992) that post hoc model modifications were not entirely stable even with sample size as high as 800. Consequently, even though a sample size of 1,200 might be considered unrealistically large for many applied researchers, it was included to provide sufficient range for revealing potential effects of methodological interest.

In order to keep the design at a manageable level, certain factors were held constant in this study. For example, loading sizes, although mixed, were kept at the same levels throughout the study. Loadings of .6, .7, and .8 were chosen to be sufficiently large to ensure proper parameter estimation but not so large as to risk the appearance of Heywood cases. Boomsma (1982) warned that the risk of Heywood cases increases when loadings are "too" large because their associated item uniquenesses are sampled around zero. The size of all secondary loadings was .4, which was at least .2 less than any of the primary loadings.

Another variable held constant in this study was the type of misspecification. Only incorrectly omitted factor loadings were modeled. Studies by Farley and Reddy (1987) and Kaplan (1988) have shown that errors of omission are considered much more serious than errors of inclusion because they not only lead to decrements in fit but they also result in biased parameter estimates.

Data Generation

Population covariance matrices were produced within LISREL VII (Jöreskog & Sörbom, 1988) using procedures described by Jöreskog and Sörbom (p. 212-213). Parameters were set equal to the values in Figures 1 and 2, with an identity matrix serving as a "dummy" input matrix for this procedure. The fitted covariance matrix generated as standard output was then used as the input matrix for the LISCOMP (Muthén, 1988) data generation procedure. Based on the population input matrix, LISCOMP generated and output into a stacked ASCII file sample covariance matrices, which the LISREL VII program read for the analyses. One hundred samples were generated per model for each of the four sample sizes for a total of 800 samples. Maximum likelihood estimation via the LISREL VII program was used for estimating parameters.

Misspecified models were created by fixing to zero parameters known to be present in the population model. Specification searches were hierarchical in nature, with parameters having the largest MI being freed at successive steps. Consistent with the technique employed by MacCallum et al. (1992), a maximum of four modifications was made for each of the 3,200 searches (i.e., 2 models x 4 misspecification levels x 4 sample sizes x 100 replications each). MacCallum et al. concluded that specification searches which extended beyond four modifications "served little purpose," and that later modifications frequently reflected chance sample characteristics. Hutchinson (1993) also found that well-fitting models were obtained with no more than four modifications even under severe misspecification. In the event that overall model nonsignificance was obtained (i.e., χ^2 with $p > .05$) or only trivial (i.e., $p > .01$) MIs remained prior to the four modifications, searches were terminated at that point. This process was automated using the "automatic modification" option available in the LISREL program. The automatic modification option essentially conducts sequential model modification internally, by successively freeing parameters with largest MIs, reestimating the model, and repeating this cycle until no more significant (i.e., $p < .01$) MIs remain. It should be noted that the rather mechanical search procedure used in this study is not optimal, but reflects what is frequently seen in CSM applications. It was chosen to demonstrate the potential severity of capitalization on chance when these types of data driven search procedures are used.

Data Analysis

To assess stability of post hoc model modifications search histories were recorded and the results reported descriptively in tables. Recorded information included for each of the 3,200 searches, values of largest MIs at each step, p-values for overall model fit, along with the parameters freed. To obtain the desired information from the LISREL analyses, a program was written in REXX which "stripped" the target information from the LISREL output listings and saved it into separate ASCII files for each cell in the study.

Based on this information, dependent variables included the number of different parameters freed in the final models per cell across all replications, the frequency of the most commonly occurring final model per cell, and the number of times per cell that the model generating the data was recovered. This latter measure is for continuity with previous studies of specification searches, which have focused on recovery of "true" underlying models. A somewhat arbitrary criterion for model modification stability was established as the occurrence of 90 or more of the same respecified models for a given cell.

Results

Recovery of Population Models

Recovery of population models, while not a measure of modification stability per se, was included as a criterion in this study to provide comparability with previous research on model modifications. The number of times per cell that specification searches identified population models is presented in Table 1. Examination of the table reveals that level of misspecification did have an effect on recovery of population models as expected, with greater recovery at Levels 1 and 2. With the exception of Level 3, which was imposed by incorrectly omitting 4 (two-factor model) or 8 (four-factor model) secondary loadings, rates of population model recovery were quite similar for the two- and four-factor models across sample size and levels of misspecification.

Insert Table 1 Here

For the two-factor model at Level 3, modifications resulted in the population model only 4 times out of 400 replications; however, for the four-factor model at Level 3, population model recovery was considerably higher and increased rather dramatically with sample size. This suggests that there might have been something anomalous about the behavior of specification searches in the two-factor model when the Level 3 misspecification was imposed.

Larger sample sizes resulted in greater recovery of population models across most conditions. Using a criterion of 90 out of 100 replications as a descriptive index of consistency, population models were consistently recovered for both the two- and four-factor models when sample size was at least 800 at Levels 1 and 2 and for the four-factor model when sample size was 1,200. Sample size had no effect on recovery of the population model for the two-factor/Level 3 condition.

Because the recovery of population models was so poor under some conditions, the potential success of stopping criteria other than overall nonsignificance or the four modification maximum was examined. At the smaller sample sizes, for example, it appeared that low statistical power may have hindered population model recovery in some cases when nonsignificant overall χ^2 values were reached prior to freeing all omitted population parameters. In some samples, searches resulted in no modifications at all because initially misspecified models were associated with nonsignificant χ^2 tests. In addition, there were instances where specification searches detected subsets of the population parameters, but the stopping criteria precluded detection and subsequent freeing of the remaining parameters. In many of these cases, had searches continued beyond nonsignificance, additional population models would have been found. For example, in the two-factor model at Levels 1 and 2 when N=200, in each case an additional 26 population models would have been identified during extended searches.

Overall, liberalization of stopping criteria would have improved population model recovery more for the four-factor model than for the two-factor model in all sample size and misspecification conditions. In every cell, the corresponding number of population models found after extended searches was higher for the four-factor model, even when the recovery based on the original stopping criteria had been better for the two-factor model. This disparity was greatest at Level 3 where extended searches would have raised rates of recovery for the four-factor model to 98 or higher for all sample sizes greater than 200. In contrast, extended searches in the two-factor model would only have retrieved an additional two population models across all four sample sizes. It appears that alternative stopping criteria would not have ameliorated the poor performance of specification searches for the two-factor model at Level 3.

To determine if differences in power might offer a plausible explanation for some of the findings, power analyses were conducted on selected samples for several cells in the design, using the procedures described by Saris et al. (1987). Within conditions, power was not especially enlightening in terms of explaining differences in population model recovery. When N=200, values of power were so erratic that they did not appear to exhibit any particular pattern. For example, in the two-factor model at Level 1, in one of the samples in which the population model was recovered, power for both omitted population parameters was > .80, which is what one might expect. Similarly, in one of the samples in which only one of the two population parameters was freed, power was .83 for the correctly identified parameter and only .49 for the parameter not found in the specification search. This is also a pattern one might expect. However, there were also samples which successfully recovered the population model despite having low power (i.e., < .60) to detect either parameter. Apparently, power to detect particular parameters alone does not determine rates of population model recovery. Moreover, examination of power values suggests that levels of power not only depend upon model characteristics (Matsueda & Bielby, 1986) or upon location of misspecifications (Saris et al., 1987), but they also appear to depend upon the particular sample used to test the model.

Between conditions there did appear to be some systematic differences in power that were related to differences in population model recovery, with sample size having the most dramatic effect as one might expect. Whereas power values were noticeably inconsistent when N=200, power to detect omitted population parameters was close to or equal to 1.0 when N=1,200 at all levels for the four-factor model, and at Levels 1, 2, and 4 for the two-factor model. Although power also increased with sample size for the two-factor model at Level 3, the increased power did not result in greater population model recovery. In these cells, it appeared that in most samples power was highest for correlated residuals in general, with one particular correlated residual (i.e., $\theta_{\delta 87}$) consistently displaying power close to 1.0. In contrast, power to detect the two-factor population parameters was relatively low in most samples.

In comparing the two- vs. four-factor models at the same level, the four-factor model had greater power to detect omitted parameters of the same magnitude. For example, λ_{93} and λ_{52} both had population loadings of 0.6. However, the power to detect λ_{93} was generally around .95 or higher in the four-factor model while the power to detect λ_{52} in the two-factor model was less consistent with power ranging between about .60 and .96. It appears, therefore, that there might be some sort of sample size by model interaction in terms of power, with power increasing with sample size but at a higher rate for the larger model.

Sampling Stability of Model Modifications

From Table 2 it can seen that for both the two- and four-factor models, modifications were quite stable when sample size was at least 800 and misspecification was at Level 1 or 2. In addition, for the two-factor model, at Level 4 misspecification with sample size of 1,200, the same model was recovered 89 times, which was just under the a priori cutoff of 90. For both models, stability was greatest for the Level 2 misspecification, which was characterized by the omission of either one (two-factor) or two (four-factor) primary loadings in addition to the omission of an equal number of secondary loadings.

Insert Table 2 Here

Conversely, respecified models were least consistent at Level 3 misspecification across virtually all sample size and model conditions. Even at a sample size of 1,200, the most times a single model was recovered at Level 3 for the two-factor model was 25. Stability was greater for both models at Level 4, which omitted either 2 (two-factor) or 4 (four-factor) primary loadings in aduition to omitting an equal number of secondary loadings, than at Level 3 which omitted only secondary loadings.

Another measure of stability, presented in Table 3, is the number of different parameters freed per cell. On this basis, there was a fairly clear superiority of the four-factor model over the two-factor model in terms of modification stability. In virtually every condition, MIs detected fewer different parameters in the four-factor model, with the exception of Levels 1 and 3 at sample size 200, and Level 4 at sample sizes of 800 and 1,200. The discrepancy is even more pronounced when one considers the far greater number of potential parameters to be freed in the four-factor model. In 9 of 16 cells, MIs detected only population parameters in the four-factor model. Three of these cells were at Level 3 misspecification which appeared to be quite unstable on the basis of number of different final models. The reason for this apparent contradiction in findings is the imposition of a maximum of four modifications per respecified model. While specification searches recovered over 45 different models for each sample size of 400 or greater for the Level 3 misspecification, all of these models were comprised exclusively of various subsets of the eight omitted population parameters. Had these searches been continued to eight modifications, MIs would have recovered the underlying population model in 98, 100, and 100 percent of the replications, for sample sizes of 400, 800, and 1,200, respectively. In contrast, the corresponding cells of the two-factor model exhibited considerable inconsistency in the parameters freed. An interesting finding with respect to the type of parameter freed was the far greater number of correlated residuals identified by MIs in the two-factor than the four-factor model for every cell in the study.

Insert Table 3 Here

Sample size also had a considerable effect on modification stability across most conditions in the expected direction, i.e., large samples generally exhibited more stable model modifications, with the notable exception of the Level 3 misspecification. At Levels 1 and 2, using the aforementioned cutoff of 90 or more same final models, modifications were quite stable when sample size was at least 800. Table 3 shows similar patterns in terms of different parameters freed, with numbers decreasing as sample size increased. In the two-factor model sample size seemed to have the greatest impact at Level 4 where MIs freed 26 different parameters across the 100 samples when N=200 compared with only 5 when N=1,200.

Discussion and Conclusions

Results of this study in terms of recovery of population models is somewhat supportive of findings by MacCallum (1986), Silvia and MacCallum (1988), Tippets (1991), and Hutchinson (1993) that more severe levels of misspecification result in less successful searches. Clearly overall recovery of population models was higher for Levels 1 and 2 when compared with Levels 3 and 4, where Levels 3 and 4 had twice as many omitted parameters as Levels 1 and 2. However, recovery was slightly higher for Level 2 than Level 1, and for Level 4 than Level 3, even though Levels 2 and 4 were characterized by the omission of primary as well as secondary loadings. This was surprising given that the omission of primary loadings was thought to be a more serious type of error. However, it suggests that MIs may be more sensitive to important specification errors than previously thought (Kaplan, 1989; Matsueda & Bielby, 1986; Saris et al., 1987).

A possible explanation is that despite the more serious nature of erroneously omitting the larger and more substantively important primary loadings, the greater magnitude of these parameters may have made it easier for the MIs to detect them. Thus, it is likely that there was greater power for locating omitted primary loadings. This is encouraging given that Matsueda and Bielby (1986) and Saris et al. (1987) had found the MI to be unreliable in detecting large errors, sometimes exhibiting high power for trivial parameters and relatively low power for larger omitted parameters.

This study also supports the sample size effect found by MacCallum (1986), whereby larger samples tended to result in greater recovery of population models. It appears that for some conditions, at least part of the sample size effect might have reflected a lack of statistical power especially when misspecification levels were relatively low. When searches were allowed to continue beyond overall model nonsignificance, recovery of population models improved dramatically at sample sizes 200 and 400 at Levels 1 and 2 for both the two- and four-factor models and for the four-factor model at Levels 3 and 4. However, even with this improvement, rates of population recovery never reached satisfactory levels for the two-factor model. When sample size was 800 and 1,200, extended searches added no additional population models in the two-factor case, while they did improve recovery rates for the four-factor model. The findings imply that the four-factor model might have been more sensitive to changes in power and that sample size might have been a more influential factor in determining rates of population model recovery for the four-factor model.

Regarding the stopping criteria employed in this study, use of a fixed number of modifications is clearly not recommended as a general practice. Results of this study demonstrated that such a criterion might result in adequate recovery of population models under some conditions but not under others. A decision to limit modifications to a particular number would depend on a variety of factors including size of the model, level of perceived misspecification, number of subjects, etc. In addition, the decision to extend searches beyond nonsignificance in this study should not be construed as an endorsement of this practice. It was merely done to determine what effect the criterion of nonsignificance might have had on the successful recovery of population models given the sample size dependency of the overall χ^2 test.

The strong sample size effect on stability of post hoc model modifications seen in the study by MacCallum et al. (1992) was also found in the present study. Specification searches conducted with larger samples were clearly more consistent in terms of producing the same modified models across repeated samples. The present study differed from MacCallum et al.'s study in determining the number of subjects required to minimize capitalization on chance to an acceptable level. In the MacCallum et al. study they found that even with a sample size of

1,200, only 6 out of 10 searches resulted in the same final model. And stability deteriorated rapidly with smaller samples until there were no two same final models recovered when N was 250. In contrast, the present study found that over 90% of respecified models were the same at Levels 1 and 2 when sample size was at least 800. Even at Level 4, modifications were more consistent at all sample sizes than they had been in MacCallum et al.'s study. The disparity in findings could be attributed to a number of factors, including differences in model size/characteristics, differences in initial model misfit, or the nature of the misspecifications. However, because MacCallum et al. used an empirical data set for which the underlying population model was unknown, and because they reported so little about the characteristics of their models, any definitive statements concerning the disparate results would be unwarranted.

The strong sample size effect seen both in this study and in the study by MacCallum et al. (1992) is undoubtedly related, at least in part, to the influence of sample size on the sampling variability of either the MIs, model parameter estimates, power, or a combination of the three. As mentioned previously, a limited power analysis conducted to explain differences in recovery of population models revealed that power estimates were highly inconsistent when N was small. It should be noted that determination of power using Saris et al.'s (1987) procedure involves use of the MI as an estimate of the noncentrality parameter needed when using the power tables. Consequently power to detect a particular parameter is directly related to the magnitude of the corresponding MI. This suggests that the apparent effect of sample size on the sampling variability of power estimates was actually mirroring the sampling variability of the MIs.

The other related factor which possibly contributed to differences in stability of specification searches across sample size is the decrease in sampling variability of parameter estimates found to be associated with increasing N (Boomsma, 1982; Gerbing & Anderson, 1985). MIs might have identified fewer different parameters at larger sample sizes because the estimated models were actually more consistent. However, it is most probable that a combination of these factors operated to produce the effect of sample size on stability of modifications.

In addition to the sample size effect, the results showed some evidence of differences due to model complexity, with model modifications being somewhat more stable in the four-factor than in the two-factor model. Again, the slightly greater overall consistency of modifications might have been a function of the higher power found in the four-factor model when compared with the two-factor model holding sample size and level of misspecification constant. Differences between the two models were most pronounced for the Level 3 misspecification.

A recent paper by Kaplan and Wenger (1993) offers some possible insight into the deviant specification searches seen in the two-factor model at Level 3. They demonstrated that the pattern of covariances among parameter estimates may be responsible for Saris et al.'s (1987) observation that power can be unequal for different misspecifications within a single model even when the size of the misspecified parameters is the same. It is possible, therefore, that the high power in the two-factor model to detect correlated residuals might have indicated an unusual pattern of covariances among the parameter estimates, resulting in capricious modifications across different samples.

As mentioned earlier, results of the population model recovery did not offer any support for the use of a fixed number of modifications. Similarly, limiting modifications to a maximum of four did not result in lower risk of capitalization on chance as MacCallum et al. (1992) had suggested. In fact, doing so actually resulted in less stable results when the number of specification errors exceeded the number of allowable modifications.

Despite the criticism of the χ^2 test as being inflated at large sample sizes, the stopping criterion of overall model nonsignificance proved to be fairly useful when $N \geq 800$. Use of the p-value associated with the χ^2 test did not lead to overfitting the model as one might expect in the presence of large samples. However, when sample size was less than 800, specification searches halted at nonsignificance did tend to end before modifications could stabilize. Apparently, the MI did not necessarily identify the same parameters in the identical order across different samples, but did identify a set of parameters fairly consistently even if the order changed somewhat from sample to sample.

Based on the results of this study, an alternative to the number of modifications or overall nonsignificance as stopping criteria might be inspection of the MIs for a distinctive drop in successive values of the maximum MI. In the present study, patterns of MIs closely paralleled the results achieved using model nonsignificance as the stopping criteria. Conditions with marked declines in values of MIs were also those that exhibited greater modification consistency, while conditions with more ambiguous patterns of MIs were the ones found to be unstable using the other stopping criteria. This finding challenges the argument by Saris et al. (1987) and Kaplan (1988; 1990) that the MI does not reliably detect important misspecifications.

Because MIs share the same sample size dependency with the overall χ^2 test, it is the relative rather than absolute size of the MIs that should be considered. When values of MIs seem to gradually decrease, even if still statistically significant, it suggests that there may be a number of specification errors present, but none of substantial size. Errors of this type are more likely to reflect chance characteristics of the data. Consequently, in practice one should probably try to limit model modifications to correction of noticeably large specification errors which would be more apt to replicate in other samples.

Several limitations of this study should be noted. First, the decision to limit population models to two CFA models will restrict generalizability of results to models of this type. Related to this was the imposition of only one kind of misspecification, i.e., incorrect omission of complex factor loadings.

Another limitation was the use of only the MI for conducting specification searches. Kaplan (1989) and Tippets (1991) have shown that the use of the expected parameter change statistic (EPC) or its standardized version (SEPC) in tandem with the MI offers a promising new strategy for conducting searches. The question of interest is whether or not the EPC and SEPC will be as subject to sampling variability as the MI. Thus far studies of the EPC/SEPC have been conducted with population data or in single empirical samples, where sampling variability was absent.

A potential weakness in the design was possible confounding of model and level of misspecification resulting from the nesting of number of specification errors within model. While it is unlikely that, say, two errors would have had the same impact on the four-factor model as it did on the two-factor model, it is also not known if doubling the number of errors in the four-factor model provided the same relative level of misspecification. While rates of population model recovery and stability of model modifications were slightly better in the four-factor model, rates were not so different as to suggest substantial disparity in misspecification. Conversely, measures of fit suggested better fit for the two-factor model at all levels of misspecification. However, there is no way to tell if this was due to more severe misspecification in the four-factor model resulting from a design flaw, or if the differences in fit were simply reflecting a model complexity effect.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analyses. Psychometrika, 49, 155-173.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411-423.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction part 1 (pp. 149-173). Amsterdam: North-Holland Publishing Co.

Boomsma, A. (1987). The robustness of maximum likelihood estimation in structural equation models. In P. Cuttance & R. Ecob (Eds.), Structural modeling by example: Applications in educational, and social research (pp. 160-188). Cambridge, U.K.: Cambridge University Press.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? Psychological Bulletin, 107, 260-273.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. Multivariate Behavioral Research, 24, 445-455.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, lagrange multiplier, and wald tests. Multivariate Behavioral Research, 25, 115-136.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. Multivariate Behavioral Research, 18, 115-126.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-157.

Farley, J. U., & Reddy, S. K. (1987). A factorial evaluation of effects of model specification and error on parameter estimation in a structural equation model. Multivariate Behavioral Research, 22, 71-90.

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. Journal of Consumer Research, 9, 443-448.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. Journal of Consumer Research, 11, 572-580.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. Multivariate Behavioral Research, 20, 255-271.

Hutchinson, S. R. (1993). Univariate and multivariate specification search indices in covariance structure modeling. Journal of Experimental Education, 61, 171-181.

Jöreskog, K. G., & Sörbom, D. (1988). LISREL 7: A guide to the program and applications. Chicago: SPSS Inc.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. Multivariate Behavioral Research, 23, 69-86.

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. Multivariate Behavioral Research, 24, 285-305.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. Multivariate Behavioral Research, 25, 137-155.

Kaplan, D., & R. N. Wenger (1993). Asymptotic independence and separability in covariance structure models: Implications for specification error, power, model modification. Multivariate Behavioral Research, 28, 467-482.

Leamer, E. E. (1978). Specification searches: Ad hoc inference with nonexperimental data. New York: John Wiley and Sons.

MacCallum, R. (1986) Specification searches in covariance structure modeling. Psychological Bulletin, 100, 107-120.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, 111, 490-504.

Matsueada, R. L., & Bielby, W. T. (1986). Statistical power in covariance structure models. In N. B. Tuma (Ed.), Sociological Methodology 1986 (pp. 120-158). San Francisco: Jossey Bass.

Muthén, B. O. (1988). LISCOMP analysis of linear structural equations with a comprehensive measurement model. Mooresville, IN: Scientific Software, Inc.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. Clogg (Ed.), Sociological Methodology 1987 (pp. 105-129). San Francisco: Jossey Bass.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

Silvia, E. S., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. Multivariate Behavioral Research, 23, 297-326.

Sörbom, D. (1989). Model modification. Psychometrika, 54, 371-384.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

Tippets, E. A. (1991). A comparison of methods for evaluating and modifying covariance structure models. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Table 1

Number of Times Population Models Recovered Per Cell

	Level of Misspecification			
N	1	2	3	4
		Two-Factor Model		
200	23	26	0	6
400	64	78	2	35
800	94	93	2	76
1,200	94	93	0	89
		Four-Factor Model		
200	19	30	8 (44)	8 (29)
400	64	71	36 (98)	40 (78)
800	96	99	85 (100)	78 (91)
1,200	100	100	93 (100)	92 (99)

Note. Numbers in each cell are out of 100 possible. Levels 3 and 4 of the four-factor model represent number of population models recovered if searches had been continued to nonsignificance rather than being terminated at four modifications. With the limit of four modifications imposed, the eight omitted population parameters could not have been recovered by four modifications. Values in parentheses represent number of population models recovered if searches were continued to eight modifications.

Table 2
Frequency of Most Commonly Occurring Final Model Per Cell

	Level of Misspecification			
N	1	2	3	4
		Two-Factor Model		
200	23[a]	41	6	10
400	64[a]	78[a]	8	35[a]
800	94[a]	93[a]	11	76[a]
1,200	94[a]	93[a]	25	89[a]
		Four-Factor Model		
200	19[a]	30[a]	7	16
400	64[a]	76[a]	6	44
800	96[a]	99[a]	6	57
1,200	100[a]	100[a]	7	60

Note. Due to the limit of four modifications, the four-factor Level 3 and 4 models could not have recovered the population model. Numbers are out of 100 possible.
[a] denotes population model.

Table 3
Number of Different Parameters Freed Per Cell

	Level of Misspecification			
N	1	2	3	4
		Two-Factor Model		
200	13	16	24	26
400	10	10	24	18
800	8	6	20	9
1,200	8	8	17	5
		Four-Factor Model		
200	15	6	17	33
400	4	4	8	16
800	4	4	8	12
1,200	4	4	8	9

Note. Smaller numbers reflect greater modification stability.

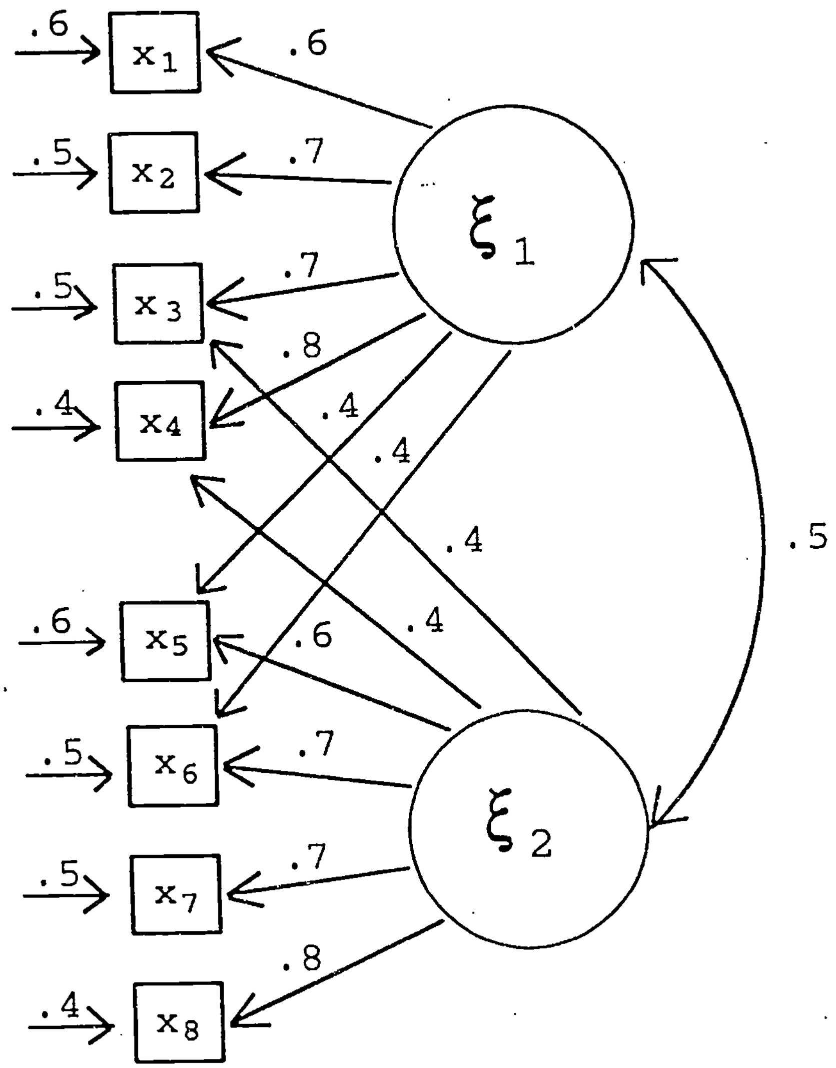
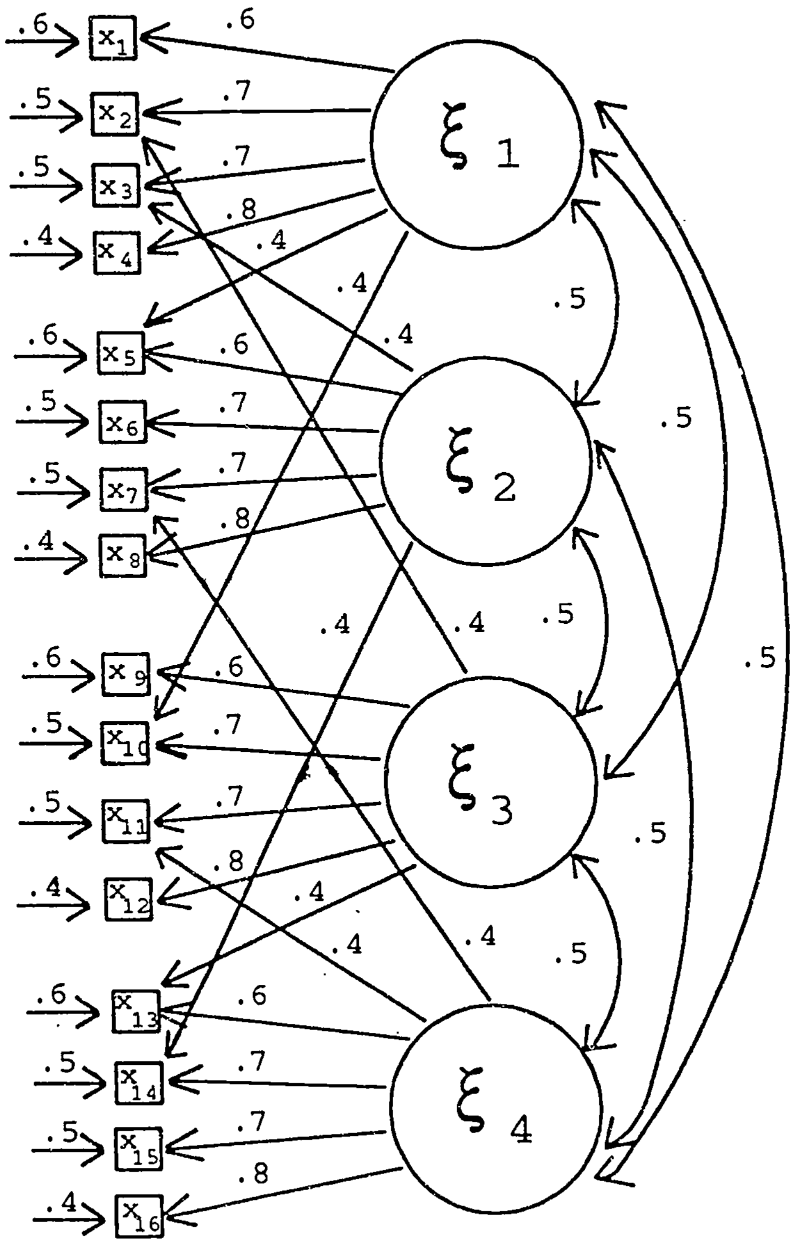


Figure 1. Population Model A



<u>Figure 2</u>. Population Model B