Model Modifications in Covariance Structure Analysis: The Problem of Capitalization on Chance

Robert C. MacCallum, Mary Roznowski, and Lawrence B. Necowitz
Ohio State University

In applications of covariance structure modeling in which an initial model does not fit sample data well, it has become common practice to modify that model to improve its fit. Because this process is data driven, it is inherently susceptible to capitalization on chance characteristics of the data, thus raising the question of whether model modifications generalize to other samples or to the population. This issue is discussed in detail and is explored empirically through sampling studies using 2 large sets of data. Results demonstrate that over repeated samples, model modifications may be very inconsistent and cross-validation results may behave erratically. These findings lead to skepticism about generalizability of models resulting from data-driven modifications of an initial model. The use of alternative a priori models is recommended as a preferred strategy.

During the past 10 to 15 years, covariance structure modeling (CSM) has become a widely used and important quantitative method in psychology, as well as in many other fields. We examine one particular commonly used procedure in CSM: the process of modifying models. In applications of CSM, researchers typically begin with one or more initial models and use a computer program such as LISREL (Jöreskog & Sörbom, 1988) or EQS (Bentler, 1989) to fit those models to sample data and obtain estimates of model parameters. If the fit of an initial model is considered inadequate, it has become common practice to modify the model so as to improve its fit to the data. This process, sometimes called a *specification search* (Kaplan, 1988; Leamer, 1978; Long, 1983; MacCallum, 1986), usually involves adding one or more parameters to the model in such a way as to improve goodness of fit maximally. The resulting modified model is then interpreted and offered as a good-fitting explanation of the relations among the measured and latent variables.

We wish to examine some fundamental issues involved in the specification search process. One issue concerns the consistency of model modifications over repeated samples. That is, if a specification search were conducted using the same initial model on different samples from the same population, how consistent would the specific model modifications be from sample to sample? (We use the term *stability* to refer to this notion of consistency of model modifications across repeated samples.) A second issue involves cross-validation of modified models. When a model is modified on the basis of results from a particular sample, how well will that modified model fit an independent sample from the same population? More generally, how well do models arising from specification searches generalize to the entire population? Clearly, these questions have important implications with respect to the validity of modified models. If model modifications are unstable, do not cross-validate well, and do not generalize to the population, researchers can have little confidence in the validity of models arising from specification searches.

We discuss the basis for these concerns and review published applications of CSM using specification searches. Using two large sets of empirical data, we investigate how well modified models generalize to other samples and to the population. Finally, we discuss implications of our results for the evaluation of models produced by specification searches in empirical applications of CSM.

Capitalization on Chance in Specification Searches

A desirable outcome in a CSM analysis is to find that the model under investigation fits well, that it cannot be simplified substantially without significant loss of overall fit, and that its fit cannot be improved to any great extent by making the model more complex. Under such circumstances, the model could be viewed as a plausible explanation of the relations among the measured and latent variables in the population. It is critical to understand that one does not seek or expect to find a model that may be considered as precisely true or correct in the population. No model fits real-world phenomena exactly. This fact cannot be overcome by increasing sample size, improving measurement precision, or modifying a model. The best one can hope for is to show that a model provides a good approximation to real-world phenomena, as represented in an observed set of data.

When the initial model of interest does not satisfy this objec-

We wish to thank Michael W. Browne and three anonymous reviewers for their helpful comments on drafts of this article. We thank Howard Miller for his efforts in collecting the first data set analyzed in this article. We also wish to thank the Human Sciences Research Council, Pretoria, South Africa, for permission to use data collected under their auspices by W. Verhoef and W. L. Roos.

Correspondence concerning this article should be addressed to Robert C. MacCallum, Department of Psychology, Ohio State University, Columbus, Ohio 43210.

Psychological Bulletin, 1992, Vol. 111, No. 3, 490-504
 0033-2909/92/$3.00

tive, researchers often conduct a specification search so as to alter the model to improve its fit to the data. As indicated above, our concern regarding this process involves whether such modified models generalize to other samples or to the population. At the heart of this concern is the issue of *capitalization on chance.* A specification search must be recognized as a data-driven process in that selected modifications of the initial model are based in part, if not entirely, on results obtained from fitting the model to a particular sample. Therefore, the specification search process is inherently susceptible to capitalization on chance in that idiosyncratic characteristics of the sample may influence the particular modifications that are performed. As emphasized by Cliff (1983), given the complexity of covariance structure models and correlational data, it is highly likely that in any particular case there will be model modifications available that would substantially improve the fit of the model to the data. However, such modifications may merely fit chance characteristics of the original sample, rather than represent aspects of the model that generalize to other samples and to the population.

Whenever a researcher modifies a model using a strategy that is in any way data driven, the issue of potential capitalization on chance must be of concern and must be addressed in some manner. This problem exists regardless of the type of search strategy used and regardless of the nature of the initial model and purpose of the research. Many different data-driven search strategies can be defined, all of which are inherently susceptible to this problem. For instance, alternative strategies could define different priorities as to what aspects of a model should be considered for modification first (e.g., measurement vs. structural portions of the model), or might exclude certain parameters from consideration for inclusion in a model (e.g., factor loadings fixed at zero in the initial model). Although future studies might show that certain strategies are less susceptible to capitalization on chance, it cannot be argued reasonably that a particular strategy that is in some way data driven is immune to this problem.

Likewise, the issue of capitalization on chance is relevant whether CSM is used in research that is primarily confirmatory or exploratory in nature. In a more exploratory study a researcher may construct a rough model in a substantive area in which little is known about the structure of relationships among variables and then explore variations of that model. In a highly confirmatory study a researcher may use a model that is well grounded in theory and prior research. Both cases represent legitimate uses of CSM, proceeding from exploratory model development to the evaluation of well-founded models. Considering the issue of specification searches, it is reasonable that model modifications might be approached less conservatively in more exploratory studies. That is, consideration of model modifications that would enhance the fit of a model to a given set of data would be a more natural aspect of research involving exploratory model development. However, this fact does not imply that modifications made to models in these types of studies are any less susceptible to capitalization on chance than in more confirmatory studies. Exploratory research is a desirable and necessary part of the process of model development. Nevertheless, the issue of capitalization on chance must be of concern whenever data-driven model modifications are carried out, regardless of whether the research is primarily confirmatory or exploratory in nature.

There are, of course, factors operating in empirical applications that heighten the concern about this problem. One of these factors is sample size. When sample size is small, the increased sampling variability in sample correlations or covariances could have substantial effects on results of CSM analyses, including the selection of model modifications. That is, the particular modifications carried out in a specification search could be quite unstable from sample to sample. This phenomenon is explored in the sampling studies to be presented. An important question involves what is an adequate sample size to gain some protection against this problem. The general issue of sample size in CSM is recognized to be a complex problem, with necessary sample size being dependent on a number of factors including the complexity of the model. In the context of specification searches, results presented by MacCallum (1986) show uniformly poor outcomes of searches that were based on sample sizes of 100 and only mediocre success when sample size was 300. We present results showing that the problem of capitalization on chance in specification searches may be quite severe even in samples of 300–400 cases and may have some impact even when sample size is quite large (e.g., 1,200 cases). Thus, in most empirical applications, sample sizes may not be sufficiently large to give much protection against this problem.

Another factor that affects the degree of concern about capitalization on chance in specification searches involves the number of modifications made to an initial model. Modifications are often made sequentially, with each successive modification selected so as to provide maximum improvement in overall model fit. Thus, the early modifications correct for more severe aspects of lack of fit, and later modifications correct for smaller sources of lack of fit. In reasonably large samples, large discrepancies between the model and the data would tend to be more stable, and smaller discrepancies would tend to be less stable. An important implication of this phenomenon is that searches characterized by relatively more modifications are likely to be more strongly influenced by chance characteristics of the data. Supportive evidence for this claim is provided by MacCallum (1986), who demonstrated that longer searches have little chance of correctly identifying model misspecifications.

A third factor that affects the degree of concern about capitalization on chance is the interpretability of model modifications. The methodological literature in CSM is replete with warnings that modifications must be substantively justified (e.g., Jöreskog & Sörbom, 1988; Long, 1983; MacCallum, 1986; Saris & Stronkhorst, 1984; Sörbom, 1989). If a parameter is to be added to a model, the researcher must be able to provide a clear substantive interpretation of that parameter. This recommendation is intended to prevent the addition of meaningless parameters to a model simply for the purpose of improving goodness of fit to a particular sample.

Unfortunately, this issue of interpretability is problematic in practice. Even when substantive justifications for model modifications are offered, one may be concerned as to the rigor and validity of those justifications. Steiger (1990) expresses strong concern about this problem as follows: "What percentage of

researchers would find themselves unable to think up a 'theoretical justification' for freeing a parameter? In the absence of empirical information to the contrary, I assume that the answer . . . is 'near zero' " (p. 175). Even more troubling is the fact, to be supported later, that in the vast majority of empirical studies little or no substantive interpretation is offered for model modifications. In this context, there appears to be a belief that it is acceptable to add parameters to a model so as to improve the fit of the model to an adequate level, without the need to attach substantive meaning to these parameters. Such parameters have been termed *wastebasket* parameters (e.g., Browne, 1982); a common example is the covariances among error terms. Our view is that users of such an approach are sometimes trying to have it both ways: They want a model that fits their data well, but without the responsibility of interpreting the changes made to achieve that fit. Clearly such an approach must raise serious concerns about the possibility that such modifications are merely capitalizing on chance characteristics of the data. A requirement that a clear and well-founded interpretation be offered for any modification would provide some protection against this problem.

However, such a requirement raises an important question: If a model modification is highly interpretable, then why was it not represented in the initial model? In response, one cannot rule out the possibility that interpretable modifications to a model may occur to a researcher only in the process of evaluating an analysis of that model in a particular sample. This may be the case even in highly confirmatory studies. In more exploratory studies a researcher may have a variety of ideas about potential relationships among variables and may wish to use the specification search process to evaluate what relationships are supported by the data. In any case, however, it is clear that model modifications should be supported by clear substantive interpretation. Given the difficulties associated with this issue, it is worthwhile to consider an alternative strategy: constructing and evaluating alternative a priori models. This strategy is discussed and strongly endorsed in the final section of this article.

We have discussed three factors that are directly relevant to the phenomenon of capitalization on chance in specification searches: sample size, the number of model modifications, and the interpretability of those modifications. The worst circumstance would involve a search conducted on a small sample and introducing many modifications that are given little or no interpretation. In such a case, there must be considerable skepticism about the generalizability of the resulting model beyond the sample at hand. However, even under the best circumstances, when sample size is large and only one or two interpretable modifications are made, one can neither ignore nor avoid the possibility that the modifications were the result of chance characteristics of the sample. As stated earlier, whenever model modifications are made using a data-driven process, this concern must be addressed.

There are two primary ways to confront this problem. One way would be to provide evidence that the model modifications and final model generalize beyond the sample at hand. Such evidence could be provided by some type of cross-validation analysis. In the methodological literature on CSM, researchers are routinely warned that modified models must be cross-validated (e.g., Bentler, 1980; Bollen, 1989; Breckler, 1990; Cliff, 1983; Hayduk, 1987; MacCallum, 1986; Saris & Stronkhorst, 1984; Sörbom, 1989). However, our review of published applications shows that this advice is rarely followed in practice. The sampling studies to be presented later include results of cross-validation analyses. On the basis of those findings, our recommendation is that cross-validation of a model resulting from a specification search should involve parallel searches conducted on independent samples. That is, the initial model should be fit to independent samples, and the model modification process should be conducted in both samples. Specific modifications can then be evaluated for consistency between samples, and final modified models arising in each sample can be fit to the other sample. Cudeck and Browne (1983) developed a two-sample cross-validation index that can be used for this purpose. Results of such analyses would provide important information about stability and cross-validity of model modifications.

If sample size or other factors make a cross-validation analysis impractical, researchers must clearly state the limitations and the need for further evaluation of models produced by specification searches. Such a statement should include the points (a) that the initial model was modified to improve its fit to one sample, (b) that the generalizability of those modifications to other samples and to the population remains to be determined, and (c) that the resulting model cannot be considered plausible until it is further evaluated using independent samples. Such a statement may often be applicable in exploratory studies using CSM but should not be considered a "kiss of death" that invalidates results of such studies. Rather, such a statement is a simple and responsible acknowledgment that further evaluation of models produced by specification searches is necessary. In summary, it is inappropriate to evaluate a model resulting from a specification search in a single sample as if it has been validated and is a plausible explanation of relations among measured and latent variables in the population.

Strategies for Specification Searches

An abundance of information is produced in a typical CSM analysis, and model modifications are usually based on some specific aspects of this information. A variety of general principles and strategies could be defined for conducting specification searches, but a few simple approaches are commonly used. Researchers using LISREL (Jöreskog and Sörbom, 1988) almost always focus on *modification indexes* (Sörbom, 1989), which are provided for each model parameter that is assigned a fixed numerical value in the initial model. The value of a given modification index indicates the minimum magnitude by which the overall likelihood ratio χ^2 value for the model would decrease if the corresponding parameter were freed. Researchers often use this information to conduct a sequence of model modifications. At each step, a parameter is freed so as to produce the largest improvement in fit, and the process is continued until adequate fit is achieved. We refer to this widely used approach as *sequential model modification.*

Whereas LISREL remains the most widely used software for CSM, other programs are available that provide other types of information that can be used for model modification. For in-

stance, Bentler's (1989) EQS program provides the Lagrange multiplier test (Lee & Bentler, 1980; Satorra, 1989) for testing whether a set of fixed parameters, if freed, would significantly improve the fit of the model. Researchers might also wish to consider measures of expected parameter change, as recommended by Kaplan (1990), in determining whether to free a fixed parameter.

Besides the variety of information that can be used to determine optimal model modifications, there are also a variety of strategies available for selecting model modifications. For instance, a researcher may establish priorities with regard to the sequence in which types of parameters are considered for addition to the model. One such approach (Anderson & Gerbing, 1982; Silvia & MacCallum, 1988) is based on the notion that new parameters should be added to the measurement model first, then to the structural model. Other strategies might call for the exclusion of certain parameters from consideration for addition to the model. For example, one might argue that the measurement model should be considered fixed, thus disallowing any new parameters representing effects of latent variables on indicators.

We wish to make several points about alternative strategies for model modification. First, each strategy will have limitations in terms of its capability for diagnosing and correcting misspecifications. Some types of misspecifications (e.g., nonlinear influences among variables) would be quite difficult to diagnose by conventional procedures. Second, any strategy that is in any way data driven is susceptible to problems arising from capitalization on chance. Regardless of the particular information and strategy used, specification searches can almost always be conducted so as to achieve an adequate fit through a sufficient number of modifications of the initial model. We do not address the question of whether some search strategies produce modified models with better stability and cross-validity than models produced by other strategies. Rather, we argue that stability and cross-validity are critical issues whenever data-driven model modifications are conducted, regardless of strategy. We present results of an investigation of one particular strategy: sequential model modification. We focus on this approach because it is widely recognized due to its availability through LISREL and widely used in published applications.

Review of Published Applications

It is important to consider how the issues we have discussed are reflected or addressed in published applications of CSM. Breckler (1990) reviewed 72 applications of CSM in four journals (*Journal of Personality and Social Psychology, Journal of Experimental Social Psychology, Personality and Social Psychology Bulletin,* and *Psychological Review*) for the period 1977–1987. To augment Breckler's data, we identified an additional 28 applications of CSM published in the *Journal of Applied Psychology* during the period 1988–1990, yielding a total of 100 applications. A list of the 72 papers considered by Breckler can be found in an appendix to his article; a list of the 28 additional papers considered presently can be obtained from Robert C. MacCallum by request. Keep in mind that there are many more applications published in other psychology journals and journals in other fields such as sociology, political science, marketing, and education.

Of the 100 studies examined, 37 contained acknowledgments of having modified an initial model to improve its fit to the data. A number of these applications were based on relatively small samples. Eleven of the 37 studies in which specification searches were conducted were based on samples of less than 200, and 7 of those 11 were based on samples of less than 125. Although such samples may not seem especially small from an experimental perspective, they may be too small to obtain stable results from CSM analyses and specification searches, as discussed earlier.

A second concern regarding these published applications is the fact that many searches involved substantial numbers of modifications of initial models. It is not unusual for researchers to incorporate a large number of new parameters into a model to improve its fit to an adequate level. For example, in a confirmatory factor analysis of measures of job characteristics, Kulik, Oldham, and Langer (1988) freed 8 parameters on the basis of modification indexes. Farkas and Tetrick (1989) freed 11 parameters in each of two analyses in their study of models of turnover decisions. Newcomb, Huba, and Bentler (1986) freed 37 covariances among error terms to improve fit of a model of sexual and dating behaviors. We found many cases where 5 or more parameters were added to a model through a specification search. As discussed earlier, concerns about capitalization on chance and lack of generalizability must be heightened as the number of model modifications increases.

A third factor previously discussed involves the interpretability of model modifications. Of the 37 applications examined, 31 contained modifications made strictly on the basis of the data (almost always using modification indexes or closely related statistics). Thus, only 6 of the 37 provided a justification of modifications on substantive grounds. These findings indicate that the warning regarding substantive justification of modifications is routinely ignored. Furthermore, we found almost no evidence of researchers declining to make a specific modification because of lack of interpretability. In fact, we found only one study (Meyer & Gellatly, 1988) in which it was stated that a model modification was not made because it would not make sense theoretically to do so.

The picture with regard to cross-validation of modified models is gloomy also. Of the 37 studies that acknowledged model modifications, we found only 4 that provided some type of information about cross-validation. Two of these studies (Reisenzein, 1986; Tanaka & Huba, 1984) fit a model resulting from a specification search to a new sample. Two other studies (Hinkin & Schriesheim, 1989; Schriesheim & Hinkin, 1990) used confirmatory factor analysis in one sample to aid in selecting items for a scale, then validated the resulting item set and model in an independent sample. The details of these analyses are not critical here. What is important is that cross-validation of modified models is rarely done in practice. Given the clear risk of capitalization on chance, coupled with the usual neglect of substantive justification of modifications, the lack of information about cross-validity raises serious concerns about the generalizability of modified models.

To summarize, our examination of 37 published applications

of CSM involving specification searches demonstrates that the cautions and warnings regarding model modification are routinely ignored or overlooked. Model modification in practice is usually done with no substantive justification and no cross-validation, often involves a substantial number of modifications, and is often based on samples that may be too small for such analyses. Every published application we examined could be faulted on at least two of these counts.

Nevertheless, model modification in this manner seems to have become accepted practice. We consider this to be an unfortunate state of affairs, representing a dangerous and misleading methodological trend. Our objective is to shed light on some serious negative consequences of common approaches to model modification. Our study of this general problem focused on two issues. The first involves the degree to which model modifications are consistent across repeated samples. The second involves the degree to which models modified to fit one sample will fit an independent sample. We investigated how these characteristics of stability and cross-validity are affected by sample size. Our study of these issues was conducted using two large sets of empirical data.

Study 1

The initial model used in the analyses in Study 1 was a model of employee behavioral and cognitive responses to job attitudes and perceptions of job conditions. The model was in part an adaptation of the heuristic model of employee responses to affect from Hulin, Roznowski, and Hachiya (1985). A general hypothesis of the Hulin et al. model is that work role dissatisfaction motivates individuals to do something to alleviate the dissatisfaction. The three independent latent variables in the model were employee satisfaction toward pay, employee satisfaction toward work, and perceptions of job characteristics. The four dependent latent variables were withdrawal, citizenship, change-oriented behavior syndromes or behavioral "families," and cognitions/intentions regarding physical withdrawal from the organization in the future. The latent variables were defined as follows:

Pay satisfaction: Attitude toward the individual's pay and the pay system in the organization.

Work satisfaction: Attitude toward the individual's work.

Job conditions: Individual perceptions of the physical working conditions in the workplace and job environment.

Withdrawal: A general syndrome reflecting psychological withdrawal or passive withdrawal of the individual from the workplace. Behaviors intended to distance the employee from the organization or the work itself are represented by this variable (Rosse & Miller, 1984; Roznowski & Hulin, 1991).

Citizenship: A general syndrome of individual behavior reflecting positive, pro-organizational acts such as volunteering and extra-role behaviors (C. A. Smith, Organ, & Near, 1983).

Change: A broad-based behavioral syndrome reflecting individual behaviors oriented toward changing the work situation to improve work conditions, the job itself, or the workplace.

Withdrawal cognitions: A general measure reflecting the employee's thoughts and intentions about future withdrawal from the organization in the form of quitting, being late, being absent, or transferring.

A relatively complex model was constructed to represent relations among these latent variables. A path diagram depicting the model, including a representation of multiple indicators for each latent variable, is shown in Figure 1. The relations to be tested were chosen to represent various hypotheses from the attitude and behavior literature relevant to individuals in complex organizational settings. First, the three independent latent variables (work satisfaction, pay satisfaction, and job perceptions) were allowed to correlate. Work satisfaction was hypothesized to influence all four behavior syndromes (change, citizenship, withdrawal, and cognitions). Theoretical justification for these links is extensive (Hulin, 1991; Hulin et al., 1985; Rosse & Miller, 1984). Perceptions of job conditions were hypothesized to influence only the most extreme form of behavior (cognitions/intentions to withdraw physically from the organization). Pay satisfaction was hypothesized to lead to both withdrawal behaviors and cognitions about future withdrawal. Finally, the citizenship behavior latent variable was hypothesized to influence change behaviors. Likewise, the withdrawal latent variable was hypothesized to lead to cognitions about withdrawal in the future.

The indicators for the latent variables were obtained from a large set of self-report measures assessed through questionnaire. The sample consisted of 3,694 employees of two large hospitals located in the Midwest. The two organizations were similar in size, employee demographic makeup, unionization, and other features.

All latent variables had multiple indicators. Each indicator was a *parcel,* which is a simple unit-weighted sum of a number of items. The use of parcels served to reduce the total number of items to a manageable level and to provide indicators with higher reliability than that of single items. Items for a particular indicator parcel were chosen so as to balance content as well as psychometric characteristics of the items across indicators. Attitude measures were taken from the Job Descriptive Index (JDI; P. C. Smith, Kendall, & Hulin, 1969). For the two satisfaction latent variables (work and pay), the item-grouping procedure used by Drasgow and Kanfer (1985) was used to obtain three indicators each for work and pay satisfaction. The work satisfaction indicator parcels each contained five JDI work items; the pay indicator parcels each contained three JDI pay items. Items suggested by Roznowski (1989) were included in place of a few JDI items with poor measurement properties. The two indicators for the job perceptions variable were constructed by summing, respectively, four items assessing perceptions of job conditions and eight items assessing perceptions of the physical environment and equipment.

A variety of self-report items was used for the dependent variables to assess a broad range of employee behaviors from very negative to less severe behaviors to positive, pro-organizational types of acts (Rosse, 1983). Eighteen items reflecting passive or psychological withdrawal were combined to form three parcels to serve as the indicators of the withdrawal latent variable. Examples of these items include "doing poor quality work," "arguing with co-workers," and "refusing to do assigned work." Twelve pro-organizational acts were used to define three

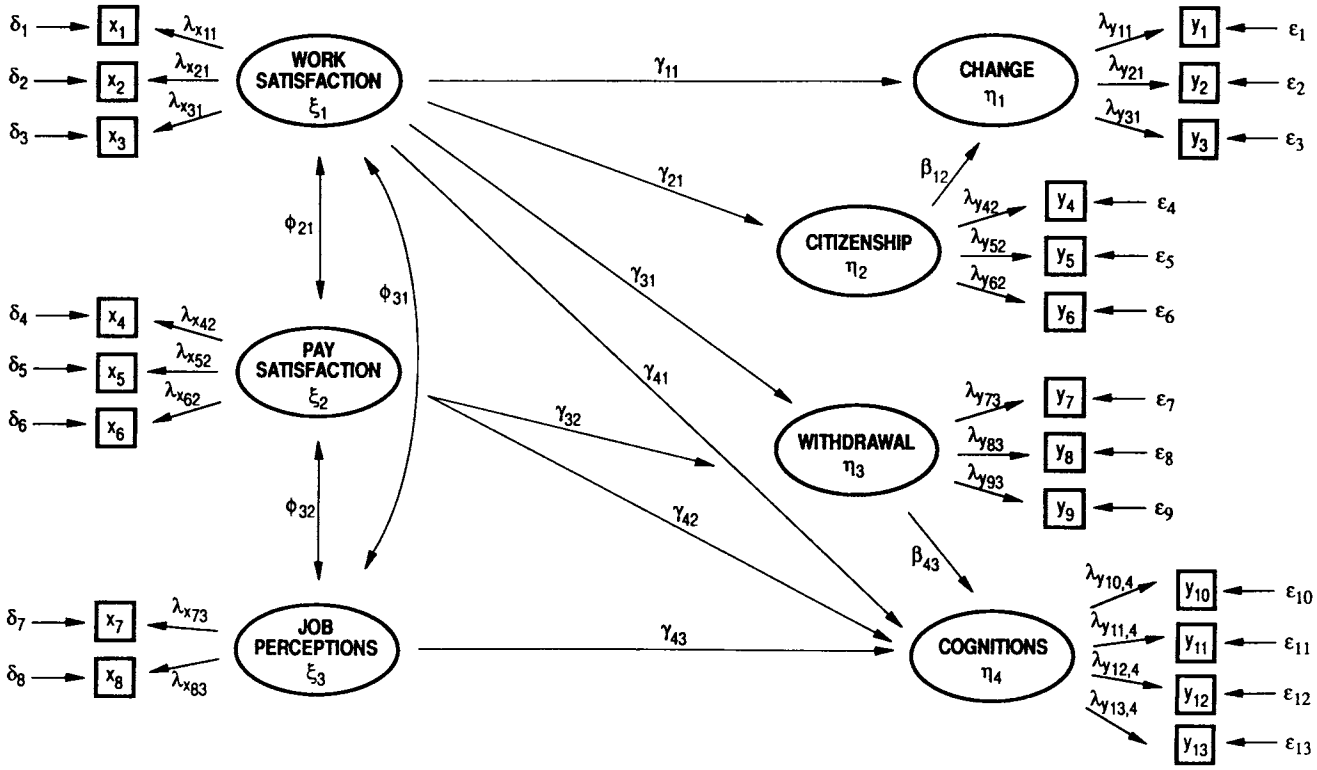

Figure 1. Path diagram of model of employee responses to job attitudes and perceptions of job conditions.

parcels to serve as indicators of the citizenship latent variable. Such items contained content reflecting positive acts committed by employees, such as "volunteering to do extra work" and "giving encouragement to new employees." Next, parcels reflecting change-oriented behaviors were constructed following notions from Hulin et al. (1985). Three indicators each containing three change-oriented behaviors were used to define the change latent variable. Examples of these behaviors include "filing a grievance" and "making suggestions for needed change." Finally, four indicators representing general cognitions and intentions about future withdrawal from the organization were created. These four parcels consisted, respectively, of cognitions and intentions about quitting, being late, being absent, or transferring out of the workplace.

The data matrix for the full sample consisted of scores on a total of 21 manifest variables for the entire sample of 3,694 persons. Both organizations were combined, because we decided that the employees in the two hospitals were similar enough to be a single population.

Method

This initial model and data set were used in a sampling study to investigate stability and cross-validity of model modifications. We used eight levels of sample size: 100, 150, 200, 250, 325, 400, 800, and 1,200. These levels were selected on the basis of (a) our review of published applications of CSM, which showed sample sizes ranging from very small to very large, but usually between 100 and 350 and (b) our pilot studies, which showed a need to represent a wide range of sample sizes to reveal important phenomena about the issues under study.

For each level of sample size, 10 pairs of random samples were drawn from the total sample of 3,694. Within each pair of samples, we designated one sample as the calibration sample (sample *a*) and the other as the cross-validation sample (sample *b*). Sample covariance matrices $\mathbf{S}_a$ and $\mathbf{S}_b$ were obtained. The initial model was then fit to $\mathbf{S}_a$, using the maximum likelihood method provided in LISREL 7.16 (Jöreskog & Sorbom, 1988). Various information about goodness of fit of the initial model was recorded. For purposes of the present study, we used two descriptive indexes of fit. The first was the nonnormed fit index (NNFI; Bentler & Bonett, 1980), which is a generalization of the Tucker-Lewis coefficient developed for maximum likelihood factor analysis (Tucker & Lewis, 1973). The second was Steiger's (1989) root-mean-square error of approximation, or RMSEA. This measure is a standardized root-mean-square residual, corrected for model complexity. It provides an estimate of badness of fit in the population.

After the initial model was fit to a given calibration sample covariance matrix, a sequence of modifications of the model was carried out using modification indexes provided by LISREL. At each step of the modification process, the fixed parameter with the largest modification index was freed. We are aware that this mechanical process violates the often repeated warning that model modifications must be substantively justified. (Note that an attempt was made to prevent meaningless modifications of the initial model. Before carrying out any specification searches, fixed parameters were identified that

would not be freed, because doing so would be uninterpretable. However, no search produced a modification index that indicated that any of these parameters should be freed. Therefore, the net effect was a search procedure that was essentially mechanical.) Thus, we used a specification search process that we believe is representative of procedures often used in practice, namely, a mechanical approach without rigorous justification for specific modifications. In each calibration sample, a sequence of four modifications was conducted. Results of pilot work indicated that a four-step search was adequate to reveal relevant phenomena. Additional steps served little purpose.

After each modification in each search, the modified model was refit to the calibration sample covariance matrix, $\mathbf{S}_a$. Each specific modification was recorded, and measures of goodness of fit of each modified model to $\mathbf{S}_a$ were calculated. Obviously, goodness of fit to the calibration sample would improve with each modification.

For the initial model and each modified model, two cross-validation indexes were calculated. The first was the two-sample index proposed by Cudeck and Browne (1983). When a particular model, k, is fit to $\mathbf{S}_a$, one can obtain the reconstructed covariance matrix, $\hat{\mathbf{\Sigma}}_{k|a}$. Given $\mathbf{S}_a$, the value of the maximum likelihood fitting function, $F(\mathbf{S}_a; \hat{\mathbf{\Sigma}}_{k|a})$ is minimized. This function measures the correspondence between $\mathbf{S}_a$ and $\hat{\mathbf{\Sigma}}_{k|a}$, and parameters are estimated so that the resulting $\hat{\mathbf{\Sigma}}_{k|a}$ yields the minimum value of the fitting function. The two-sample cross-validation index is defined as the value of the fitting function measuring the correspondence between $\hat{\mathbf{\Sigma}}_{k|a}$ and the covariance matrix for the cross-validation sample, $\mathbf{S}_b$, namely, $F(\mathbf{S}_b; \hat{\mathbf{\Sigma}}_{k|a})$. We refer to this index as CV2, indicating a two-sample cross-validation index. Smaller values of CV2 indicate better cross-validity. Logistically, this index could be obtained by fitting model k to $\mathbf{S}_b$, where all parameters are fixed at values obtained when model k is fit to $\mathbf{S}_a$ (though this was not the computational method used in the present study). This approach corresponds to the tight method of cross-validation mentioned by Bentler (1980).

To overcome the problem of needing two samples to obtain CV2, Browne and Cudeck (1989) proposed a single-sample cross-validation index. They developed the following measure:

$$\text{CV1} = F(\mathbf{S}_a; \hat{\mathbf{\Sigma}}_{k|a}) + 2q_k/n - p - 2) \quad (1)$$

where q_k is the number of free parameters in model k, p is the number of measured variables, and n is sample size. Browne and Cudeck show that CV1 provides a close approximation to values of CV2.

For each pair of samples drawn, we calculated CV1 and CV2 for the initial model and each of the four modified models. We also calculated the value of CV2 for each model using the covariance matrix for the total sample, $\mathbf{S}_t$, in place of $\mathbf{S}_b$. This index provided a measure of how well a solution obtained from a calibration sample would fit the total sample. Results obtained by computing these various cross-validation indexes allowed for the evaluation of cross-validity of models as they were modified to improve fit to the calibration sample. If the successively modified models cross-validate more poorly, then the cross-validation indexes should increase from one model to the next. Conversely, if cross-validation improves as modifications are carried out, then the indexes should decrease from one model to the next. Note that we used these indexes in a context different from that originally intended by Cudeck and Browne (1983) and Browne and Cudeck (1989). Their context involved assessing cross-validity of alternative a priori models. They urged researchers to use cross-validation indexes to evaluate which of a set of alternative a priori models would generalize best to other samples. We used the same indexes to evaluate relative cross-validity of alternative models produced by modification of an initial model. Our use of these indexes in this context produced one interesting phenomenon regarding CV1, which is discussed in the Results section.

To summarize the method used in Study 1, we defined an initial model and conducted a sampling study of the model modification process. For each of eight different levels of n, we drew 10 pairs of samples, each pair consisting of a calibration sample and a cross-validation sample. The initial model was fit to each of the 80 calibration samples and was then modified four times, each modification involving freeing the parameter with the largest modification index. Each specific modification was recorded. For the initial model and each modified model, we obtained goodness-of-fit measures, as well as two-sample and single-sample cross-validation indexes. Finally, the same procedure was followed for the full sample of $n = 3{,}694$. All of the same information was obtained for this sample, except for the two-sample cross-validation index. Thus, a total of 81 specification searches was conducted (80 subsamples, plus the total sample).

Results

Consider first the goodness of fit of the initial and modified models in the calibration samples. Tables 1 and 2 show summary statistics for the NNFI and RMSEA fit measures, respectively. Each table shows statistics for the 10 replications at each sample size, as well as for the full sample, for the fit of the initial model and the final model (i.e., after four modifications).

The mean values of the fit indexes for the initial model show that the initial model did not fit badly but was clearly in need of improvement. The mean values of fit indexes for the final model show that definite improvement in fit was achieved and that the final models fit the calibration sample moderately well. This pattern is fairly typical of applications using specification searches. That is, the fit of the initial model is inadequate, and modifications are conducted resulting in a final model that fits the data reasonably well. An interesting and somewhat surprising result was that the means of the fit measures for the initial model improved with sample size. This result was especially surprising with regard to the NNFI index, because it is generally considered to be relatively independent of sample size (Anderson & Gerbing, 1984; Balderjahn, 1988; Bollen, 1990; Marsh, Balla, & McDonald, 1988). The effect of sample size on mean NNFI for the initial model was quite pronounced. This issue is tangential to the focus of the present study but probably merits further investigation.

The range and standard deviations of the fit indexes in Tables 1 and 2 show that the fit of both the initial and final models was quite unstable at small sample sizes and still somewhat unstable even through moderately large sample sizes. For instance, note the rather wide range in values of NNFI for the final models at $n = 325$ (.874–.932). These results indicate that the benefits of a specification search, in terms of the fit of the final model, may be quite dependent on matters of sampling unless n is quite large. For instance, in the present study, if a specification search were conducted on a sample of $n = 325$, a lucky investigator might obtain a final model that fits quite well after a relatively short search, but an unlucky investigator might have to conduct a much longer search to achieve adequate fit.

Consider next the particular modifications made in each specification search. Table 3 shows the sequence of parameters freed in each of the 81 searches conducted. Parameters are indicated in Table 3 using LISREL program designations. For every modification in every search, the corresponding modification

Table 1
Summary Statistics for NNFI for Initial and Final Models

	NNFI for initial model				NNFI for final model			
n	High	Low	*M*	*SD*	High	Low	*M*	*SD*
100	.875	.781	.814	.024	.951	.842	.888	.030
150	.850	.803	.830	.015	.922	.860	.894	.017
200	.890	.834	.858	.016	.946	.890	.916	.016
250	.904	.811	.856	.030	.938	.882	.911	.020
325	.877	.816	.849	.019	.932	.874	.910	.016
400	.887	.834	.855	.015	.931	.898	.915	.011
800	.880	.832	.865	.014	.933	.903	.921	.010
1,200	.887	.857	.869	.009	.930	.909	.921	.007
3,694			.875				.930	

Note. NNFI = nonnormed fit index.

index had a value that was statistically significant at the .01 level, with almost all being significant at a much lower level. Thus, the modifications shown in Table 3 represent statistically significant steps in a search.

There are some striking results in Table 3. The specific modifications made in each search were found to be highly inconsistent across repeated samples, even in sample sizes as large as 400. In fact, although the modifications became more consistent as n increased, they were not completely consistent across repeated samples even at $n = 1,200$. The bottom row of entries in Table 3 shows the four modifications made when a specification search was conducted for the full sample. Considering the sequence of modifications made in the subsamples, and ignoring their order, note that at $n = 1,200$ only 6 of the 10 searches resulted in the same four modifications as in the total sample. For $n = 800$, this proportion was 4:10; for $n = 400$, 2:10; for $n = 325$, 1:10; and for n of 250 or less, none of the 40 searches produced the same four modifications as did the search in the full sample. Thus, modifications in subsamples were not found to generalize well to the total sample, which can be thought of as analogous to the population for present purposes. Another indication of instability in model modifications can be seen by counting the number of different parameters freed in the 10 searches at each sample size. Proceeding from $n = 100$ to $n = 1,200$, these numbers were as follows: 23, 17, 20, 15, 15, 12, 10, 7. By any criterion, Table 3 shows high inconsistency in model modifications across repeated samples of even moderately large size, as well as poor correspondence between modifications in samples and modifications in the population.

We next consider the behavior of the cross-validation indexes for the initial and modified models. Because values of CV1 and CV2 cannot be compared across different data sets, it would be inappropriate to compute and compare mean values of these indexes across replications within the various conditions. Instead, we examine the pattern of change in these indexes across each sequence of model modifications. For the two-sample cross-validation index, CV2, Table 4 shows the frequency of increases and decreases, across the 10 replications, at each modification and each sample size. For example, for the 10 replications at $n = 100$, CV2 increased four times at the first modification and decreased the other six times. The results in Table 4 show considerable instability in CV2 for sample sizes through 250. Though results for single samples are not shown in Table 4, inspection of such results showed CV2 to behave quite errati-

Table 2
Summary Statistics for RMSEA Fit Index for Initial and Final Models

	RMSEA for initial model				RMSEA for final model			
n	High	Low	*M*	*SD*	High	Low	*M*	*SD*
100	.097	.069	.086	.007	.081	.043	.067	.011
150	.089	.076	.082	.004	.071	.055	.065	.005
200	.082	.066	.074	.005	.067	.047	.057	.006
250	.085	.060	.074	.009	.068	.048	.058	.007
325	.083	.066	.076	.006	.069	.049	.058	.006
400	.080	.063	.074	.005	.063	.050	.057	.004
800	.079	.066	.071	.004	.060	.050	.054	.003
1,200	.074	.064	.070	.002	.057	.051	.054	.002
3,694			.070				.053	

Note. RMSEA = root-mean-square error of approximation.

Table 3
Sequence of Model Modifications for Each Sample In Study 1

		Sequence of modifications			
n	Rep	1	2	3	4
100	1	BE 4 1	LY 10 1	TE 2 1	LX 3 3
	2	TE 11 10	TE 13 4	LY 9 2	LX 3 3
	3	LY 10 3	BE 1 4	LY 9 2	TD 4 1
	4	LY 9 2	TE 13 10	TE 10 9	TD 7 4
	5	LY 3 4	LY 7 2	LY 11 3	BE 1 3
	6	LY 12 3	BE 4 1	LY 7 2	TE 8 4
	7	LY 9 2	LX 3 3	TD 5 1	TE 9 2
	8	LY 9 2	LY 10 3	LY 11 3	LY 3 4
	9	LX 3 3	LY 11 3	LY 7 2	LY 4 1
	10	LY 9 2	LX 3 3	BE 2 4	TD 5 1
150	1	LY 9 2	TE 9 5	LY 10 1	LY 13 1
	2	TE 13 10	LY 9 2	TE 11 8	TE 10 3
	3	LY 9 2	BE 1 4	LY 12 3	LY 11 3
	4	LY 9 2	TE 9 4	TE 13 10	LX 3 3
	5	LY 6 3	LY 11 3	LY 7 2	TE 10 3
	6	TE 9 4	LY 12 3	TD 5 3	LY 11 3
	7	TE 9 4	LY 11 3	LY 9 2	LY 13 1
	8	LY 9 2	TE 9 4	LY 12 3	LY 11 3
	9	TE 13 10	LY 9 2	LY 7 2	TE 9 5
	10	TE 9 4	LY 3 4	TE 13 10	TE 5 3
200	1	LY 9 2	LY 11 3	LY 12 3	BE 4 1
	2	LY 11 3	LY 12 3	BE 2 4	LY 7 2
	3	LY 9 2	LY 3 4	TE 13 10	TD 8 3
	4	LY 9 2	LX 3 3	LY 13 3	LY 3 4
	5	TE 8 4	LY 9 2	TE 13 10	TE 11 8
	6	LY 3 4	LY 9 2	LY 12 3	LX 3 3
	7	TE 9 4	LY 3 4	LX 3 3	LY 10 3
	8	LY 9 2	TE 11 10	LY 10 1	TE 12 7
	9	TE 9 4	LY 7 1	TE 13 11	LY 6 3
	10	TE 9 4	TE 13 10	LX 3 3	LY 3 4
250	1	TE 9 4	LY 3 4	LY 4 1	LX 3 3
	2	PS 3 1	LY 9 2	LY 11 3	TE 13 8
	3	LY 3 4	LY 9 2	LY 11 3	PS 3 1
	4	TE 13 10	LY 9 2	LY 3 4	LX 3 3
	5	LY 9 2	LY 3 4	LY 4 1	LY 11 3
	6	LY 9 2	LY 12 3	LY 11 3	LY 4 1
	7	TE 13 10	LY 9 2	LX 1 3	LY 3 4
	8	TE 13 10	LY 9 2	TE 6 5	LX 3 3
	9	LY 11 3	LY 9 2	BE 2 4	LY 12 3
	10	LY 9 2	TE 13 10	BE 1 4	LY 11 3
325	1	LY 9 2	LY 3 4	LY 11 3	LX 3 3
	2	LY 9 2	LY 11 3	BE 2 4	LY 8 4
	3	LY 3 4	TE 9 4	LY 11 3	LY 12 3
	4	LY 9 2	LX 3 3	LY 11 3	LY 3 4
	5	LY 9 2	LY 11 3	PS 3 1	LY 7 2
	6	LY 9 2	TE 8 4	TE 12 10	TE 11 6
	7	LY 9 2	LY 12 3	LY 11 3	LY 3 4
	8	LY 11 3	LY 9 2	LY 6 1	BE 2 4
	9	LX 3 3	LY 9 2	LY 11 3	LY 3 4
	10	LY 9 1	LY 3 4	LY 11 3	TE 9 4
400	1	LY 9 2	LY 11 3	LY 12 3	LY 7 2
	2	LY 9 2	TE 13 10	LY 3 4	LY 11 3
	3	LY 9 2	TE 13 10	LX 3 3	TE 9 4
	4	LY 11 3	LY 3 4	LY 9 2	LY 12 3
	5	LY 9 2	TE 13 10	LX 3 3	LY 7 2
	6	LY 11 3	LY 3 4	LY 9 2	LY 4 1
	7	LY 9 2	LY 3 4	TE 9 4	LY 11 3
	8	LY 11 3	LY 12 3	LY 9 2	BE 2 4
	9	LY 9 2	BE 2 4	TE 9 4	LY 10 3
	10	LY 9 2	LY 3 4	LY 11 3	LY 12 3

Table 3 (*continued*)

		Sequence of modifications			
n	Rep	1	2	3	4
800	1	LY 9 2	LY 3 4	LX 3 3	LY 11 3
	2	LY 9 2	LY 3 4	LY 11 3	LY 12 3
	3	LY 9 2	LX 3 3	LY 12 3	LY 11 3
	4	LY 9 2	LY 3 4	LY 11 3	LY 12 3
	5	LY 9 2	BE 1 3	TE 13 10	LX 3 3
	6	LY 9 2	LY 11 3	LY 12 3	LX 3 3
	7	LY 9 2	LY 11 3	LY 12 3	TE 9 4
	8	LY 9 2	LY 11 3	LY 12 3	LY 3 4
	9	LY 9 2	LY 3 4	LY 11 3	LY 12 3
	10	LY 9 2	LY 11 3	LY 12 3	LY 4 1
1,200	1	LY 9 2	LY 11 3	LY 12 3	LY 3 4
	2	LY 9 2	LY 3 4	LY 11 3	LY 12 3
	3	LY 9 2	LY 11 3	LY 12 3	LY 3 4
	4	LY 9 2	LY 11 3	LY 12 3	LY 3 4
	5	LY 9 2	LY 3 4	LY 11 3	LY 12 3
	6	TE 9 4	TE 13 10	LY 9 2	LY 3 4
	7	LY 9 2	TE 13 10	LY 3 4	LX 3 3
	8	LY 9 2	LY 11 3	LY 12 3	LY 3 4
	9	LY 9 2	TE 13 10	LY 3 4	LX 3 3
	10	LY 9 2	TE 13 10	LY 3 4	LY 11 3
3,694		LY 9 2	LY 3 4	LY 11 3	LY 12 3

Note. Parameter labels use matrix designations from Jöreskog & Sörbom, 1988: BE = Beta, LX = Lambda-X, LY = Lambda-Y, PS = Psi, TD = Theta-delta, TE = Theta-epsilon.

cally at low levels of *n*, going up and down across the sequence of modifications. This index behaved consistently only in samples of 800 or larger, where it improved steadily from step to step in each specification search.

Cudeck and Browne (1983) recommend using CV2 to determine which of several models yields the best cross-validity. Table 5 shows the frequency with which each model in the specification search yielded the lowest value of CV2 at each sample size. For instance, at $n = 100$, a model with one modification yielded the lowest value of CV2 in 4 of the 10 replications analyzed. Note, however, that those 4 cases do not necessarily result in the same model. As seen in Table 1, the first modification was highly inconsistent. Thus, Table 5 shows which step in the searches produced the lowest CV2, rather than which model did so. Once again, results show considerable inconsistency at low levels of *n*, becoming fairly consistent only for *n*s of 325 or larger. Complete consistency is seen at *n*s of 800 and 1,200, when the last model considered in each search produced the lowest value of CV2. This observation is, of course, redundant with the results in Table 4 showing CV2 to decrease at each step when *n* was 800 or 1,200.

Results in Tables 4 and 5 indicate that in terms of the two-sample cross-validation index, one cannot have confidence in the cross-validity of a modified model unless sample size is quite large. Furthermore, at small-to-moderate sample sizes, support for cross-validity of modified models is highly subject

Table 4
Frequency of Increases and Decreases of Two-Sample Cross-Validation Index Over Successive Modifications

	Modification							
	1		2		3		4	
n	I	D	I	D	I	D	I	D
100	4	6	9	1	5	5	7	2
150	3	7	5	5	4	6	6	4
200	2	8	2	8	2	8	4	6
250	2	8	0	10	4	6	0	10
325	1	9	2	8	1	9	1	9
400	0	10	2	8	0	10	1	9
800	0	10	0	10	0	10	0	10
1,200	0	10	0	10	0	10	0	10

Note. I = increase, D = decrease.

Table 5
Frequency for Each Model Having the Lowest Two-Sample Cross-Validation Index

	Model				
n	M0	M1	M2	M3	M4
100	3	4	0	1	2
150	1	1	1	3	4
200	1	0	1	3	5
250	0	0	1	0	9
325	0	1	0	1	8
400	0	1	0	1	8
800	0	0	0	0	10
1,200	0	0	0	0	10

to sampling fluctuations. Some samples may produce supportive results; others may not.

Recall that CV2 was also calculated for each model using the total sample covariance matrix, $\mathbf{S}_t$, as the cross-validation sample. Tables for this measure, analogous to Tables 4 and 5, were constructed but need not be presented. Results followed essentially the same pattern as seen in Tables 4 and 5, but with an important difference: In comparison with Table 4, results for cross-validation to the total sample tended to show a slightly higher frequency of improvement in CV2, especially for the first two modifications in each search. As a consequence, in comparison with Table 5, results for cross-validation to the total sample showed slightly higher frequencies for the more complex models producing the lowest values of CV2.

Results for the single-sample cross-validation index were so consistent that they need not be presented in detail. At each step in every specification search, at all sample sizes, CV1 decreased. Thus, in every search, the final model considered showed the lowest value of CV1. Results of Browne and Cudeck (1989) show that the model with the most parameters does not always produce the lowest value of this index. Considering the formula given in Equation 1, models with different numbers of parameters will have different values of both terms in the formula. In the special case of a specification search, as the number of parameters increases from step to step, the value of the fitting function will decrease, but the value of the second term in Equation 1 will increase. Because each step involves selection of the new parameter that will produce the maximum decrease in the fitting function, it would be expected that CV1 would routinely decrease from step to step in a sequential specification search. Only when the decrease in the fitting function becomes smaller than the increase in the second term of Equation 1 would this not be the case. In the present study, this never occurred. This is not at all surprising and simply demonstrates that CV1 is probably not appropriate for evaluating sequences of models produced by specification searches. This index is not used in the second sampling study.

Study 2

We conducted a second sampling study using a large data set collected by Verhoef and Roos (1970). A subset of these data was used by Cudeck and Browne (1983) and Browne and Cudeck (1989) to illustrate the use of cross-validation indexes. We used the same data, consisting of scores for 2,677 students on six mental ability tests, measured on three different occasions corresponding to age levels of approximately 14, 16, and 18 years. These data fit the framework of a multitrait–multimethod problem, in which occasions are analogous to methods. We attempted to use several different multitrait–multimethod confirmatory factor analysis models (Widaman, 1985). However, substantial difficulty was encountered in obtaining convergent and proper solutions for many of these models in the full sample or in subsamples. These problems have been cited often in the literature (e.g., Marsh, 1989; Wothke, 1984) and caused us to limit our sampling studies in terms of levels of n and number of replications. Nevertheless, we achieved limited success with two models, and we briefly describe our analyses and results.

Part A

The first initial model used was a seven-factor model, consisting of one general factor (with loadings on all 18 variables), and six trait factors, each with loadings on the three replications of a given test. The trait factors were correlated with each other, but not with the general factor. This model corresponds to Widaman's (1985) Model 2C. Following the design in Study 1, a small sampling study was conducted by drawing three pairs of samples of size 400 and three pairs of size 800. For each pair of samples, the initial model was fit to the calibration sample and then modified three times on the basis of the highest modification index at each step. We calculated goodness-of-fit and cross-validation indexes for the initial model and each modified model.

Results were generally similar to those obtained in Study 1, with one important difference. In every case, the fit of the initial model was quite good, for example, the mean value of NNFI across samples was approximately .95. The specific modifications carried out at each step were highly inconsistent at both levels of n, and did not correspond well with modifications made in analyses of the full sample. The two-sample cross-validation index, CV2, behaved erratically across model modifications at $n = 400$ but decreased fairly consistently at $n = 800$.

Part B

The second initial model used was a four-factor model, with one general factor and three method factors. Each method factor had loadings on the six tests at a given occasion. The method factors were correlated with each other but orthogonal to the general factor. This model corresponds to Widaman's (1985) Model 3B. Another small-scale sampling study, following the same procedure, was conducted. Sample sizes were 200 and 400. Three pairs of samples of each size were drawn, and three modifications were conducted in each specification search.

Results of these analyses were more encouraging with regard to stability and cross-validity. The fit of the initial model was rather poor (mean NNFI approximately .84) but improved substantially by means of the specification search (mean NNFI of final model approximately .91). In one sense, the particular parameters freed at each step in these searches were quite consistent across replications. Although the particular parameters and their sequence varied across replications, the parameters freed almost always came from a well-defined subset. In almost every case at both levels of n, as well as in the full sample, the parameter with the highest modification index at each step represented a covariance between error terms for the same test measured at two different times. Clearly, the four-factor model was not adequate to account for relationships between the same tests measured at different times. This systematic misspecification was revealed in the modification indexes, resulting in some consistency in the nature of the model modifications, if not their sequence. In addition, cross-validity of modified mod-

els was generally better, with CV2 decreasing through the sequence of modifications in almost every case, except when a particular modification did not fit the subset just defined.

Discussion

An evaluation of our results must be approached with the purpose of our study in mind. We are attempting to shed light on some phenomena that can occur when model modifications are conducted in practice. We will not argue that all of these phenomena must occur in any given application. However, the fact that a number of them can be seen quite clearly in our results provides strong evidence and reason for concern about what can happen in practice.

From this perspective, our results clearly reveal several important phenomena. The first is that the specific modifications carried out in a sequential specification search can be highly unstable in small-to-moderately-large samples. They may not be completely stable even when sample size is very large. This phenomenon was revealed in Study 1 and in Part A of Study 2. These results imply that when a sequential specification search is conducted in practice using data from a single sample, researchers cannot have great confidence that the specific model modifications would generalize beyond that sample. Unless sample size is very large, modifications may be quite idiosyncratic to that particular sample. Analyses of other samples of the same size may produce a quite different sequence of modifications. Furthermore, our results show that model modifications that are based on sample data may not correspond well to modifications that would be made in the population, as approximated here by the very large full sample in each study.

A second general finding, which is based on Study 1 and Part A of Study 2, is that modified models may not cross-validate consistently well unless sample size is quite large. As measured by the two-sample cross-validation index, cross-validity of modified models was found to be quite unstable at small-to-moderate levels of sample size. This finding implies that researchers can have little confidence in cross-validity of models produced by sequential specification searches in practice, unless sample size is very large.

Another important result from Study 1 and Part A of Study 2 is that cross-validation results themselves can be quite unstable across repeated samples, especially in small-to-medium-sized samples. Thus, even if a researcher conducts a two-sample cross-validation in practice and finds supportive evidence for a modified model, the same procedure repeated on a new pair of samples may produce a very different outcome. That is, cross-validity may not be supported in a different pair of samples.

Recall that our findings from Study 1 showed that two-sample cross-validation results were slightly more encouraging when CV2 was calculated using the total-sample covariance matrix rather than the covariance matrix from a completely independent sample of the same size. Using this approach, cross-validity seemed to improve more consistently for the early modifications in the searches. Although this may seem to indicate that validity of modified models in the population may be better than indicated by the use of CV2 for independent samples, we caution against being overly encouraged. First, results for cross-validation to the total sample were only slightly better than those for cross-validation to independent samples, shown in Tables 4 and 5. Second, this improvement could be an artifact to some extent, because the total sample obviously includes each subsample in which model fitting and modification were conducted. Given these factors, we believe results from cross-validation to independent samples should be given more weight.

Results from Study 1 also revealed an important phenomenon related to goodness of fit. Recall that goodness-of-fit measures were rather unstable for both initial and modified models across repeated samples. Thus, a researcher who modifies a model in practice to improve its fit may be making decisions that are highly influenced by sampling fluctuations. Fitting and modifying an initial model in another sample may result in substantially more or fewer, and perhaps quite different, model modifications to achieve adequate fit.

Focusing next on results of Study 2, there is evidence of two additional phenomena of interest. Results of Part A, where the seven-factor model was used as the initial model, showed very good fit of the initial model, followed by highly unstable modifications. This finding suggests the logical conclusion that when an initial model fits well, it is probably unwise to modify it to achieve even better fit because the modifications may simply be fitting small idiosyncratic characteristics of the sample. Of course, in such circumstances a specification search is generally unnecessary, though it may still be tempting for some researchers to take advantage of opportunities to improve fit that is already good.

Results of Part B of Study 2 provided the only encouraging support for stability and cross-validity of model modifications. In that study we found fairly consistent and interpretable model modifications in relatively small samples, with good cross-validation indexes. This finding indicates that the negative phenomena observed in other parts of our study do not necessarily occur in every case. Furthermore, it can be seen that these more encouraging results were found when the initial model was systematically misspecified and inadequate to explain the data. Thus, there are circumstances under which a specification search conducted in practice can yield a modified model that would be stable over repeated samples and would cross-validate well to an independent sample and to the population. However, in practice it may be quite difficult to determine whether these circumstances exist in any given case. It is tempting to draw the conclusion that these circumstances are supported by highly interpretable model modifications. However, we have already discussed concerns about how readily a researcher might find an interpretation for a modification that offered substantial improvement in fit. Thus, although results of Part B of Study 2 offer some encouraging signs, we do not see them as providing support for stability and cross-validity of model modifications in most empirical studies.

The findings reported here call into question the process and outcomes of specification searches in CSM, except in those few cases in which samples are extremely large. Consider the following scenario: Suppose a researcher began by fitting an initial model to a calibration sample, found that it fit at a mediocre level, modified it so as to improve its fit to an adequate level,

then cross-validated the modified model on a second sample and found the modified model to produce a better value of CV2 than the initial model. Our findings would make us quite skeptical of the support for the final model on several grounds. First, the instability of fit measures (Tables 1 and 2) means that the fit of the initial and final models in the calibration sample might be substantially different in another sample, meaning in turn that the specification search might be much shorter or longer. Second, the instability of particular modifications across samples (Table 3 and Part A of Study 2) means that a specification search in a different sample may well have produced a very different sequence of modifications. Third, the instability of cross-validation results (Tables 4 and 5 and Part A of Study 2) means that supportive findings with regard to cross-validity might well be another chance characteristic of the sample data and that quite different results might occur if different calibration and cross-validation samples were analyzed.

We believe these concerns are relevant to the vast majority of studies in which model modification is carried out. As discussed in the introduction, most such applications focus on improving the fit of an initial model by conducting a mechanical specification search. Such applications usually are not based on data from large samples and seldom provide information about cross-validity. An important implication of our findings, though, is that we would generally not be persuaded by supportive two-sample cross-validity evidence anyway (unless sample size is very large), because model modifications and cross-validity results are themselves quite unstable across repeated sampling.

Although we are quite confident in the legitimacy of these concerns, we recognize at least two aspects of the current study that merit further attention. The first is the generalizability of our findings, in that they are based on analyses of samples drawn from only two large data sets. Could these data be atypical in some sense? Would support for cross-validity of modified models be more favorable in other data sets? Although this scenario is possible, we consider it rather implausible. We have found nothing unusual about the data or the initial model used here. Furthermore, we reemphasize the point that our objective has been to reveal phenomena that can occur quite readily in empirical applications of specification searches. From that perspective, our results should cause serious concern. We suggest that the burden of proof is on those who may wish to prove us wrong, that is, to show that model modifications and cross-validity results are normally quite stable over repeated samples. We would be most interested to see sampling studies on large sets of empirical data showing such results.

A second limitation of the present study involves the manner in which model modifications were conducted. We used a mechanical process whereby an initial model was modified sequentially to achieve optimal improvement in overall fit. As emphasized above, we recognize that this approach violates the routine recommendation that modifications be substantively justified. However, our study was designed to mimic the manner in which models are often modified in practice. Nevertheless, a legitimate question involves whether stability and cross-validity of modified models would improve if an alternative approach to model modification were used. We suggest again that the burden of proof is on those who wish to show that highly generalizable modified models may be obtained using a different strategy. As stated earlier, any strategy that is data driven is inherently susceptible to problems arising from capitalization on chance. Therefore, we expect that most procedures for model modification would exhibit problems of the kind we have discussed and demonstrated.

A final issue raised by our findings concerns the question of how a researcher should evaluate cross-validity of modified covariance structure models. The single-sample cross-validation index, CV1, should not be used for this purpose. Browne and Cudeck (1989) proposed that measure for the comparison of alternative a priori models and demonstrated its usefulness. However, CV1 is not useful for evaluating cross-validity of a sequence of models produced by a specification search, because it will virtually always show the most complex model to have the best cross-validity. That is, CV1 is not designed to be sensitive to capitalization on chance in the specification search process. With regard to two-sample cross-validation, we are concerned by the unstable results found with regard to the behavior of CV2 in most of the conditions we examined. This index behaved erratically when sample size was not large, which is when cross-validation is most needed.

Rather than use one of these indexes to assess cross-validity of modified models, we recommend a *parallel specification search* procedure. This procedure would involve conducting the specification search process on independent samples and obtaining goodness-of-fit measures for each model in each sample. One would also obtain two sets of two-sample cross-validation measures by carrying out a double cross-validation analysis, exchanging the designation of calibration and cross-validation samples. We believe this approach would provide the most relevant information by allowing the investigator to assess the consistency of the model modifications as well as goodness-of-fit and two-sample cross-validation measures in each sample. Inconsistent results would cast severe doubt on the validity of the final model(s) obtained; consistent results would provide strong support. On the basis of our findings, we expect that such support would be obtained primarily when sample size is very large or the initial model is systematically misspecified.

In summary, our results bring us to a position of considerable skepticism with regard to the validity of the model modification process as it is often used in practice. We believe models produced by mechanical specification searches in samples that are not extremely large are likely to be highly influenced by chance characteristics of the sample. For researchers who proceed with model modifications in single samples in the face of the problems and concerns we have raised, it is essential that they take a very conservative approach, that is, make few modifications and require clear interpretability. This requirement might be relaxed slightly in studies of a more exploratory nature. In any case, however, researchers must clearly acknowledge the questionable validity of a modified model by stating that the initial model has been modified to improve its fit to a single sample and that the resulting modified model may not generalize to other samples or to the population. Such models

must be evaluated in subsequent studies using independent samples.

Finally, given the concerns we have raised regarding specification searches in CSM, we encourage researchers to consider an alternative strategy for the development and evaluation of covariance structure models: the use of multiple a priori models. In confirmatory studies there may be conflicting theoretical positions or diverse research findings that call for the construction of competing models. In exploratory studies alternative models may be constructed as a result of mere uncertainty about the pattern of relationships among variables or a variety of ideas about the nature of that pattern of relationships. Alternative models could then be fit to sample data and evaluated in terms of overall fit, cross-validity, and interpretability of results. This approach is encouraged by Cudeck and Browne (1983) and Browne and Cudeck (1989) in their development of measures of cross-validity of covariance structure models. We strongly endorse this approach and consider it a more defensible method of model development than the construction of a single initial model followed by a number of data-driven modifications of that model. Because the construction of alternative a priori models is not driven by the data at hand, the process of model comparison and selection should be somewhat less influenced by chance characteristics of the data. Furthermore, the serious problem of interpretation of model modifications would be circumvented by this approach.

References

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 19,* 453–460.

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

Balderjahn, I. (1988). A note to Bollen's alternative fit measure. *Psychometrika, 53,* 283–285.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419–456.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Software.

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

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

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229–242.

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

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in multivariate analysis* (pp. 72–141). Cambridge, England: Cambridge University Press.

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

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,* 147–167.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

Farkas, A. J., & Tetrick, L. E. (1989). A three-wave longitudinal analysis of the causal ordering of satisfaction and commitment on turnover decisions. *Journal of Applied Psychology, 74,* 855–868.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore: Johns Hopkins University Press.

Hinkin, T. R., & Schriesheim, C. A. (1989). Development and application of new scales to measure the French and Raven (1959) bases of social power. *Journal of Applied Psychology, 74,* 561–567.

Hulin, C. L. (1991). Adaptation, persistence, and commitment in organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 445–505). Palo Alto, CA: Consulting Psychologists Press.

Hulin, C. L., Roznowski, M., & Hachiya, D. (1985). Alternative opportunities and withdrawal decisions: Empirical and theoretical discrepancies and an integration. *Psychological Bulletin, 97,* 233–250.

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

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

Kulik, C. T., Oldham, G. R., & Langer, P. H. (1988). Measurement of job characteristics: Comparison of the original and the revised job diagnostic survey. *Journal of Applied Psychology, 73,* 462–466.

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

Lee, S. Y., & Bentler, P. M. (1980). Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal, 41,* 121–136.

Long, J. S. (1983). *Covariance structure models: An introduction to LISREL.* Beverly Hills, CA: Sage.

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

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

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.

Meyer, J. P., & Gellatly, I. R. (1988). Perceived performance norm as a mediator in the effect of assigned goal on personal goal and task performance. *Journal of Applied Psychology, 73,* 410–420.

Newcomb, M. D., Huba, G. J., & Bentler, P. M. (1986). Determinants of sexual and dating behaviors among adolescents. *Journal of Personality and Social Psychology, 50,* 56–66.

Reisenzein, R. (1986). A structural equation analysis of Weiner's attribution–affect model of helping behavior. *Journal of Personality and Social Psychology, 50,* 1123–1133.

Rosse, J. G. (1983). *Employee withdrawal and adaptation: An expanded framework.* Unpublished doctoral dissertation, University of Illinois.

Rosse, J. G., & Miller, H. E. (1984). Relationship between absenteeism and other withdrawal behaviors. In P. S. Goodman & R. S. Atkin (Eds.), *Absenteeism: New approaches to understanding, measuring, and managing employee absence.* San Francisco: Jossey-Bass.

Roznowski, M. (1989). Examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology, 74,* 805–814.

Roznowski, M., & Hulin, C. L. (1991). The scientific merit of valid measures of important constructs with special reference to job satisfaction and job withdrawal. In C. J. Cranny, P. C. Smith, & E. F. Stone (Eds.), *Job satisfaction: Advances in theory and research.* New York: Free Press.

Saris, W. E., & Stronkhorst, H. (1984). *Causal modeling in nonexperimental research.* Amsterdam: Sociometric Research Foundation.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54,* 131–151.

Schriesheim, C. A., & Hinkin, T. R. (1990). Influence tactics used by subordinates: A theoretical and empirical analysis and refinement of the Kipnis, Schmidt, and Wilkinson subscales. *Journal of Applied Psychology, 75,* 246–257.

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

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68,* 653–663.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand-McNally.

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

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

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

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology, 46,* 621–635.

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

Verhoef, W., & Roos, W. L. (1970). *The aim and experimental design of Project Talent Survey* (Report No. MT1). Pretoria, Transvaal, Republic of South Africa: Human Sciences Research Council.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Wothke, W. (1984). *The estimation of trait and method components in multitrait–multimethod measurement.* Unpublished doctoral dissertation, University of Chicago, Chicago.

Received September 30, 1990
Revision received August 12, 1991
Accepted August 16, 1991 ■