JOURNAL OF ORGANIZATIONAL BEHAVIOR, VOL. 18, 667–683 (1997)

Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives

AMY E. HURLEY[1], TERRI A. SCANDURA[2], CHESTER A. SCHRIESHEIM[2], MICHAEL T. BRANNICK[3], ANSON SEERS[4], ROBERT J. VANDENBERG[5] AND LARRY J. WILLIAMS[6]

[1]*Department of Professional Studies, Chapman University, U.S.A.*
[2]*Department of Management, University of Miami, U.S.A.*
[3]*Department of Psychology, University of South Florida, U.S.A.*
[4]*Department of Management, Virginia Commonwealth University, U.S.A.*
[5]*Department of Management, The University of Georgia, U.S.A.*
[6]*Department of Management, University of Tennessee, U.S.A.*

J. Organiz. Behav. **18**: 667–683 (1997)
No. of Figures: 0 No. of Tables: 3 No. of References: 28

Introduction

> *'Most uses of "confirmatory" factor analyses are, in actuality, partly exploratory and partly confirmatory in that the resultant model is derived in part from theory and in part from a respecification based on the analysis of model fit.'*
>
> (Gerbing and Hamilton, 1996, p. 71)

The above quote illustrates the debate over the use of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) in organizational research. Recent articles appearing in the major organizational research journals (Brannick, 1995; Stone-Romero, Weaver and Glenar, 1995) concluded that the use of CFA is steadily increasing while the use of EFA is declining. The merits of exploratory and confirmatory factor analysis have long been debated and have resulted in some extremely energetic exchanges on both the research methods and the structural equation modeling networks. Further, the *Journal of Organizational Behavior* recently highlighted this debate with a three piece exchange on covariance structure modeling in May of 1995 (Brannick, 1995; Kelloway, 1995; Williams, 1995). These exchanges have been useful to researchers in deciding which type of factor analysis to use.

In general, proponents of CFA believe that researchers need to have a strong theory underlying their measurement model before analyzing data (Williams, 1995). CFA is often used in data analysis to examine the expected causal connections between variables. Supporters of EFA believe that CFA is overapplied and used in inappropriate situations. Despite the rhetoric to the

Addressee for correspondence: Amy E. Hurley, Department of Professional Studies, Chapman University, Orange, CA 92866. We would like to thank Mark Gavin for his question on using EFA and CFA on the same data sets.

CCC 0894–3796/97/060667–17$17.50

Accepted 18 February 1997

contrary, some researchers believe that CFA is still being used with little theoretical foundation, and that reviewers may be requiring CFA where a simpler alternative would be as or more appropriate (Brannick, 1995). EFA is often considered to be more appropriate than CFA in the early stages of scale development because CFA does not show how well your items load on the nonhypothesized factors (Kelloway, 1995).

Others believe each method is appropriate in different situations. EFA may be appropriate for scale development while CFA would be preferred where measurement models have a well-developed underlying theory for hypothesized patterns of loadings. A line of research would start out with studies utilizing EFA while later work would show what can be confirmed. A recent study by Gerbing and Hamilton (1996) using Monte Carlo methods found that EFA can contribute to model specification when used prior to cross-validation using CFA.

Because of the continued interest in the topic, we assembled a panel of five experts at the 1996 SIOP annual meeting in San Diego California to discuss the underlying issues and provide guidance for researchers interested in utilizing factor analytic procedures. The discussion underscored the need for clarification in the use of EFA and CFA in organizational research. The format of this paper generally follows the panel discussion. Questions were posed by the session facilitators and the panelists responded. In addition, the authors were asked to comment on guidelines for scale development and goodness-of-fit indices (GFIs). For this paper, the authors also provided a concluding statement summarizing their position. Hence, our purposes were to provide a general overview of the valuable discussion that took place during the session on exploratory–confirmatory procedures and to go beyond this discussion to provide guidelines for practice.

Debate

Amy Hurley and Terri Scandura: Let's begin with the topic of the current debate: EFA versus CFA. What are your views on the issue? Do you think the debate is a fruitful discourse or just making hay?

Mike Brannick: Different people can mean very different things when they discuss CFA and EFA. Factor analytic techniques currently in use vary in such features as the loss function used for parameter estimation (e.g. maximum likelihood, least squares), the type of rotation (e.g. VARIMAX, Procrustes), the type of test for the number of factors (e.g. parallel analysis, chi-square), degree of restriction of parameter estimates (fixed, freed, or constrained within boundaries), and what sources of error are thought to be most important (i.e., error due to the sampling of people or the sampling of variables).

To me, CFA is best defined as a decision rule to accept or reject one or more hypotheses about a population factor structure based on sample data. That is, CFA is about hypothesis testing. Unfortunately, techniques commonly employed in CFA do not always correspond to the hypotheses we have in mind when we employ them. The strongest possible use of CFA would be to hypothesize a set of parameters (factor loadings, correlations, and uniquenesses) and test the fit of a reproduced matrix to sample data without estimating any parameters based on sample data (This is also the strongest form of cross validation). I don't recall ever seeing this done, but it would be the quintessential application of CFA. Instead, common practice is to estimate a set of factor loadings, one for each observed variable so that simple structure is achieved. In addition, it is common practice to set all other factor loadings to zero. Note that what is tested in such a case

is essentially whether lots of observed variables have zero loadings on lots of factors in the population. With any real data, the probability of finding such a structure in the population is about zero, and of course with large samples the null hypothesis is proved false. The significance test does not correspond too well to the hypothesis we have in mind, however, which is something like 'Are there K important factors in our data?'

This question is not well addressed by a traditional hypothesis testing approach. It is analogous to the question of statistical significance versus magnitude of effect where we can have significant but trivial effects. We would like to know whether some small number of factors will account for the meaningful variance in the correlations of interest. What we actually test for, however, is whether there are other nonzero factors in the data, and of course there nearly always are in real data.

Another concern I have is that researchers automatically turn to CFA without considering the merits of EFA. For example, current reviewers for major journals seem to routinely require CFA analyses of MTMM matrices. Although CFA has been used to analyze the MTMM matrix, there are well known estimation problems when fitting data to such models (e.g. Brannick and Spector, 1990; Kenny and Kashy, 1992) and the analysis rarely provides meaningful parameter estimates. Reviewers seem to assume that CFA is better than a variety of older methods, and that the older models really do not tell us much. But in some cases, the methods traditionally employed for EFA provide a better basis for examining what we really want to know.

Chet Schriesheim: The debate over CFA versus EFA is like arguing about the 'best' flavor of ice cream: you pay the price and you take your choice. Arguments such as these often remind me of some I took part in as a child, growing up in a family where we each had decided preferences for different flavors of ice cream. However, I believe that this discussion is generally healthy for organizational research, provided that each of the various advocates and factions are listening to and hearing what the other is trying to say. Virtually all data-analytic methods have both strengths and corresponding weaknesses and these vary considerably by the particular purpose and application involved (cf. Wherry, 1975). Both EFA and CFA are useful and both have some serious requirements if they are to be performed and used reasonably well. I think this needs to be kept in mind by all parties in the CFA–EFA debate.

Anson Seers: Comparison and contrast of the merits of CFA in relation to those of EFA implies that researchers should be expected to choose one form of analysis over the other. In a sense, researchers must make such a choice within the context of any single research study. Given the well-documented differences between the techniques and their ideal applications, only one, or the other, should truly be appropriate to a given combination of data and research questions.

Bob Vandenberg: Prior to being asked to participate on the panel, I can honestly admit that I hadn't put much thought into my position on the topic. The invitation to participate, however, changed that. Suddenly, I was being asked to take a position on something that I took for granted in my own work. My first reaction, therefore, was one of believing that I didn't really have a position. The more I thought about it, though, the more it became apparent to me the strong feelings I did possess. It's just that those feelings were anchored to a perspective that I am going to call 'traditional'. In part, this traditionalist perspective means that I take a middle-of-the-ground approach. For example, when addressing whether this discussion on EFA and CFA is a 'fruitful discourse or just making hay', I would say that it is both. It is fruitful discourse from two perspectives. First, it is quite useful to remind ourselves periodically of what constitutes appropriate or inappropriate practices. Second, it is fruitful to point out the folly of anchoring

oneself to one side or the other. From this same perspective, however, I would say that we are also 'just making hay', because to me the techniques are complementary. In other words, it is cut and dry as to the appropriate application of the techniques, and as such, discussing them equates to 'much ado about nothing'.

Larry Williams: I think it is inevitable that as new statistical techniques come to be adopted by organizational researchers, their relative merits in comparison to traditional tools will be considered. Examples can be found with various types of latent variable structural equation models (LVSEM) which go beyond the simpler CFA models discussed presently. For instance, the advantages of LVSEM over partial correlation and multiple regression approaches to model testing have been discussed (e.g. Williams, Gavin and Williams, 1996), as have the strengths of LVSEM over alternative approaches to the analysis of reciprocal relationships with longitudinal field data (e.g. Williams and Podsakoff, 1989). The same is true for the analysis of multitrait–multimethod matrices (e.g. Williams, Cote and Buckley, 1989). I feel that proponents of new statistical methodologies are obligated to consider such comparisons, as more is required before one can advocate the use of a new tool than to simply say that the tool is new. I believe that the comparison of different methodologies should focus (at least in part) on the different conditions under which the techniques should be used, the types of output or diagnostics provided, and the degree to which conclusions about theoretical questions are likely to be different. Consideration of these issues with respect to EFA versus CFA will lead to better and more informed use of these two techniques by organizational researchers.

AH and TS: All of you mentioned that CFA and EFA have their own strengths and restrictions. Would you discuss some of the important differences researchers should keep in mind while deciding which technique to use?

MB: I would like to present a small simulation to illustrate some important differences and similarities of EFA and CFA. It begins by assuming a very clean population factor structure in which there are three factors and 10 variables per factor for a total of 30 variables. Each observed variable has a loading of 0.55 or more on the appropriate factor and a loading of no more than 0.20 on any other factors. The factors are either orthogonal or show small correlations. Two common techniques were applied to a sample of 500 observations from this simulated data. The first is the SAS factor analysis employing principal axes with prior communality estimates computed by squared multiple correlations. The initial factor pattern was rotated by the SAS program PROMAX, which attempts to maximize simple structure while allowing the factors to become correlated. The second technique was to use LISREL as is commonly employed in CFA. Each variable was hypothesized to load on a single factor; the other loadings were fixed at zero. Factor correlations and uniqueness were also estimated. Maximum likelihood was the loss function employed.

Population factor loadings and their estimates are shown in Table 1 (the PROMAX factor loadings are standardized regression coefficients) and Table 2 shows population factor correlations and their estimates. Some points to notice about the results of the simulation are that (1) none of the sets of estimates of factor loadings or factor correlations equal the parameters (and this would be true even if the sample size were infinite), (2) all of the estimates of the factor correlations are too high (biased), and (3) the results are very similar across EFA and CFA estimates.

Another concern I have is that my understanding of maximum likelihood is that statisticians prefer it because it is efficient and consistent. As the sample size increases, maximum likelihood

Table 1. Population factor pattern compared to EFA and CFA estimates

Variable	Population parameters			PROMAX			LISREL		
1	0.70	0.12	0.04	0.71	0.06	−0.07	0.70	0.00	0.00
2	0.65	0.15	0.20	0.65	0.07	0.05	0.72	0.00	0.00
3	0.55	0.02	0.15	0.50	−0.03	0.09	0.55	0.00	0.00
4	0.70	0.10	0.10	0.71	−0.00	0.02	0.72	0.00	0.00
5	0.75	0.03	0.12	0.72	−0.09	0.07	0.73	0.00	0.00
6	0.60	0.12	0.13	0.67	−0.01	−0.02	0.64	0.00	0.00
7	0.55	0.17	0.02	0.54	0.09	−0.07	0.53	0.00	0.00
8	0.58	0.10	0.20	0.56	0.04	0.12	0.65	0.00	0.00
9	0.62	0.03	0.18	0.58	−0.07	0.15	0.64	0.00	0.00
10	0.65	0.12	0.01	0.62	0.01	−0.03	0.60	0.00	0.00
11	0.04	0.65	0.20	−0.04	0.63	0.15	0.00	0.67	0.00
12	0.08	0.70	0.04	−0.04	0.75	−0.10	0.00	0.67	0.00
13	0.04	0.64	0.13	−0.08	0.61	0.08	0.00	0.60	0.00
14	0.20	0.60	0.13	0.09	0.55	0.12	0.00	0.64	0.00
15	0.15	0.70	0.02	0.13	0.69	−0.09	0.00	0.72	0.00
16	0.12	0.72	0.10	0.06	0.71	−0.01	0.00	0.74	0.00
17	0.04	0.60	0.03	−0.02	0.58	−0.04	0.00	0.55	0.00
18	0.02	0.70	0.01	−0.08	0.71	−0.10	0.00	0.62	0.00
19	0.20	0.68	0.14	0.10	0.73	0.03	0.00	0.81	0.00
20	0.14	0.58	0.12	0.04	0.59	0.12	0.00	0.66	0.00
21	0.15	0.10	0.65	0.13	0.01	0.65	0.00	0.00	0.75
22	0.08	0.04	0.70	0.05	−0.08	0.69	0.00	0.00	0.68
23	0.10	0.20	0.55	0.06	0.07	0.56	0.00	0.00	0.62
24	0.04	0.03	0.58	−0.07	−0.00	0.63	0.00	0.00	0.58
25	0.10	0.10	0.70	0.01	−0.03	0.69	0.00	0.00	0.68
26	0.12	0.17	0.65	0.05	0.12	0.62	0.00	0.00	0.70
27	0.02	0.09	0.66	−0.01	−0.03	0.67	0.00	0.00	0.64
28	0.04	0.20	0.59	−0.05	0.11	0.64	0.00	0.00	0.65
29	0.18	0.03	0.60	0.05	−0.03	0.65	0.00	0.00	0.66
30	0.09	0.07	0.62	0.04	−0.03	0.63	0.00	0.00	0.64

Table 2. Population factor correlations and sample estimates

Population factor correlations				PROMAX factor correlations			
	1	2	3		1	2	3
1	1			1	1		
2	0.20	1		2	0.43	1	
3	0.30	0.00	1	3	0.56	0.33	1
LISREL estimates				Estimates based on observed correlations of unit weighted composites			
	1	2	3		1	2	3
1	1			1	1		
2	0.48	1		2	0.42	1	
3	0.63	0.39	1	3	0.55	0.34	1

estimates converge on the population values relatively quickly and the bias in the estimate is asymptotically zero. But I understand this is only true when the hypothesized model is true in the population. If it is not true, the parameter estimates cannot be shown to converge on the correct values. Any misspecification tends to be spread throughout the model. We can be pretty certain

with real data that our model is misspecified going into the analysis, so that our parameter estimates will be wrong. I wonder how wrong they will be and under what conditions examining them will be misleading.

CS: I'd echo the above points and further emphasize that CFA requires *a priori* hypotheses or clear 'theory', while EFA is theoretically less demanding. Thus, one can always subject a data set to an EFA but not necessarily a CFA. However, needless to say, theory-based research is more compelling in many ways than is more exploratory work (cf. Kerlinger, 1986). CFA therefore helps ensure that the researcher considers relationships between data and theory and doesn't just collect data and 'grind it' though exploratory procedures. On the other hand, sometimes the theory being tested doesn't 'fit'. What does one do then? One approach is to use modification indices to help a misspecified CFA become a good representation of a data set. Technically, if one changes a few parameters based upon the data themselves, the CFA actually becomes an EFA. One can argue then that perhaps a more honest approach is simply to conduct a traditional EFA and report one's findings as such. This may not be pleasant but I believe that it's essential to be extremely accurate (and honest) in reporting EFA/CFA methods and results.

LW: There are some important differences between these two techniques. For example, EFA places great emphasis on eigenvalues as indicators of dimensionality, while the CFA emphasizes goodness-of-fit. Also, with EFA all factor loadings (for every item on every factor) are estimated, while with CFA nearly all factor loadings are constrained to zero (only one loading per item is typically estimated). Finally, with EFA the researcher has to treat all factors the same, in that they are all proposed to be either correlated or uncorrelated. With CFA, the researcher may allow some factors to be related and others to be unrelated.

AH and TS: There seem to be clear guidelines in your minds of the appropriate times when EFA and CFA should be used in research. Would you state what you feel these guidelines are?

MB: I would use CFA whenever I had a specific hypothesis to test. For example, if I had some theoretical reason to believe that the correlation between the Verbal and Numerical Factors measured by the SAT were different for males and females, I would use CFA to test the equality of the factor correlations. I would also use the technique for testing hypotheses about scale translations such as whether the factor loadings were equal for representative workers in the U.S.A. and Brazil.

I would use EFA for scale development and evaluation. There are reasons to avoid using EFA for scale development, such as difficulty factors and poor item distributions (e.g. Nunnally and Bernstein, 1994). However, these reasons apply to both EFA and CFA when used to model items as variables. Despite such problems, EFA helps develop scales that show good internal consistency while minimizing overlap with other scales.

There is nothing to stop one from using CFA in scale development to test whether the newly written items conform to the hypothesized structure the scale architect had in mind. However, it seems pretty much a waste of time to me because it is virtually certain that the CFA as commonly employed will not fit the data well. Common advice in psychometrics is to begin with 2–3 times as many items in the initial scale as are to be included in the final scale. We simply cannot write items that behave perfectly in a psychometric sense; using empirical data (item analysis) we need to choose the best items from a set of items that are equally face valid. Even after selection of items by EFA to maximize convergent and discriminant validity of items in scales, the CFA model is too restrictive to expect a good fit. That is, we cannot reasonably expect many loadings

to be zero in the population. The typical CFA model may symbolize what we mean by good measurement; however to expect such a model to hold with a real scale is simply asking too much of our fallible measures. Therefore, the hypothesized structure will always provide poor fit to the data if the sample size is large enough. If one does apply CFA to the data and subsequently finds poor fit, then I think that one should revert to EFA and that reviewers should not penalize the author for doing so.

AS: A classic distinction, between the context of confirmation, in which our main purpose is to test formal hypotheses, and the context of discovery, in which our main purpose is to generate hypotheses, is quite useful in relation to the issue of EFA versus CFA. Often, we seek to hasten the progress of our discipline by attempting to use a single data set to identify a new construct and to assess construct validity, when cross-validation is preferable. No single study ever stands alone. Disciplines make their most rapid progress when programmatic lines of research are pursued such that earlier work explores and later work shows what can be confirmed.

We should not assume that use of the term 'exploring' means we have no preconceived ideas about what we may find. The most fruitful explorations involve a great deal of forethought. Even the most well-designed exploration would be presumptive if couched in terms of confirmation. We do need to present a reasonable accumulation of evidence as a foundation for any analysis to be described as confirmatory.

CS: I think that it makes a lot of sense to be very cautious in proposing statistical/analytic 'guidelines' of any kind, since these tend to become reified and take on an air of greater precision than is often meant by their proposers. 'Guidelines' are also sometimes misused by the naive, who take refuge in following 'rules' rather than in developing the expertise which is necessary to exercise informed judgment. Having advanced this disclaimer, let me now offer some brief observations about when EFA and CFA might be best used. Of course, these comments assume that the data being employed are suitable for use of either approach (there are an adequate number of observations and manifest variables, distributional assumptions are met, etc.). Simply put, I'd use CFA when I had some basis for developing one or more *a priori* hypothesized structures for the data and EFA either when I had no such basis or when my *a priori* structures were not confirmed. The source of an *a priori* structure could be an existing or newly-developed theory, a literature review or meta-analysis, or previous empirical work (which did not employ the same data). Additionally, I'd like to add the observation that replication substantially enhances the scientific contribution of any factor-analytic work, so that researchers might want to consider using EFA and CFA in multi-sample studies, perhaps first exploring and then confirming one or more particular structural hypotheses.

BV: Addressing this begins with a reminder to us all that for nearly 100 years (Pearson, 1901; Spearman, 1904), the purpose of factor analysis has remained exactly the same. To quote from Bollen (1989, p. 206), 'Its goal is to explain the covariances or correlations between many observed variables by means of relatively few underlying latent variables'. Thus, the issue underlying a discussion of CFA and EFA is not one of differing goals or outcomes. We are all interested in identifying the common ground upon which our observations rest. In this case, the common ground is the observed variables before us that have some practical or research value to us.

We need to understand how the observations before us were generated. Were the observations the end product of some TRULY *a priori* process in which the conceptual basis was stated prior to data collection? Or are these observations from some source of unknown origin or at best, the result of some brainstorming session in which we made a wish about wanting to measure

something without articulating why? The reason I bring it down to such a simple level is as follows. As part of my traditional perspective, I have adopted the position expressed by experts for many, many years. Namely, confirmatory factor analysis is just that, confirmatory, a method for confirming theory in which a detailed and identified initial model has been specified before data collection. These sentiments with all of the underlying rationale have been succinctly expressed by Mulaik (1972, 1987), Gorsuch (1983), Bollen (1989), Velicer and Jackson (1990) and most recently, by Gerbing and Hamilton (1996). My point in citing these individuals is not to claim that they have a corner on 'the truth' or that they have provided 'the answer'. Namely, my point is that the fundamental difference between CFA and EFA hasn't changed in 25 years (over 30 years actually if one includes Joreskog's work in the 1960s at ETS). And my understanding of the situation is that this difference is not going to change for another 25 or 30 years. Thus, what does this mean for us as researchers? It means that we need to possess a thorough understanding of our data.

LW: As a result of my experiences over the past few years, I have come to believe that in many (if not most) data analytic situations involving measurement issues, EFA and CFA can provide complementary perspectives on one's data. This opinion is based on the assumption that in these instances, two important questions that researchers face are (a) the underlying dimensionality of data, and (b) the adequacy of individual items. For these two questions, the contributions of EFA should not be dismissed too quickly. For example, the use of eigenvalues as diagnostics in judgments about dimensionality has a rich tradition, and when properly used provides a more direct picture of dimensionality than goodness-of-fit measures used with CFA. Also, researchers often want to show that their items contain relatively little common variance other than that associated with a single, substantive factor. The use of CFA addresses this question only indirectly, in that the inappropriate restriction to zero of factor loadings (for factors other than the intended one) is reflected in relatively lower goodness-of-fit values. In these two instances, I feel that EFA provides important diagnostics which should be considered along with the results of CFA in judging a scale and its items.

AH and TS: Let's specifically discuss scale development/validation. What guidelines can you give us in that area?

CS: I'd like to disagree with one of Mike Brannick's earlier statements. Where I've seen scales carefully constructed, I mean *painstakingly* with respect to the issue of content validity (cf. Schriesheim, Powers, Scandura, Gardiner and Lankau, 1993), CFAs have generally worked out fine. I've rarely seen a scale that was well-constructed which didn't produce pretty good CFA results on a fairly consistent basis. Obviously, you may get a quirky sample. But, if you do the preliminary work well, using CFA should pose few problems. Parenthetically, if you're interested in doing work in an area where the scales were developed many years ago, it might make sense for you to develop your own measures—to go back to square one and do a careful job rather than relying on someone else's work (in part, because standards have changed). Thus, if you're doing serious work in an area it's probably worth the investment. Of course there are problems with doing this. The field doesn't 'pay' for that type of work (in terms of reputation, publication in top-tier outlets, etc.). It's very hard to get scale development work published and it's often forced to remain part of an unpublished technical report.

BV: Assuming that the primary question is, 'which procedure should a researcher start out with to analyze data?', we need to know does the researcher have control over the design of the survey

and its administration (researcher–control), or did the researcher inherit some organization's database representing an in-house developed survey in which items are anchored to known conceptual bases (researcher–inherit). With respect to the 'researcher–inherit' situation, there is, in my opinion, little recourse other than the use of EFA procedures. In most 'researcher–inherit' situations, the items have no known conceptual basis, and thus, any logical grouping of items by the researcher even before undertaking the analyses is still a *post-hoc* exercise. If the observations are the end product of some truly *a priori* process, then CFA is warranted.

AH and TS: What about using both techniques on the same data set? For example, you *a priori* develop a measure that taps three dimensions and have hypotheses about which items load where. We assess that in a sample where you get supervisor and subordinate responses to the same items and we're interested in not only whether the *a priori* factor structure holds, but if that measurement model holds across both samples. Since we're developing new measures we might need to use exploratory since it is a first test with data on that measurement instrument. We might also be able to use confirmatory because we've got these *a priori* hypothesized patterns. The two give you different pieces of information. For example, in exploratory you can see the actual magnitudes of the cross loadings but you can't pick that up in CFA. Could you do an exploratory, and assuming that your data structure pretty much fits in both samples, then resubmit it to a confirmatory to get further item diagnostics. Where confirmatory does a little bit better than the exploratory is it allows you to test with the multiple groups the invariance of those measurement models and how they compare.

AS: Obviously the situation is an attempt to say 'if I can get a really big sample and hold half of that out, can I use that for cross validation'. I think that's a step in the right direction but we have to be concerned about capitalizing on chance. All capitalizing on chance doesn't involve just the particular chance relations that are in a particular subset of data that you use as a sample. Subsets from an original, common data set should be quite likely to share common method variance and I'd be a little more inclined to think that different data sets collected over a period of time have less probability of chance variation occurring there to be picked up on in the analysis. We're talking about one study collected this year and another 2 years from now, and so on. So I'd be more optimistic for checking for similarity of results of additional samples if they weren't collected simultaneously by the same researchers with the same methods. The other issue is the question of generalizability, i.e. external validity. People often misinterpret this as comparing results from one sample to those from another sample, but that's not what generalizability addresses. Generalizability applies results from a sample to a population. Without careful consideration of just how our samples represent a population, we don't adequately address the question of generalizability. So I'm skeptical.

MB: This issue is really one of cross validation. It turns out that as validation and cross validation samples grow large, they converge upon a common covariance matrix. Because of this convergence, cross validation becomes certain as the sample size increases. Such a cross validation effort is a check against parameter fluctuations due to sampling error. This type of cross validation is a check on the magnitude of the parameters estimated and not upon the choice of parameters to estimate. This means that you could have two different models generated from the same data that would cross validate equally well. The bottom line, therefore, is that such a cross validation fails to ensure that you have a good representation of the factor structure. A stronger cross validation in the sense of understanding what the factors really mean in some population would be achieved by using new variables (items or measures) to measure the same constructs in a

new sample. Finding a similar structure under such a circumstance would be strong evidence in support of the model.

BV: On the one hand this is the philosophical approach I would take, if you're a good researcher you're not going to do something that's inappropriate. Statistically if your sample size is large enough and your scales are at least different enough that you're going to get some discriminant validity between them, there's basically no difference in the statistical outcomes of EFA and CFA. In that case, I would say why not look at these things simultaneously and thus, use CFA. You had a theory that generated the items in the first place. You could basically accomplish both goals. For the situation described, you would want to test a series of nested models using the multi-sample analysis feature. The first model in that series addresses the question, 'is the latent factor structure underlying the set of items the same in the samples?' *If it isn't*, then evidence exists that the samples were using different conceptual frames of reference when interpreting and responding to the items, and thus, the samples cannot be directly compared. At this point, the less restrictive EFA procedures may be employed to determine where the 'breakdowns' occurred, particularly if the expectation was that the samples would interpret the items using identical frames of reference. *If the first model is supported*, that is the latent factor structure underlying the items is the same across samples, then other models in the series can be tested. For example, the next model may ask the question, 'are the factor loadings equivalent across samples?' By specifying invariant factor loadings in this model, one is directly testing whether the samples are calibrating the items in a similar fashion. Namely, does responding with a '3', for example, to an item mean the same thing to supervisors as it does to subordinates? Or does the 3 (on a 5-point Likert scale) mean different things (i.e. for the subordinate it truly reflects a neutral position, but for the supervisor it reflects a moderately positive position)? If this model results in a dramatic loss in fit, then an exploratory approach may be employed to uncover which item or items result in the greatest differences. Continuing in the series of tests, the next model may restrict item intercepts to be invariant across samples. Doing so addresses whether one group had different response tendencies (i.e. biases) across the items than the other groups (Bollen, 1989). Do supervisors, for example, characteristically respond across the items with lower values than do subordinates? Again, if this model results in a dramatic loss in fit, one can resort to some exploration of the data to determine where the loss was greatest? Was it due to just one or two items, or is there a bias evident across all of the items? The next model in the nested sequence could specify the latent means of the constructs to be invariant between groups. Expectations as to the outcome of this test would be set up by the hypotheses. Namely, is one group supposed to be significantly higher or lower than the other group, or should they possess equal means?

Examples of the application of these procedures can be found in the *Journal of Applied Psychology* (Vandenberg and Self, 1993), and the *Journal of Management* (Riordan and Vandenberg, 1994). My overall point, though, is that EFA and CFA can each have a role in addressing the issues raised, but my recommendation would be to start with CFA since it provides a strict test of equivalence across the groups. If nonequivalencies are found, then EFA and other exploratory mechanisms may be employed to discover where the anomalies are in the database. For the interested reader, I would also encourage her/him to look at Byrne, Shavelson and Muthen (1989), and Marsh and Hocevar (1985).

CS: I guess my reaction to this is that EFA might make some sense if you ran a CFA and could not find a good fitting model. We might then drop back to the position of asking, 'What's going on in these data?' For this question, EFA makes sense. However, let me add that one of the

'niceties' of doing CFA lies in its potential to do multi-sample or stacked modeling which looks at the similarity of parameter estimates across samples (and further decreases concern about parameter estimates which capitalize on chance error). This is potentially quite useful in this domain.

AS: I'm going to disagree with Bob. I think he started out by saying if you're a good researcher you're not going to do something inappropriate and I think the very existence of the disagreements among researchers indicate that we may be well advised to be a little more cautious here. One of the maxims in which I was schooled, even if I sometimes fail to heed it, was put to me as the no. 1 rule in research and that is, don't fool yourself. We all try to look at models of the world that fit what we think we see and it's altogether too easy given the sophistication of our tools to fall into the seduction of thinking that sophisticated research techniques overcome the weaknesses of human judgment. We still can see what we want to see whether it's there or not.

AH and TS: What about the frequent changes people make to their models based on their CFA results?

BV: I find that using CFA to explore is one of the traps researchers often fall into. The following example typifies this trap. A researcher sets out an hypothesis, examines the results with that hypothesis, and then revises the hypothesis and reanalyzes the same data. To expand upon this trap, let me just quote directly from Velicer and Jackson (1990, p. 21).

'Researchers are prejudiced against the null hypothesis and tend to persevere by modifying procedures and/or fail to report when results do not support the theory. When results do agree with the theory, researchers tend to over-generalize, that is, independent of conditions and procedures. This reflects a phenomenon that has been labeled a confirmation bias. A confirmation bias has been demonstrated to result in a wide variety of effects including a perseverance of belief in discredited hypotheses. A review of some of the recent articles employing the LISREL analysis will provide the reader with examples of initial rejection of the hypothesis followed by changes indicated by the modification indices, such as extensive correlated error terms for non-longitudinal studies, to get the model to fit'.

Two of the many problems created by this trap are (Gorsuch, 1983): one, because there is explicit examination of the data for rewriting the hypotheses, there is capitalization upon chance. This makes the significance tests not just meaningless, but actually misleading. The second problem flows simply from the subjectivity of the procedure. The rewriting of hypotheses will be influenced by the investigator's own personal opinions, and as such, this practice borders on being nonscientific. Further, to writeup the results as CFA at this point is misleading to readers.

LW: There is considerable evidence that modification indices capitalize on chance and that models that are revised based on them will not replicate. So, while I tend to be a relativist, when it comes to modification indices I see them to be of little value. Also, as we think about the progression of what we're doing as we proceed from EFA to CFA, one would hope that the problems associated with high modification indices would have manifested themselves at earlier stages in EFA.

CS: Probably the greatest single mistake that was done in producing LISREL 8 is that modification indices are now routinely given and nobody knows what havoc that's probably wrought in terms of people not reporting 'snooping' but actually doing so. Call me what you will but I'm suspicious. CFA is supposed to be theory driven and, with modification indices, you're letting the

data tell you what you should be doing. That's exploratory, not confirmatory. Now, if you've tried CFA and your model didn't confirm, you're scratching your head. My argument about EFA is that if a CFA doesn't work out you may be compelled to do EFA and say 'I give up, so tell me'. However, at that point in time you can't purport to have done a CFA—you've done an EFA.

AH and TS: Let's discuss the Goodness-of-Fit Indices.

MB: Nobody pays any attention to the statistical tests the CFA approach provides. Instead, there is a fit statistic of the day (or maybe 10 or 20 of them) that will be inspected to determine whether the model provides a good fit to the data. I object to these on several grounds, including (1) they misbehave because of sensitivity to sample size, (2) they do not necessarily agree with one another when it seems they should, (3) they allow the researcher to choose one to suit the conclusion they wish to draw rather than being chosen *a priori* to provide a conclusion.

It is often stated that an advantage of CFA and structural equation models generally is that such an approach provides a statistical test of the hypothesized model. Referring to the simulated data mentioned before, Table 3 shows some LISREL fit statistics for this sample data. The χ^2 test in Table 3 convincingly shows, as it should, that the sample data were not likely drawn from the hypothesized population. Based on the statistical test, we can reject the model. Conventional wisdom is to ignore the χ^2 test and to examine other fit indices. Most of the rest of the fit statistics shown in Table 3 support the viability of the model according to current convention.

Table 3. LISREL fit statistics

Statistic	Value
χ^2 ($df = 402$)	567.02 ($p = 0.0000001$)
RMSEA	0.029 (test of close fit $p = 1.00$)
Root mean square residual	0.048
Goodness of Fit index (GFI)	0.93
Adjusted GFI	0.92
Parsimony GFI	0.80
Non-normed fit index	0.97
Parsimony normed fit index	0.84
Comparative fit index	0.97
Incremental fit index	0.97
Relative fit index	0.91

BV: I both agree and disagree with Mike's perspective. My disagreement is simply the fact that I believe he is overgeneralizing the concern, and the situation is not as 'poor' as it appears. I agree, however, with his general observation as to the number of Goodness-of-Fit Indices (GFIs) available, and the apparent inattention researchers have given to their relative strengths and weaknesses before selecting one or more for their purposes. This inattention, however, is not necessarily the fault of the researcher *per se*, but rather, in my opinion, being unaware of the literature on GFIs, particularly that literature comparing the characteristics of each GFI with one another. For example, there is an excellent chapter by Tanaka in Bollen and Long's (1993) edited Sage book in which he compares a number of GFIs. I find this chapter to be quite useful because it examines the appropriateness of certain GFIs relative to the design characteristics of your study. Thus, one can at least look to see if a certain GFI or GFIs fit with the design of the study.

CS: Paraphrasing an observation made in another domain, the 'thing' about goodness-of-fit indices is that many people have a favorite. Personally, my observations suggest the following two points. First, there are no clear-cut 'superior' indices so that, when possible, one should not rely on a single index but use several. Second, the various indices differ in terms of what they are trying to index. Thus, one needs to be familiar with the various options and select one which seems to best fit the use at-hand. One good source which may be helpful in making a selection decision is the Medsker, Williams and Holahan (1994) article in the *Journal of Management*. A key caution in this area is, of course, that the indices used to assess fit should be selected (based upon theoretical considerations of appropriateness) before they are computed. Unfortunately, my suspicion is that people routinely scan the fit indices which are automatically provided by their software and then use those which they like.

AS: I would suggest that the pragmatic concern would be for authors to be expected to explicitly address their choice among the options for evaluating how well models fit the data. In parallel, paper reviewers should consider how clearly each manuscript conveys its selection of fit indices just as they would expect a publishable manuscript to clearly articulate the appropriateness of its methodology with respect to other elements of research design. Just as the lack of an appropriate reason to conduct a CFA in the first place would constitute a weakness of a particular study, the lack of any reason for the appropriateness of chosen fit indices would constitute a methodological weakness. In all cases, what we owe to the readers of our journals is the assurance that a rigorous approach to ferreting out what can be learned from the evidence at hand provides us with confidence that our findings transcend the artifactual.

LW: My main concern about the use and interpretation of GFIs associated with CFA is that many organizational researchers do not fully understand how they work and what they reflect (see Medsker *et al*. (1994) for a general discussion). For the present discussion, it is important to note that the values for GFIs reflect the appropriateness of the constraints (restrictions of paths, mostly to zero) that are proposed with the specification of a model. Everything else being equal, the more constraints in a model, the lower the obtained values of GFIs will be (e.g. Williams and Holahan, 1994). This will be especially true when the paths that are proposed to be zero have non-zero values when estimated in a different analysis. Now let's consider the typical measurement situation which confronts an organizational researcher. The associated CFA model will have a large number of factor loadings constrained to zero, and this number will increase as the number of items and the number of proposed factors increase. This happens because for every item that is added to the analysis, the associated loadings on the factors other than the one proposed to be linked to the item are set to zero. Similarly, for every factor or latent variable that is added to a model, the loadings linking the factor with items associated with the other factors are set to zero. Thus, many measurement situations for organizational researchers involve CFA models with a relatively large number of items and constrained factor loadings.

These highly constrained models, with many factor loadings restricted to zero, are predisposed to have relatively lower fit values. This outcome is most likely to occur when those factor loadings that are constrained to zero in CFA are shown with EFA to have non-zero acceptable values. Thus, one can begin with an EFA and show that all of these factor loadings (those other than the loadings which link a specific item and its intended factor) are relatively small (in many cases we are thrilled if they are less than 0.20). However, in the associated CFA the constraints of these loadings to zero will necessarily lower the GFIs, which increases the chances that the researcher will reject the model. I would think that in many such cases, the rejection of the CFA model based only on the GFI values would be premature and ill-advised. In short, it is imperative for

organizational researchers to use the full range of diagnostics to evaluate a CFA model (e.g. analysis of residuals, calculation of item and latent variable reliabilties) and not just rely solely on GFI values.

AH and TS: Here's the opportunity for each of you to tell us your closing thoughts.

MB: Throughout the paper I have tried to make three points. First, the difference between exploratory and confirmatory analyses is not about the superiority of newer versus older computer programs and loss functions. Confirmatory techniques test hypotheses about population factor structures based on sample data. Exploratory techniques attempt to describe, summarize, or reduce data to make them more easily understood. Both older and newer computer programs can be used for either EFA or CFA, although the older programs tend to be more useful for EFA and the newer programs tend to be more useful for CFA. Second, the newer methods do not necessarily do a better job of answering research questions than do older methods. The better technique most closely matches the intended use. Finally, in the case of a 'clean' factor structure, there will be little difference in the interpretation of estimates of factor pattern, factor correlation and uniqueness matrices from the newer and more traditional factor analytic methods.

I would encourage both the researcher and the reviewer to pay close attention to the intent of each particular study. Rather than providing progress, the newer techniques may only provide more powerful tests of hypotheses that were not intended to be tested.

BV: In keeping with my traditionalist perspective, I absolutely disagree with those who feel compelled to anchor themselves to one technique or the other. Doing so fails to help us 'commoners' recognize that in reality both are part of one overarching paradigm (Gorsuch, 1983). That is, we are all interested in identifying the common ground upon which our observations rest. Both EFA and CFA are tools for doing so. Like all tools their applicability is totally dependent upon our own understanding of whether our situation meets the assumptions underlying those tools.

Thus, I totally agree with Mike's encouragement above, and would add to it in the following manner. If we as researchers do not fully understand the appropriateness of each tool, then I would encourage those researchers to engage in some 'continuing education' and obtain the knowledge. I firmly believe that what continues to fuel this debate is in part due to researchers themselves not fully understanding that these tools belong to this one overarching paradigm. Since our current students will be the future researchers and reviewers, it is imperative that we train them in the proper application of these tools. This begins when we as the mentors of those students are engaging in appropriate practices ourselves.

CS: Again, I'd like to echo Bob Vandenberg's sentiments about how important it is that we are not wedded to any one data-analytic technique. All have their strengths and weaknesses and there are appropriate and inappropriate situations for the use of each. However, let me add the closing note that I'm excited about the use of SEM and that I view it as one of the major advancements in our field in the past several decades. SEM forces the scientist to be theoretical; it also forces the scientist to consider the linkages between theory, data collection, and measurement. These are vital to enhancing the rate of progress in our collective enterprise of advancing knowledge about human behavior in work organizations.

AS: In closing, what comes to mind regarding the notion of EFA versus CFA is an analogy that I'm sure my students tire of hearing. Would we ask a carpenter to choose between carrying a saw versus carrying a hammer? A useful toolbox needs to be big enough to carry both, and

a competent carpenter needs to recognize when a board is too long versus when one board needs to be nailed to another. Yet the structures we build as researchers are conceptual rather than material. In our building process the most consequential matter is whether those who would treat our contributions as a foundation upon which to build further find that our work clearly articulates a conceptual basis. In this respect I would second the recommendations made by the authors that both researchers and reviewers emphasize the extent to which assumptions are explicated and that methodological choices adhere to the required assumptions.

LW: I feel that exchanges such as those provided by this paper advance the understanding of methodologies that organizational researchers use. There should be no disagreement that the state of organizational science, across its various subdisciplines, can only be improved with increased attention to measurement issues. I hope that the present discussion of two important tools for examining measurement concerns will contribute to the advancement of this cause. We need to move beyond simple arguments that EFA is not needed with CFA, or that CFA should be ignored because its assumptions are never met. The informed combined use of these two techniques should be our objective.

Discussion

We have learned how important it is for researchers to scrutinize their decision regarding the analysis they will utilize in their studies. It is necessary to examine the purpose of our study and our data along with its collection method before proceeding to analyzing data. Researchers must be able to logically substantiate their rationale for utilizing EFA or CFA in their data analysis.

We also learned that this debate will probably never reach an end where EFA or CFA is declared to be the 'winner'. There was a consensus among the authors that there is a place for both types of factor analysis in our field, and that the appropriateness of each depends on the study context. All of the authors agreed on the importance of *a priori* theory before beginning CFA. The authors also concurred that CFA should not be misused as an exploratory technique. However, if a decision is made to perform CFA, the next issues to be considered are the GFIs and respecification of the model. There are enormous misunderstandings regarding what fit statistics tell us about our data. Although the authors did not agree on the merits of various GFIs, they did agree that the rationale for choosing specific GFIs should be included in manuscripts. One area that there was clear consensus on between our authors was not using modification indices to respecify models and then run the analysis again. All of the authors agreed that, at the very least, a new sample would need to be collected for verification.

There is a need for agreement on a set of standard EFA and CFA procedures that researchers can use as a guide, that can be taught in doctoral programs, and employed by reviewers in evaluating manuscripts. While our debate did not produce this, it did frame the significant issues. Below we have identified the areas which need to be addressed in guidelines for EFA and CFA. (1) When should EFA be used? CFA? (2) The role of CFA in scale development. (3) Should both EFA and CFA be used on the same data set? (4) Should models be changed based on CFA results? (5) Appropriateness of 'forcing' models into a preset number of factors.

Our session also offered hope that publishing exploratory or confirmatory factor analysis in scholarly journals would not rely on the luck of the draw as to reviewers but more on the researcher's ability to support their decision to use EFA or CFA. We hope this exchange has informed researchers as well as reviewers.

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.

Bollen K. A. and Long, J. S. (1993). *Testing Structural Equation Models*, Sage, Newbury Park, CA.

Brannick, M. T. (1995). 'Critical comments on applying covariance structure modeling', *Journal of Organizational Behavior*, **16**, 201–214.

Brannick, M. T. and Spector, P. E. (1990). 'Estimation problems in the block-diagonal model of the multitrait–multimethod matrix', *Applied Psychological Measurement*, **14**, 325–339.

Byrne, B. M., Shavelson, R. J. and Muthen, B. (1989). 'Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance', *Psychological Bulletin*, **105**, 456–466.

Gerbing, D. W. and Hamilton, J. G. (1996). 'Viability of exploratory factor analysis as a precursor to confirmatory factor analysis', *Structural Equation Modeling*, **3**, 62–72.

Gorsuch, R. L. (1983). *Factor Analysis*, 2nd edn, L. Erlbaum Associates, Hillsdale, NJ.

Kelloway, K. E. (1995). 'Structural equation modeling in perspective', *Journal of Organizational Behavior*, **16**, 215–224.

Kenny, D. A. and Kashy, D. A. (1992). 'Analysis of the multitrait–multimethod matrix by confirmatory factor analysis', *Psychological Bulletin*, **112**, 165–172.

Kerlinger, F. N. (1986). *Foundations of Behavioral Research*, 3rd edn, Holt, Rinehart & Winston, New York.

Medsker, G., Williams, L. and Holahan, P. (1994). 'A review of current practices for evaluating causal models in organizational behavior and human resources management research', *Journal of Management*, **20**, 439–464.

Marsh, H. W. and Hocevar, D. (1985). 'Application of confirmatory factor analysis to the study of self-concept: First- and higher-order factor models and their invariance across groups', *Psychological Bulletin*, **97**, 562–582.

Mulaik, S. A. (1987). 'Toward a conception of causality applicable to experimentation and causal modeling', *Child Development*, **58**, 18–32.

Mulaik, S. A. (1972). *The Foundations of Factor Analysis*, McGraw-Hill, New York.

Nunnally, J. C. and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd edn, McGraw-Hill, New York.

Pearson, K. (1901). 'On lines and planes of closest fit to systems of points in space', *Philosophical Magazine*, **6**, 559–572.

Riordan, C. M. and Vandenberg, R. J. (1994). 'A central question in cross-cultural management research: Do employees of different cultures interpret work-related measures in an equivalent manner?' *Journal of Management*, **20**, 643–671.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C.C. and Lankau, M. J. (1993). 'Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments', *Journal of Management*, **19**, 385–417.

Spearman, C. (1904). 'General intelligence, objectively determined and measured', *American Journal of Psychology*, **15**, 201–293.

Stone-Romero, E. F., Weaver, A. E. and Glenar, J. L. (1995). 'Trends in research design and data analytic strategies in organizational research', *Journal of Management*, **21**, 141–157.

Vandenberg, R. J. and Self, R. M. (1993). 'Assessing newcomers' changing commitments to the organization during the first 6 months of work', *Journal of Applied Psychology*, **78**, 557–568.

Velicer, W. F. and Jackson, D. N. (1990). 'Component analysis vs. common factor analysis: Some issues in selecting an appropriate procedure', *Multivariate Behavioral Research*, **25**, 1–28.

Wherry, R. J. (1975). 'Underprediction from overfitting: 45 years of shrinkage', *Personnel Psychology*, **28**, 1–18.

Williams, L. J. (1995). 'Covariance structure modeling in organizational research: Problems with the method versus applications of the method', *Journal of Organizational Behavior*, **16**, 225–234.

Williams, L. J. and Holahan, P. (1994). 'Parsimony based fit indices for multiple indicator models: Do they work?' *Structural Equation Modeling: A Multidisciplinary Journal*, **2**, 161–189.

Williams, L. J. and Podsakoff, P. M. (1989). 'Longitudinal field methods for studying reciprocal relationships in organizational behavior research: Toward improved causal analysis'. In: Staw, B. and Cummings, L. (Eds) *Research in Organizational Behavior*, JAI Press Inc, Greenwich.

Williams, L. J., Cote, J. A. and Buckley, M. (1989). 'The lack of method variance in self-reported affect and perceptions at work: Reality or artifact?', *Journal of Applied Psychology*, **74**, 462–468.
Williams, L. J., Gavin, M. B. and Williams, M. L. (1996). 'Investigating measurement and non-measurement processes with method effect variables: An example with negative affectivity and employee attitudes', *Journal of Applied Psychology*, **81**, 88–101.

