Multivariate Behavioral Research, 25 (1), 41-45

Some Additional Thoughts on Components, Factors, and Factor Indeterminacy

James H. Steiger
University of British Columbia

Velicer and Jackson (1990, hereafter referred to as VJ) have performed a valuable service in presenting this intelligent and very useful summary of information on an important practical issue. Velicer has, of course, been a leader in developing this information. The fact that he and Jackson are sophisticated practitioners of multivariate analysis in psychometric applications lends added force to their arguments.

VJ (1990) demonstrate with a wealth of evidence that there are few data sets where component analysis and factor analysis would lead to different substantive conclusions. Hence, although it would seem logical to say that in most cases a theoretical problem like factor indeterminacy would have little impact on what people use factor analysis for, it would also appear that the advantages of factor analysis are largely illusory.

VJ (1990) are surprisingly gentle in their conclusions. After reading the article and keeping score of the various points they raise, I was left with the distinct feeling that factor analysis had taken a beating, and that VJ were taking pains to be kind. If the article were a football game, the score (in a metric I know Velicer is fond of) would be something like — Component Analysis 49, Factor Analysis 14.

In making their case, VJ (1990) left few stones unturned. Most importantly, they based their analysis on objective information, rather than cliches. Their assertions will be difficult to refute.

But common factor analysis sometimes evokes a kind of religious fervor in its proponents (is Velicer the next factor analytic Salmon Rushdie?), and keepers of the factor analytic flame will inevitably offer their rebuttals. Some of these rebuttals will deal with areas touched on only lightly by VJ (1990). Here are two I've heard already, and my responses to them.

1. *Factor analysis is a model. Component analysis is a tautological data reduction system. Factor analysis admits statistical testing. Component analysis does not.*

Factor analysis *is* a model. The probability that it fits any data set exactly, for a number of factors less than Lederman's number, is essentially zero. Hence,

for real data, the factor analysis model is an approximation. In many cases it is not a good approximation, as evidenced by the extraordinary number of Heywood cases encountered in practice. Nevertheless, in some cases the factor analysis model seems a natural one to consider. For example, when one is recording a signal at several locations each subject to independent random noise interference, the single factor model seems appropriate. Interestingly, it is precisely in such a case where factor scores might be of paramount interest.

Obviously, one can conceive of countless hypothetical examples where either the common factor model or some component model (not necessarily principal components) would be more appropriate. But usually either approach will furnish only an approximation to reality.

Component *models* can be proposed, and can be tested, so long as one does not suffer from a terminal lack of imagination while contemplating the meaning of the word *model*. In fact, addressing the question of how path models involving indeterminate latent variables can be rephrased as component models sheds interesting light on the general role of statistical testing in latent variable models. Space does not permit me to expand on this notion here. Component models and factor models do not overlap completely. But they certainly have important elements in common.

2. *Factor Indeterminacy has been resolved. Read the writings of eminent statisticians like Bartholomew (1981) and Williams (1978).*

I urge everyone to read the interesting work of Bartholomew and Williams, rather than relying on someone else's interpretation. Careful readers will find neither has *solved* indeterminacy. They have simply rephrased it with overlays of statistical sophistication.

Williams (1978) shifts the foundation of the factor model from a finite set of p variables, to an infinite set. In this "behavior domain," either a unique solution for the common factors exists, no solution exists, or multiple solutions exist. Williams' article was mathematically important in the sense that it provided a rigorous formalization of the behavior domain notion.

Practically, the article offered no solution to the problem of factor indeterminacy. Factor indeterminacy exists when one has p variables, and the m factors are not determinate. To say that these p variables could have been sampled from an infinite domain where there is no factor indeterminacy is hardly any help. How does one know what is out there in the behavior domain? Moreover, (this is one of the more closely guarded secrets of the factor analytic sect) how does one define a *behavior domain* in a non-circular fashion?

Williams, beginning with great optimism, promises a "complete solution" to indeterminacy problems in his opening paragraph. He is still riding high on page 303, when he refers derisively to the published history of the "so-called

problem" of factor indeterminacy. Up to this point, Williams has enjoyed the luxuries of engaging in his mathematical exercise, and demeaning his predecessors without even bothering to cite them. Now it is time to deliver the goods.

It turns out Williams has nothing to deliver. He says "no adequate model has ever been set out before..." In other words, the model that everyone else was studying (i.e., the one based on the p variables they had to factor analyze) was the wrong model. The right model was the one which included infinitely many variables which they would never be able to analyze. Why was this the right model? Presumably because one could imagine factor indeterminacy going away in this model.

Williams, referring to his predecessors' shortcomings, says their problem was "that the concept of a random variable was not understood well enough..."

"If Ω contains no finite subset having unit probability, then it is possible to define at least countably many linearly independent random variables on ... [the probability space]. The vast majority of these have no extra-mathematical meaning at all. Such meaning derives only from observations which can be made on a random variable, and this is not possible for factor scores, or from extra-mathematical interpretation of all properties of a well-defined method for constructing the random variable, an approach which can be adopted for factor scores. All other random variables, which can be defined, and therefore exist, may as well be ignored." (Williams, 1978, p. 303.)

The anxious reader searching for a key to this Rosetta stone should be reassured. Williams was merely, by fiat, declaring the "construction approach" (Steiger, 1979) to factor indeterminacy to be off bounds.

By page 305, it is clear Williams has no solution to the old-fashioned indeterminacy problem. He advises users that the observed variables "should be selected to make the characteristic roots of $[\mathbf{F}_p'(\mathbf{U}_p\mathbf{U}_p')^{-1}\mathbf{F}_p]^{-1}$ equal and as small as possible." Apparently membership in the factor analytic sect confers special powers of clairvoyance. Imagine knowing F and U before you even perform the factor analysis!

By the last pages of his article, Williams is reduced to (a) giving an (obviously) incorrect theorem on unidentifiability, and (b) advising users to check indeterminacy indices. In retrospect, the whole effort seems to have fizzled badly.

Bartholomew (1981) fares only slightly better. Like Williams, he attempts to defuse indeterminacy by casting it in a revised statistical framework. His key device is Bayesian statistics. We learn that factors based on the factor analysis of p variates should not be thought of as indeterminate, but, rather, as having a "posterior distribution" which is not a point distribution. To understand what Bartholomew is saying, consider the following simple example which exemplifies most of the key elements of the indeterminacy issue.

There is a model, called Model A. It expresses observed variable X in terms of a single underlying unobservable variable Y. It has many important characteristics in common with the factor model. It is that

$$Y^2 - 100Y + X = 0$$

X is known to have a discrete distribution. Only two pairs of values ever occur. They are 99 and 2100. They occur equally often. In this case, what can be said about latent variable Y? Suppose, for example, X is 99. Obviously, Y, to fit X and the model, must be either 1 or 99. Equally obviously, it cannot be both. Suppose that our model is correct. Now, it could well be the case that, on all occasions where X is 99, Y is also 99. Or, alternatively, Y could always be 1 on the occasions when X is 99. We say that Y is indeterminate.

Bartholomew's way of expressing this fact is to say that the conditional distribution of Y, given X, is not a point distribution. The conditional distribution of Y given X=99 is easily calculated from Bayes' theorem. It depends, of course, on the prior distribution for Y, which is somewhat arbitrary. If one assumes equal prior probabilities for all possible values of Y (there are four in this simplified example) then of course the posterior probability distribution for Y given X=99 assigns probabilities of .50 to Y=1 and .50 to Y=99. The "conditional mean" (corresponding to the regression estimates in factor analysis) is 50.

But 50 does not fit the model! Suppose Y were to represent a percentage performance of some kind. We are, in Bartholomew's vision, uncertain as to whether Y=99 or Y=1. (Peter Schönemann and I would say we are absolutely certain that Y is *either* 99 *or* 1!) In other words, performance is *either* great *or* terrible. It is certainly not mediocre! What sense, then, does it make to put forth the conditional mean as some kind of compromise solution?

The answer is that it frequently makes no sense at all. To say, for example that you are *either* very heavy *or* very light, and that you are not sure which, is quite different from saying that you are of average weight.

Bartholomew concedes that the unknown prior distribution is a serious barrier to the use of factors, and he concludes components are more useful than factors themselves. Bentler (1985) gently reminds him that this fundamental point was anticipated by Bentler (1976) and Schönemann and Steiger (1976).

Some of Bartholomew's readers seem to believe that he has somehow solved factor indeterminacy with his approach. For example, Aitkin (1985) states that the "re-interpretation of factor score 'estimates' as the conditional expectation of random variables is an important step forward, however, and lays to rest the past arguments over the status of factor scores." In view of the above simple Model A example, it is difficult to imagine why Aitkin might feel this

way. In the style of Williams, he is not at all explicit about what doubts he thinks have been laid to rest. Perhaps Aitkin was confused by statements like "indeterminacy is simply a reflection of the fact that Y is still a random vector after X has been observed ... Only if the posterior probability were to be concentrated on a single point ... would the ys be determinate. To speak of indeterminacy as a 'problem' is thus to overlook the essentially random character of the quantities concerned." (Bartholomew, 1981, p.97) Bartholomew certainly seems to have convinced himself. But why, after observing an X of 99, does thinking of Y as a random variable eliminate the indeterminacy problem? We still don't know whether Y=1 or Y=99.

References

Aitkin , M. (1985). Comments on D. J. Bartholomew, Foundations of factor analysis: some practical implications. *British Journal of Mathematical and Statistical Psychology, 38*, 127-128.

Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 34*, 93-99.

Bentler, P. M. (1976). Multistructure model applied to factor analysis. *Multivariate Behavioral Research, 11*, 3-25.

Bentler, P. M. (1985). On the implications of Bartholomew's approach to factor analysis. *British Journal of Mathematical and Statistical Psychology, 38*, 129-131.

Schönemann, P. H. and Steiger, J. H. (1976). Regression component analysis. *British Journal of Mathematical and Statistical Psychology, 29*, 175-189.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika, 44*, 157-168.

Williams, J. S. (1978). A definition for the common factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika, 43*, 293-306.

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