Bulletin of the Psychonomic Society
1978, Vol. 12 (4), 287-290

On the validity of indeterminate factor scores

PETER H. SCHÖNEMANN and JAMES H. STEIGER
Purdue University, Lafayette, Indiana 47907

A partition of the vector space of all deviation score vectors for fixed sample size N is used to show that the (indeterminate) factors of the factor model can always be constructed so as to predict any criterion perfectly, including all those that are entirely uncorrelated with the observed variables.

STATEMENT OF THE PROBLEM

The latent variables of the factor model, the "factors," are not uniquely defined by the observed variables. This peculiar property of the factor model is called "factor indeterminacy." Some have argued that the implications of this indeterminacy are "trivial," because the covariance relations of the observed variables with the common and unique factors remain unchanged for all possible assignments of the factors. But if this were not the case, we would not have an indeterminacy.

Whether or not the consequences of indeterminacy are trivial cannot be decided in the abstract, but rather depends on what one wants to do with the factors. Presumably, one uses factor analysis to learn something about the given variables that could not be learned without it. Since the interrelations among the given variables are, by definition, known, this additional knowledge must pertain to their relation with other variables (e.g., criteria) that are not observed at the time of the analysis.

Once one accepts the fact that the factors are not uniquely defined by the model, one faces the question of how the indeterminate increment that is needed to define them can possibly enhance our knowledge of other variables. In the past, opinions were divided on this question. Some have argued that factor analysis is superior to component analysis, which defines new variables simply as linear combinations of the observed variables, precisely because the factors, in contrast to components, "go beyond the test space."

The purpose of this note is to lay the ground for a rational study of this question, which, on the surface, concerns the relation between factor indeterminacy and external prediction, and which, in the final analysis, concerns the purpose of factor analysis as a scientific method.

VECTOR SPACE NOTATION

In studying the relationship between factor indeterminacy and external validity, it will be convenient to use the language of abstract vector spaces. The needed definitions and notation will be summarized in this section, together with some elementary but useful results from the algebra of vector spaces.

Let R^N be the vector space of ordered N-tuples $x = (x_i)$, $i = 1,N$, with real-valued components x_i. Let $J \in R^N$ denote the column vector of N ones, so that its transpose is:

$$J' = (1, 1, \ldots, 1). \tag{1}$$

We call any vector $x \in R^N$ a "vector of deviation scores" if it is in the null space of J', so that:

$$J'x = 0, \tag{2}$$

that is, its sample mean is $\bar{x} = \Sigma x_i/N = J'x/N = 0$. As is well known, this null space (i.e., the space of all deviation score vectors for fixed sample size N) has dimension $N - 1$. We denote this space "Sp" and write $Sp \subset R^N$ to indicate that it is a proper subspace of R^N.

If $v_1, \ldots v_r$ are the r columns of an N x r matrix $V = (v_1, \ldots v_r)$, we denote its column space (i.e., the set of all linear combinations of its columns) by Sp(V). We write, for $V_1 \in Sp(V)$, $V_2 \in Sp(V)$:

$$Sp(V) = Sp(V_1) \oplus Sp(V_2) \tag{3}$$

iff $v \in Sp(V) \Rightarrow v = v_1 + v_2$, with $v_1 \in Sp(V_1)$, $v_2 \in Sp(V_2)$, and for a given V, V_1, V_2, and v, the vectors v_1 and v_2 are uniquely defined. We call Sp(V) the direct sum of $Sp(V_k)$, $k = 1,2$. It is well known that in this case the dimensions add,

$$\dim Sp(V) = \dim Sp(V_1) + \dim Sp(V_2). \tag{4}$$

In general, $Sp(V_2)$ is not uniquely defined by Sp(V), $Sp(V_1)$ in Equation 3. If one defines $Sp(V_2)$ as an "orthogonal complement of $Sp(V_1)$" [relative to Sp(V)] and writes

$$Sp(V_1) \perp Sp(V_2) \tag{5}$$

iff $v_1 \in Sp(V_1)$, $v_2 \in Sp(V_2) \Rightarrow v_1'v_2 = 0$, then $Sp(V_2)$ is uniquely defined in terms of Sp(V) and $Sp(V_1)$, as is well known. We therefore could define Sp simply as the orthogonal complement of Sp(J) in R^N:

$$R^N = Sp \oplus Sp(J), Sp \perp Sp(J) \tag{6}$$

THE FACTOR MODEL

Let $Y = (y_j)$ be any set of p observed, linearly independent deviation scored vectors. They span a subspace $Sp(Y) \in Sp$ of exactly p dimensions, for which Y is a basis. Suppose Y can be written

$$Y = XA' + ZU \tag{7}$$

for some $p \times m$ ($m < p - 1$) full column rank matrix A and some positive definite (p.d.) diagonal matrix U^2 of order $p \times p$,

James H. Steiger is now at the Department of Psychology, University of British Columbia, Canada.

and X(N × m) and Z(N × p) contain column vectors of deviation scores that jointly satisfy

$$\begin{bmatrix} X' \\ Z' \end{bmatrix} (X,Z)/N = I_{p+m}. \tag{8}$$

In this case, we say "Y satisfies the factor model with common factor pattern A, unique factor pattern U, common factor scores X, and unique factor scores Z." The definitions of the factor model (Equations 7 and 8) imply at once that the observed sample variance-covariance matrix

$$C = Y'Y/N \tag{9}$$

must be of the form

$$C = AA' + U^2. \tag{10}$$

Equation 10 has been called the "fundamental theorem of factor analysis." It states a necessary condition for the model (Equations 7 and 8) to hold for any observed Y. It has been used in practice to falsify this factor model for a hypothesized number of common factors m(< p − 1). Recently, Schönemann and Steiger (1976) have proven that this empirical check at the variance-covariance level does not suffice to validate the factor model (Equations 7 and 8) at the random variable level, because other ("regression component") decompositions of Y exist that also imply Equation 10, but that are quite different from the factor model, because they violate Equation 8. These regression component decompositions are falsifiable exactly to the same extent that the factor model is, and hence, they are empirically indistinguishable from it. They have the advantage that they do not suffer from the indeterminacy problems that result as a consequence of the full rank condition (Equation 8) of the factor model.

FACTOR INDETERMINACY

Wilson (1928), upon reviewing Spearman's (1927) *The Abilities of Man,* commented that the definitions of the factor model, Equations 7 and 8, imply that the "latent" random variables (i.e., in the sample, the factor score matrices X,Z in Equations 7 and 8) are not uniquely defined in terms of the observed information in Y that satisfies Equation 10 at the variance-covariance level. It was shown subsequently by several authors (Guttman, 1955; Kestelman, 1952; Piaggio, 1931) that many different pairs of deviation score matrices X,Z can be constructed for the same observed score matrix Y, and the same fixed pattern A,U, so that Equations 7 and 8 are met when Equation 10 is. If P is an arbitrary gram factor of $I - A'C^{-1}A$, that is, is defined (within rotations) by

$$PP' = I - A'C^{-1}A, \tag{11}$$

then any (X,Z) computed as

$$(X,Z) = (S_m,Y)T = (S_m,Y)\begin{bmatrix} P' & -P'A'U^{-1} \\ C^{-1}A & C^{-1}U \end{bmatrix} \tag{12}$$

will satisfy Equations 7 and 8 if C satisfies Equation 10, provided the N × m matrix of deviation scores S_m satisfies

$$S_m'Y = 0,\ S_m'S_m/N = I_m,\ p + m < N - 1. \tag{13}$$

These conditions mean that S_m, which is needed in Equation 12 to complete the "determinate parts"

$$\hat{X} = YC^{-1}A,\ \hat{Z} = YC^{-1}U \tag{14}$$

in the construction of the factor score matrices X,Z is an orthogonalized basis of an arbitrarily chosen subspace $Sp(S_m)$ of m dimensions of the N − p − 1 dimensional null space Sp(S) ⊂ Sp of (Y,J). If S is a basis of this null space, we have

$$\begin{bmatrix} Y' \\ J' \end{bmatrix} S = 0, \tag{15}$$

and hence,

$$Sp = Sp(Y) \oplus Sp(S),\ Sp(Y) \perp Sp(S). \tag{16}$$

As long as N − 1 exceeds p + m, there is considerable leeway for the definition of S_m, and thus X and Z in Equation 12. To measure the extent of this "factor indeterminacy" for a given factor score column x_j in X, Guttman (1955) proposed the minimum correlation between the corresponding columns x_j, x_j^* of two equivalent matrices X, X* under variation of all S_m. This correlation is given by

$$\min_{S_m} r_{x_j x_j^*} = 2a_j'C^{-1}a_j - 1, \tag{17}$$

where a_j is the jth column in the common factor pattern A in Equation 7. Schönemann and Wang (1972) found that this correlation is in practice frequently negative, confirming what Guttman had suspected 20 years earlier: "It seems that the sought for traits are not very distinguishable from radically different possible alternative traits, for the identical factor loadings" (1955, p. 74). This possibility has not been unduly disquieting for most practitioners of factor analysis in the past. They simply ignored the indeterminacy altogether.

A PARTITION OF Sp AND THE ASSOCIATED ORTHOGONAL PROJECTORS

Using $\hat{X}$ in Equation 14, we define $Sp(Y_o)$ by

$$Sp(Y) = Sp(\hat{X}) \oplus Sp(Y_o),\ Sp(\hat{X}) \perp Sp(Y_o), \tag{18}$$

and using S_m in Equation 12, we define $Sp(S_o)$ by

$$Sp(S) = Sp(S_m) \oplus Sp(S_o),\ Sp(S_m) \perp Sp(S_o). \tag{19}$$

In view of Equation 16, we thus decompose the space of all deviation score vectors Sp into four subspaces that are pairwise orthogonal:

$$Sp = Sp(\hat{X}) \oplus Sp(Y_o) \oplus Sp(S_m) \oplus Sp(S_o). \tag{20}$$

For the dimensions, we have

$$\dim Sp(X) = \dim Sp(S_m) = m, \tag{21}$$

and

$$\dim Sp(Y_o) = p - m,\ \dim Sp(S_o) = N - p - m - 1.$$

None of these subspaces will be empty if

$$0 < m < p < m + p < N - 1, \tag{22}$$

as we shall assume. From Equation 12, we find

$$\begin{aligned} Sp(X) &= Sp(S_mP' + YC^{-1}A) \\ &= Sp(\hat{X} + S_mP') \subset Sp(S_m,\hat{X}) = Sp(S_m) \oplus Sp(\hat{X}). \end{aligned} \tag{23}$$

Now let $w = (w_i) \in Sp$ be any criterion vector of N deviation scores which is normed to satisfy

$$w'w/N = 1, \tag{24}$$

so that its sample variance is one. As is well known (e.g., Searle, 1966; Seber, 1966), the squared multiple correlation of such a $w \in Sp$ is regressed on any set of r linearly independent predictors $v_1 \ldots v_r$ in V is simply

$$R^2_{w \cdot V} = w' P_V w/N, \tag{25}$$

where

$$P_V = V(V'V)^{-1} V' = P^2_V = P'_V \tag{26}$$

is the orthogonal projector onto $Sp(V) \subset Sp$, and V is the basis for Sp(V). As is also well known (e.g., Pease, 1965, Chapter 11),

$$x \in Sp(V) \text{ iff } P_V x = x, \tag{27}$$

so that $x \in Sp(V) \Rightarrow R^2_{x \cdot V} = 1$, by Equation 26. Since P_V is symmetric,

$$x \perp Sp(V) \text{ iff } P_V x = 0, \tag{28}$$

whence $x \perp Sp(V) \Rightarrow R^2_{x \cdot V} = 0$. Moreover, if Sp(V) is decomposed into orthogonal subspaces, one finds

$$\begin{aligned} Sp(V) &= Sp(V_1) \oplus Sp(V_2), \\ &= Sp(V_1) \perp Sp(V_2) \Rightarrow P_V = P_{V_1} + P_{V_2}, \end{aligned} \tag{29}$$

so that $x \in Sp(V_1) \Rightarrow R^2_{x \cdot V_1} = R^2_{x \cdot V} = 1$.

In the present case, we find that the orthogonal projector for the total space of deviation scores, Sp, is given by

$$P^* = I - JJ'/N, \tag{30}$$

in view of Equations 6 and 29. Some other projectors of interest are

$$P_Y = YC^{-1} Y'/N, \tag{31a}$$

$$P_{\hat{X}} = YC^{-1} A(A'C^{-1} A)^{-1} A'C^{-1} Y'/N, \tag{31b}$$

$$P_{Y_o} = P_Y - P_{\hat{X}}, \tag{31c}$$

$$P_S = P^* - P_Y, \tag{31d}$$

$$P_{S_m} = S_m S'_m/N, \tag{31e}$$

for all S_m which satisfy Equation 13,

$$P_X = XX'/N = \hat{X},S_m) \begin{bmatrix} I & P \\ P' & P'P \end{bmatrix} \begin{bmatrix} \hat{X}' \\ S'_m \end{bmatrix} /N, \tag{31f}$$

$$P_{(X,Z)} = (XX' + ZZ')/N, \tag{31g}$$

$$P_{(Y,S_m)} = P_Y + P_{S_m}. \tag{31h}$$

TWO RESULTS ON EXTERNAL FACTOR VALIDITY

To see what happens if one predicts an external criterion $w \in Sp$ from the matrix of common factor scores, X, we need:

Lemma 1: If U^2 in Equation 10 is positive definite, then PP', and hence P'P are positive definite.

Proof: If U^2 is p.d., one finds from Equation 11 that $PP'(I + A'U^{-2} A) = I_m$. Hence, PP', and thus also P, are of full rank m. Since both PP' and P'P are Gramian by definition, they are both p.d., q.e.d.

We now prove:

Theorem 1: If $N - 1 > p + m$, the factor model (Equations 7 and 8) implies the existence of criteria that, although perfectly correlated with the observed scores Y (in a multiple-regression sense), remain completely unpredictable from the common factor scores X under all choices of S_m in Equation 12. It also implies the existence of criteria that, although entirely uncorrelated with the observed scores Y, are positively correlated with suitably defined common factors X.

Proof: Let $w_1 \in Sp(Y_o)$. $Sp(Y_o) \subset Sp(Y) \Rightarrow R^2_{w_1 \cdot Y} = 1$, by Equation 29. $Sp(Y_o) \perp Sp(X)$, $Sp(Y_o) \perp Sp(S_m) \Rightarrow Sp(Y_o) \perp Sp(X) \subset Sp(\hat{X}) \oplus Sp(S_m)$, by Equation 23. By Equation 28, $w_1 \perp Sp(X) \Rightarrow R^2_{w_1 \cdot X} = 0$, for all S_m. Now let $w_2 \in Sp(S)$. $Sp(S) \perp Sp(Y) \Rightarrow R^2_{w_2 \cdot Y} = 0$, by Equation 28. We can use $w_2 \in Sp(S)$ to construct a basis S_m to define $Sp(S_m)$ as follows: We use w_2 as the first basis vector and add $m - 1$ arbitrarily chosen nonnull $s_j \in Sp(S)$ that satisfy $s'_j w_2 = 0$, $s'_i s_j = 0$, for $i \neq j$, and $s'_j s_j/N = 1$. Then $S_m = (w_2, s_1, \ldots, s_{m-1})$ satisfies Equation 13 and we have $w'_2 S_m/N = b' = (1, 0, \ldots, 0) \neq \phi'$. Using this S_m to construct X in Equation 12, we find

$$R^2_{w_2 \cdot X} = w'_2 P_X w_2/N = (\phi', b') \begin{bmatrix} I & P \\ P' & P'P \end{bmatrix} \begin{bmatrix} \phi \\ b \end{bmatrix} = b'P'Pb.$$

By Lemma 1, P'P is p.d. Hence, $R^2_{w_2 \cdot X} > 0$, q.e.d.

In short, some criteria that are entirely uncorrelated with the observed information in Y are better predictable from suitably defined common factors X then others that are perfectly predictable from Y.

We now turn to the prediction of w from both X, the common factor scores, and Z, the unique factor scores, jointly. So far as we know, no one has seriously advocated using the unique factor scores for external prediction. But neither can we think of any a priori reason why one could not also include Z in a regression equation to predict some criterion w. It would not strike us as any more unreasonable than to base such a prediction on X alone, from what we now know. If X and Z were both used to predict w, a result even more startling than Theorem 1 can be achieved. To prove this result, we need:

Lemma 3: $Sp(X,Z) = Sp(S_m,Y)$.

Proof: The mapping from Sp(X,Z) onto $Sp(S_m,Y)$ is isomorphic since T in Equation 12, which represents this mapping relative to the bases (X,Z) and (S_m,Y) has inverse

$$T^{-1} = \begin{bmatrix} P & A' \\ -U^{-1} AP & U \end{bmatrix} \tag{32}$$

as is easily verified q.e.d.

This observation enables us to prove:

Theorem 2: The common and unique factors of the factor model (Equations 7 and 8) can always be constructed so as to predict any given criterion w perfectly, including all those that are entirely uncorrelated with the observed scores in Y.

Proof: In view of Equation 16, any $w \in Sp$ can be written uniquely as $w = w_s + w_y$, with $w_s \in Sp(S)$, $w_y \in Sp(Y)$, $w'_s w_y = 0$. If $w_s = \phi$, then $R^2_{w \cdot (X,Z)} = 1$ by Equation 29 for all S_m. Let $w_s \neq \phi$. Then $w^*_s = w_s \sqrt{N/w'_s w_s}$ satisfies $w^{*\prime}_s w^*_s/N = 1$. Since $w_s \in Sp(S) \Rightarrow w^*_s \in Sp(S)$, w^*_s can be used to construct a basis S_m for $Sp(S_m)$ as before by adjoining $m - 1$ nonnull $s_j \in Sp(S)$ that satisfy $s'_j w^*_s = 0$, $s'_i s_j = 0$ for $i \neq j$, and $s'_j s_j/N = 1$. $Sp(S_m)$ for this S_m contains w_s and Sp(Y) contains w_y. Hence, $Sp(Y,S_m) = Sp(Y) \oplus Sp(S_m)$ (by Equation 3) contains $w = w_s + w_y$. Lemma 3 and Equation 27 then give $R^2_{w \cdot (X,Z)} = 1$ q.e.d.

DISCUSSION

When Wilson (1928) discovered factor indeterminacy, the problem received a fair amount of attention (see Steiger &

Schönemann, 1976). With the rise of the "Thurstone school" of factor analysis in the 1940s and 1950s, factor analysts turned away from the study of such theoretical issues and devoted most of their energies to the development of computational algorithms for fitting the factor model, regardless of any defects the model may have. The present problem is whether it still makes any sense to fit the factor model at all, because the model has a built-in indeterminacy that has been ignored over the years.

It has often been said that "factor scores must be estimated because they cannot be computed," or because "the theoretical scores are not available (Tucker, 1971). That such justifications cannot be valid follows from Equation 12, which shows how factor scores can be computed if one wants to. It is, of course, a different question whether they should be computed, and yet another whether they should be "estimated." It is not clear what exactly is meant by the word "estimate" when the criterion for estimation is not uniquely defined. When this situation arises in statistics; for example, in connection with the linear model of deficient rank, estimation is restricted to precisely those linear functions of the indeterminate regression weights that can be uniquely defined, and those are called "estimable."

In addition to the previous questions, what exactly is being estimated, and what exactly is meant by the word "estimate" in this context, we now confront a third question that cuts even more deeply into the common-sense foundation of the factor model. This new question is: Why would anyone want to estimate factors with the absurd properties described in Theorem 1 and Theorem 2? It has been said that factors are superior to components because they "go beyond the test space." Theorems 1 and 2 dramatize the arbitrariness of the increment needed to construct factor scores X,Z in accordance with the factor model (Equations 7 and 8). By Theorem 2, we can always choose it so as to predict any criterion perfectly from X,Z jointly, no matter how this criterion relates to the observed variables. These counterintuitive properties of X,Z are direct consequences of the definition of the factor model at the random variable level. When Spearman (1927) first proposed this model, his definitions seemed plausible because he did not know of the indeterminacy they implied. The present problem is whether they remain plausible in the light of our current knowledge.

REFERENCES

GUTTMAN, L. The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 1955, **8**, 65-81.

KESTELMAN, H. The fundamental equation of factor analysis. *British Journal of Psychology*, Statistical Section, 1952, **5**, 1-6.

PEASE, M. C. III. *Methods of matrix algebra.* New York and London: Academic Press, 1965.

PIAGGIO, H. T. H. The general factor in Spearman's theory of intelligence. *Nature*, 1931, **127**, 56-57.

SCHÖNEMANN, P. H., & STEIGER, J. H. Regression component analysis. *British Journal of Mathematical and Statistical Psychology*, 1976, **29**, 175-189.

SCHÖNEMANN, P. H., & WANG, M. M. Some new results on factor indeterminacy. *Psychometrika*, 1972, **37**, 61-91.

SEARLE, S. R. *Matrix algebra for the biological sciences.* New York: Wiley, 1966.

SEBER, G. A. F. *The linear hypothesis: A general theory.* New York: Hafner, 1966.

SPEARMAN, C. *The abilities of man.* New York: McMillan, 1927.

STEIGER, J. H., & SCHÖNEMANN, P. H. A history of factor indeterminacy. In Shye, S. (Ed.), *Theory construction and data analysis.* San Francisco: Jossey-Bass, in press.

TUCKER, L. Relations of factor score estimates to their use. *Psychometrika*, 1971, **36**, 427-436.

WILSON, E. B. On hierarchical correlation system. *Proceedings of the National Academy of Sciences*, 1928, **14**, 283-291.

(Received for publication April 6, 1978.)