PSYCHOMETRIKA—VOL. 44, NO. 1.
MARCH, 1979

THE RELATIONSHIP BETWEEN EXTERNAL VARIABLES AND COMMON FACTORS

JAMES H. STEIGER

UNIVERSITY OF BRITISH COLUMBIA

A theorem is presented which gives the range of possible correlations between a common factor and an external variable (i.e., a variable not included in the test battery factor analyzed). Analogous expressions for component (and regression component) theory are also derived. Some situations involving external correlations are then discussed which dramatize the theoretical differences between components and common factors.

Key words: factor analysis, factor indeterminacy, external variables.

Introduction

Since 1970, a number of articles have focused on basic theoretical and structural aspects of the common factor model, with particular emphasis on the phenomenon of factor indeterminacy. (See, for example, papers by Schönemann, 1971; Schönemann & Wang, 1972; Meyer, 1973; McDonald, 1974; Mulaik, 1976; Green, 1976; Schönemann & Steiger, 1976; McDonald, 1977; Mulaik & McDonald, 1978; Schönemann & Steiger, 1978a, 1978b; Steiger, Note 1; Williams, 1978.) Factor indeterminacy refers to the fact that, for any set of factor loadings, there exist infinitely many factor random variables, many quite different, which satisfy the definitional requirements of a common factor. Since these random variables are empirically indistinguishable (in the sense that they fit an observed set of data equally well), they represent continuing uncertainty about the identity of a factor.

The present paper explores a somewhat different aspect of factor indeterminacy, by examining the possible relationships between a common factor and an external variable (i.e., one not included in the test battery factor analyzed). The external variable need not necessarily be observed empirically—it might well be purely hypothetical.

For the sake of comparison, some analogous results for component theory and regression component theory [Schönemann & Steiger, 1976] are given. Finally, some implications for the comparison of factors (and components) across different batteries of tests are briefly discussed.

Some Basic Theory

Given some p observed random variables in the random vector y, with $\mathcal{E}(y) = \emptyset$, var $(y) = \mathcal{E}(yy') = R$, the m-factor orthogonal common factor model holds for y if

$$(1) \qquad y = Ax + Uz$$

Support by the National Research Council of Canada (NRC Grant No. A4640) and the University of British Columbia (UBC Humanities and Social Sciences Grant 26-9718) is gratefully acknowledged.

This paper is based in part on the author's Ph.D. dissertation. I am particularly grateful to my dissertation advisor, Dr. Peter H. Schönemann. Thanks also to the editor and the anonymous reviewers, who contributed many helpful comments and suggestions.

Requests for reprints should be sent to Dr. James H. Steiger, Department of Psychology, University of British Columbia, Vancouver, B.C., CANADA V6T 1W5.

0033-3123/79/0300-0093$00.75/0

where x is an $m \times 1$ random vector of m "common factors," z is a $p \times 1$ random vector of p "unique factors," A is a $p \times m$ matrix of constants, of full column rank, called the "common factor pattern," U is a $p \times p$ diagonal, positive definite matrix of coefficients called a "unique factor pattern," and x, z satisfy the conditions

$$\mathcal{E}(xx') = I,\ \mathcal{E}(xz') = \emptyset,\ \mathcal{E}(zz') = I,\ \mathcal{E}(x) = \emptyset,\ \mathcal{E}(z) = \emptyset. \tag{2}$$

It is well-known that (1), (2) may hold for y if and only if one can write

$$R = AA' + U^2. \tag{3}$$

However, for any A, U, satisfying (3) for a given R, there exist infinitely many x, z which satisfy (1), (2). Any and all of these may be written

$$\begin{aligned} x &= A'R^{-1}y + Ps, \\ z &= UR^{-1}y - U^{-1}APs, \end{aligned} \tag{4}$$

where s is any random vector satisfying

$$\mathcal{E}(ys') = \emptyset,\ \mathcal{E}(ss') = I,\ \mathcal{E}(s) = \emptyset, \tag{5}$$

and P is any Gram-factor of $I - A'R^{-1}A$, i.e., $PP' = I - A'R^{-1}A$.

One may rewrite (4) as

$$x = B'y + e = \hat{x} + e, \tag{6}$$

which shows that a factor may be thought of as composed of 2 parts, one a determinate linear combination of the observed variables in y, the other an indeterminate component e.

The Correlation between a Factor and an External Variable

Let w be an external variable (in standard score form), not included in the test battery y. The external correlation between w and a common factor x_j is given by

$$r_{x_jw} = \mathcal{E}(x_jw) = \mathcal{E}(a_j'R^{-1}yw) + \mathcal{E}(p_j'sw) = c_{\hat{x}_jw} + c_{e_jw}. \tag{7}$$

Since e_j is largely arbitrary, (7) suggests immediately that an external correlation is not determinate, and may in fact be free to vary over a wide range of values. In order to develop upper and lower bounds for r_{wx_j} (and the e_j which yield these bounds), we first need the following Lemma.

Lemma

Let

$$y^* = \begin{bmatrix} y \\ w_1 \\ w_2 \end{bmatrix}$$

be a $(p + 2) \times 1$ random vector, composed of random variables in standard score form, whose partitioned correlation matrix may be written

$$R^* = \begin{bmatrix} R & r_1 & r_2 \\ r_1' & 1 & r_{12} \\ r_2' & r_{12} & 1 \end{bmatrix} \tag{8}$$

with obvious notation. Then the correlation r_{12} between w_1 and w_2 must satisfy

$$r_1'R^{-1}r_2 - \sigma_1\sigma_2 \leq r_{12} \leq r_1'R^{-1}r_2 + \sigma_1\sigma_2 \tag{9}$$

where σ_1, σ_2 are the standard errors of w_1 and w_2 about their linear regressions on y.

Proof. By partial correlation theory, one obtains

$$-1 \leq \frac{\sigma_{12.y}}{(\sigma_{11.y}\,\sigma_{22.y})^{1/2}} \leq 1, \tag{10}$$

where

$$\sigma_{12.y} = r_{12} - r_1'R^{-1}r_2,\ \sigma_{11.y} = 1 - r_1'R^{-1}r_1 = \sigma_1^2,\ \sigma_{22.y} = 1 - r_2'R^{-1}r_2 = \sigma_2^2. \tag{11}$$

The result follows immediately by substitution. Q.E.D. (The lemma and proof are essentially due to McDonald, 1977.)

Using the lemma, one may derive upper and lower bounds between a factor and an external variable, as is shown in the following theorem.

Theorem

The correlation r_{x_jw} between common factor x_j and external variable w satisfies the bounds

$$c_{w\hat{x}_j} - (1 - R_{w.y}^2)^{1/2}\sigma_{e_j} \leq r_{wx_j} \leq c_{w\hat{x}_j} + (1 - R_{w.y}^2)^{1/2}\sigma_{e_j}, \tag{12}$$

where $R_{w.y}^2$ is the squared multiple correlation between w and the tests in y, and σ_{e_j}, the standard deviation of e_j, is given by $\sigma_{e_j} = (1 - a_j'R^{-1}a_j)^{1/2}$. The e_j yielding the upper and lower bounds given in (12) are, respectively,

$$\begin{aligned} e_{j(\max)} &= \sigma_{e_j}s_w, \\ e_{j(\min)} &= -e_{j(\max)}, \end{aligned} \tag{13}$$

where s_w is that normalized component of w which is linearly unpredictable from y, i.e., $s_w = (w - c_{wy}R^{-1}y)(1 - R_{w.y}^2)^{1/2}$.

Proof. Let R in (8) satisfy (3). Then let w_1 in the lemma be x_j, the j^{th} common factor. If $w_1 = x_j$, then $r_1 = a_j$. Let w_2 in the lemma be equal to w, the external variable. Equation (12) then follows by substitution, via (7). It is easily verified by substitution that the e_j in (13) yield the bounds in (12). Q.E.D.

Equation (12) shows that the range of external correlations, which is given by $2\sigma_{e_j}(1 - R_{w.y}^2)^{1/2}$, is a function of two influences, the indeterminacy of factor x_j and the "externality" of w, i.e., its linear unpredictability from the tests in y.

External Correlation Theory for Components and Regression Components

External correlation theory for components may be derived more directly than that for common factors, because components are uniquely defined as linear functions of the observed variables; such theory is presented here for the sake of comparison.

Schönemann and Steiger [1976] proposed, as an alternative to the common factor model, a general class of linear data reduction systems, which they call "component decompositions." Specifically, they define m linearly independent variables in x^* as "components" of $p \geq m$ random variables in y if and only if there exists a matrix of defining linear weights B such that

$$x^* = B'y, \tag{14a}$$

and

$$\text{var}(x^*) = B'RB \text{ is positive definite.} \tag{14b}$$

This rather broad definition includes as special cases "principal components," "linear discriminant functions," "partial images," "anti-images," and the various types of "factor score estimators."

A component decomposition of y in terms of m components in x^* is written

$$y = Ax^* + E = AB'y + (I - AB')y. \tag{15}$$

A "regression component decomposition," in the terminology of Schönemann and Steiger [1976], is a component decomposition in which the rows a_j' of the pattern A contain regression weights for predicting the observed y_j from x^*, i.e.,

$$A = RB(B'RB)^{-1}, \tag{16}$$

or equivalently,

$$B = \underline{R^{-1}A(A'R^{-1}}A)^{-1}.$$

From the above definitions, one may easily derive that the correlation between an external variable w and a component $x_j^* = b_j'y$ is determinate and is given by

$$r_{x_j^*w} = b_j'c_{yw}(b_j'Rb_j)^{-1/2}. \tag{17}$$

As an interesting special case of (17), one may write the correlation between an external variable w and a regression estimator $\hat{x}_j = a_jR^{-1}y$ of a common factor x_j as

$$r_{\hat{x}_jw} = a_j'R^{-1}c_{yw}(a_j'R^{-1}a_j)^{-1/2} = c_{\hat{x}_jw}\sigma_{\hat{x}_j}^{-1}. \tag{18}$$

Similarly, it follows that the correlation between a regression component x_j^* and an external variable w is given by

$$r_{x_j^*w} = a_j'R^{-1}c_{yw}(a_j'R^{-1}a_j)^{-1/2}. \tag{19}$$

External Correlational Theory—Some Illustrations and Examples

Some special cases from external correlation theory serve to illustrate possible implications of the theoretical differences between components and common factors. Consider, for example, an external variable w which is orthogonal to all the tests in y. From (17) through (19), one may readily see that such a w must be orthogonal to any of the regression components or components of y, and likewise to the regression estimators of the common factors of y.

On the other hand, (12) shows that, for such a w, the possible correlation r_{wx_j} between w and common factor x_j ranges from $-\sigma_{e_j}$ to $+\sigma_{e_j}$. Hence, if common factor x_j has minimum correlation indeterminacy index of zero (corresponding to a multiple correlation of .71 between x_j and the observed variables in y), it may correlate anywhere from $-.71$ to $+.71$ with w.

Another interesting special case involves the comparison of factors (and components) across different batteries of tests. Suppose, for example, 2 different factor analyses are performed on the same population of subjects. However, the two factor analysts are interested in entirely different aspects of behavior, and their test batteries are linearly uncorrelated, i.e., all the variables in the first experimenter's test battery are uncorrelated with all the variables in the second battery. Suppose further that each factor analysis yields a single common factor with a minimum correlation index of zero. What is the correlation between these two factors?

Since all components (including regression estimators of a common factor and regression components) are linear functions of a set of observed variables, it obviously follows that components from orthogonal test batteries must be orthogonal. However, a similar result does not hold for common factors. Indeed, one can easily derive that, in the above-mentioned situation, the correlation $r_{x_1x_2}$ can vary anywhere from -1 to $+1$.

Specifically, with obvious notation

$$\begin{aligned} x_1 &= a_1'R_1^{-1}y_1 + p_1s_1 = \hat{x}_1 + e_1, \\ x_2 &= \underline{a_2'R_2^{-1}y_2} + p_2s_2 = \hat{x}_2 + e_2, \end{aligned} \tag{20}$$

and since it is given that $\mathcal{E}(y_1 y_2') = \emptyset$, it follows by substitution that

$$r_{x_1 x_2} = \mathcal{E}(x_1 x_2) = \mathcal{E}(\hat{x}_1 e_2) + \mathcal{E}(\hat{x}_2 e_1) + \mathcal{E}(e_1 e_2). \tag{21}$$

If the minimum correlation index for x_1 and x_2 is zero, then var $(\hat{x}_1)$ = var $(\hat{x}_2)$ = var (e_1) = var $(e_2) = \frac{1}{2}$. Since the two test batteries are orthogonal, $\hat{x}_1$ satisfies all the requirements for e_2, and $\hat{x}_2$ satisfies all the requirements for e_1. If one sets $e_1 = \hat{x}_2$, and $e_2 = \hat{x}_1$, then $\mathcal{E}(\hat{x}_1 e_2)$ becomes var $(\hat{x}_1) = \frac{1}{2}$, $\mathcal{E}(\hat{x}_2 e_1)$ becomes var $(\hat{x}_2) = \frac{1}{2}$, and $\mathcal{E}(e_1 e_2)$ becomes $\mathcal{E}(\hat{x}_1 \hat{x}_2)$, which is zero, since the two test batteries are orthogonal. Hence, by setting $e_1 = \hat{x}_2$, and $e_2 = \hat{x}_1$, one obtains $r_{x_1 x_2} = 1$. Similarly, if one sets $e_1 = -\hat{x}_2$, and $e_2 = -\hat{x}_1$, then $r_{x_1 x_2} = -1$. One can always obtain an $r_{x_1 x_2}$ equal to zero by choosing mutually orthogonal e_1 and e_2 which are orthogonal to $\hat{x}_2$ and $\hat{x}_1$, respectively.

Hence, it is apparent that two common factors from *orthogonal* test batteries need not be orthogonal. In the current situation, such factors could be the same factor, the "opposite" factor, or orthogonal to each other, i.e., the correlation between the two factors could range from −1 to +1.

Conclusions

For any set of factor loadings which satisfy the common factor model, there exist infinitely many random variables, some quite different, satisfying the definition of a common factor. Components of a set of observed variables, on the other hand, are uniquely defined as linear combinations of these variables. In many practical data analytic applications, this theoretical distinction may seem of minor consequence. On the other hand, as has been demonstrated here with the theory of external correlations, factor and component models may diverge sharply in some situations. Prospective users of factor analytic and component approaches may wish to keep these distinctions in mind when evaluating the relative merits of the two methods for analyzing multivariate data.

REFERENCE NOTE

1. Steiger, J. H. *The relationship between external variables and indeterminate factors.* Unpublished doctoral dissertation, Purdue University, 1976.

REFERENCES

Green, B. F. On the factor score controversy. *Psychometrika,* 1976, *41,* 263–266.

McDonald, R. P. The measurement of factor indeterminacy. *Psychometrika,* 1974, *39,* 203–221.

McDonald, R. P. The indeterminacy of components and the definition of common factors. *British Journal of Mathematical and Statistical Psychology,* 1977, *30,* 165–176.

Meyer, E. P. On the relationship between ratio of number of variables to number of factors and factorial indeterminacy. *Psychometrika,* 1973, *38,* 375–378.

Mulaik, S. A. Comments on 'the measurement of factorial indeterminacy.' *Psychometrika,* 1976, *41,* 249–262.

Mulaik, S. A., & McDonald, R. P. The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika,* 1978, *43,* 177–192.

Schönemann, P. H. The minimum average correlation between equivalent sets of uncorrelated factors. *Psychometrika,* 1971, *36,* 21–30.

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

Schönemann, P. H., & Steiger, J. H. On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society,* 1978, *12,* 287–290.

Schönemann, P. H., & Wang, M. M. Some new results on factor indeterminacy. *Psychometrika,* 1972, *37,* 61–91.

Steiger J. H. & Schönemann; P. H. A history of factor score indeterminary. In Shye, S. (ed) *Theory construction and data analysis in the behavioral Sciences,* San Francisco: Jossey-Bass, 1978.

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

Manuscript received 8/5/77
First revision received 3/28/78
Final version received 9/20/78