DOCUMENT RESUME

ED 061 255 TM 001 157

AUTHOR	Jackson, Douglas N.; Morf, Martin E.
TITLE	An Empirical Evaluation of Factor Reliability.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-66
PUB DATE	Dec 71
NOTE	34p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Data Analysis; *Evaluation; *Factor Analysis; *Generalization; Hypothesis Testing; Orthogonal Rotation; Personality Tests; *Psychometrics; *Reliability; Response Style (Tests)

ABSTRACT

The psychometric reliability of a factor, defined as its generalizability across samples drawn from the same population of tests, is considered as a necessary precondition for the scientific meaningfulness of factor analytic results. A solution to the problem of generalizability is illustrated empirically on data from a set of tests designed to measure facets of response styles and of personality dimensions. Parallel sets of measures based on personality scales defining each of seven factors were separately factored. Independent sets of component scores derived from the orthogonal least squares fit to the oblique factor pattern matrix were computed, and these component scores were intercorrelated between the two sets, yielding factor reliabilities, whose values ranged from .65 to .85. A corresponding analysis based on scores derived from random binary data yielded nonsignificant factor reliabilities ranging from -.12 to +.07. It was recommended that such a test of factor generalizability be incorporated routinely into factor analytic investigations, particularly those employing Procrustes-type rotations. (Author/AG)

RESEARCH BULLETIN

RB-71-66

AN EMPIRICAL EVALUATION OF FACTOR RELIABILITY

Douglas N. Jackson
University of Western Ontario

and

Martin E. Morf
University of Windsor

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the authors. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
December 1971

AN EMPIRICAL EVALUATION OF FACTOR RELIABILITY

Douglas N. Jackson
University of Western Ontario

Martin E. Morf
University of Windsor

Abstract

The psychometric reliability of a factor, defined as its generalizability across samples drawn from the same population of tests, is considered as a necessary precondition for the scientific meaningfulness of factor analytic results. A solution to the problem of generalizability is illustrated empirically on data from a set of tests designed to measure facets of response styles and of personality dimensions. Parallel sets of measures based on personality scales defining each of seven factors were separately factored. Independent sets of component scores derived from the orthogonal least squares fit to the oblique factor pattern matrix were computed, and these component scores were intercorrelated between the two sets, yielding factor reliabilities, whose values ranged from .65 to .85 ($p < .0001$, for each factor). A corresponding analysis based on scores derived from random binary data yielded nonsignificant factor reliabilities ranging from -.12 to +.07. It was recommended that such a test of factor generalizability be incorporated routinely into factor analytic investigations, particularly those employing Procrustes-type rotations.

AN EMPIRICAL EVALUATION OF FACTOR RELIABILITY[1]

Douglas N. Jackson[2]
University of Western Ontario

Martin E. Morf
University of Windsor

The present study has two major aims: (a) to propose a method of estimating the psychometric reliability or generalizability of a set of factors; and (b) to apply this method to a set of empirical data whose reliability has been questioned in the literature.

The simplest and least controversial, but also the least informative, definition of a factor is: "a set of loadings." Such sets of loadings are obtained from correlation matrices by procedures like the principal factor method (Harman, 1967). If unities are inserted in the main diagonal of the correlation matrix, this procedure yields a mathematically elegant and unique solution. However, some or all of the factors obtained may reflect pseudo-relationships based entirely on chance, while others may account for real relationships in psychologically nonmeaningful ways. Such preliminary factor solutions thus raise two problems: (1) Which factors reflect true common variance as distinguished from error variance? (2) If a set of factors does not merely reflect chance relationships, how may axes be rotated to yield psychologically meaningful factors?

A variety of characteristics permit one to make inferences about the significance of a factor; among these are the size of its eigenvalue, the standard errors of its loadings, and its contribution to the communalities of the variables (Cliff & Hamburger, 1967). Since the principal factor method extracts factors in order of size (as reflected by their eigenvalues), the first question can be rephrased to: How many factors should be retained for rotation and interpretation? The question of which, or how many, factors to retain has been approached in a number of ways. First, rules of thumb have been applied (e.g., Kaiser, 1960). Second, mathematically derived statistics reflecting the significance of factors have been developed (Jöreskog, 1967, 1969; Lawley, 1943; Rao, 1955). Third, real data and random data have been factor analyzed together and only the factors based on real data with eigenvalues greater than the largest eigenvalue of a random factor retained (Horn, 1965).

None of these methods provide unequivocal answers. The answers are contingent on the assumptions of the method that underlies them; on whether it is the subjects or the variables that are treated as a population, and upon which of several properties of factors the emphasis is placed. While preliminary, unrotated factors obtained by some extraction method may reflect a pseudo-meaningful structure attributable to chance, rotated factors are even more likely to do so. Rotation has been identified as occupying a critical role in the possible capitalization upon chance implicit in the emergence of a pseudo-meaningful structure based on random data.[3] This is especially true when the method of rotation is "procrustean" and when the constraint of orthogonality does not interfere with the maximization of loadings in accordance with the investigator's theory (Horn, 1967; Humphreys, Ilgen, McGrath, & Montanelli, 1969). As a result, the factor analytic investigator is faced with the dilemma of blind rotation providing mathematically satisfactory but not necessarily psychologically meaningful solutions (Guilford & Hoepfner, 1969; Saunders, 1960) versus rotation that increases the possibility of a pseudo-meaningful structure.

Some ways around this dilemma are suggested by examination of the criteria to be met by satisfactory factor solutions. Guilford and Hoepfner (1969), for example, suggest that the factors obtained should be amenable to investigation by nonfactor-analytical means, should fit relevant psychological theory, and should be replicable. Kaiser and Caffrey (1965) have elaborated the notion of factor replicability by distinguishing between statistical replication across samples of observations and psychometric replication across samples of variables. Other properties of satisfactory solutions are low standard errors of the loadings and small deviation of the means of the sampling distributions of loadings from the population parameters (Cliff & Pennell, 1967; Pennell, 1968).

Focus on the properties of satisfactory solutions has led to the investigation of the effects on them of various independent variables by means of Monte Carlo simulation studies and to the assessment of factor invariance and other properties of factor solutions in specific analyses of real data. A number of useful rules of thumb, helpful in preventing the emergence of pseudo-meaningful structure, have emerged from the first of these two categories of studies. Horn (1967) and Humphreys *et al*. (1969) factor analyzed randomly generated data and obtained results which suggest that the number of observations and the ratio of variables to factors should be higher, and the ratio of variables to subjects lower, than they are in most studies. Cliff and Pennell (1967) and Pennell (1968) constructed population factor matrices and generated large numbers of sample correlation matrices implied by them. They extracted preliminary factors from the correlation matrices and rotated them to the best least-squares fit with the original population factor matrix (Cliff, 1966). Their results indicate that the larger sample size, communality, and factor size, the greater the consistency, and the smaller the bias, of the loadings.

Given the rules of thumb which follow from studies like these, it is possible to design reasonably sound factor analytical studies. The specific adequacy of each individual study, however, requires separate examination. The present study focuses on this problem. The adequacy of a specific factor analytical solution obtained in an earlier study (Morf & Jackson, in press) and subjected to some intuitive criticism by Block (1971) regarding its supposed chance basis, is tested by examining the invariance of the factors across two parallel subsets of measures included in the original battery.

The psychometric reliability of a factor solution is but one aspect that could be investigated. Factor invariance or replicability over samples of tests is, however, a necessary condition for drawing generalizable conclusions regarding results. Replicability has frequently been described as the "minimum requirement of science." Demonstrating factor replicability, in the pschometric sense, is tantamount to demonstrating that chance alone does not account for the results.

Method

The basic data for this study have been published by Morf and Jackson (in press). Since that report fully described procedures for data collection and substantive interpretation of primary and second-order factors, these issues will not be highlighted here. Briefly, the study was designed to elicit responses from 196 liberal arts undergraduates, 87 males and 109 females, to a personality questionnaire of 560 items comprising 49 nonoverlapping scales relevant to eight factors, four attributable to content (Exhibition, Play, Succorance, and Understanding), and four attributable to response styles (True Responding, Item Endorsement, Desirability, and Adjective Endorsement). A facet design was employed, in which each substantive personality scale was designed to load one content factor and at least one response style factor. Reference tests for response style factors were also included. The results from the Morf and Jackson principal axis analysis and rotation to a clustran (Bentler, 1971) criterion yielded unusually clear support for their hypothesized factors, with virtually all tests appropriately and substantially loading the hypothesized dimensions. The findings did not, however, convince Block (1971), who, citing Horn (1967) and Humphreys et al. (1969), attributed them to chance. Even though the Morf and Jackson study more than met the Humphreys recommendations of at least four tests defining a factor (indeed, there were more than 30 defining each of the major acquiescence dimensions), the fact that, as far as the authors were aware, no satisfactory test of psychometric reliability had been reported provided an impetus for the present investigation.

Method of Analysis

Each of the two sets of scales was separately and independently factored, subjected to an independent analytic patterned rotation, an orthogonal Procrustes rotation, and two separate matrices of component scores computed. These two sets of component scores were then intercorrelated. The correlations between corresponding component scores for a given factor could then be evaluated for statistical significance and for reliability.

The first step in the analytic treatment was to divide the set of variables into two sets. This was done in such a way so as to place an equal number of tests hypothesized to reflect each factor into each set. All test scores were centered at the mean of the test and scaled to have a unit standard deviation. These two matrices of standardized scores were intercorrelated separately within each battery, unities retained in the diagonal, and, because there were seven hypothesized factors, seven principal components factors were extracted from each battery. Because the Adjective Endorsement factor had been defined by only three variables which differed in desirability level, it was not possible to obtain parallel sets of variables to define it. Hence the scales originally defining this factor were dropped from the analysis, as were two additional variables not loading highly on any factor, Infrequency and Sex.

The basic procedure for the orthogonal Procrustes solution was analytically to place axes at the centroid of the respective hypothesized salient test vectors and then to find the orthogonal rotation fitting this oblique solution in a least squares sense. A procedure developed by Horst (1965, pp. 394-397) was employed to transform each of the principal axis factor matrices separately into alignment with their respective hypothesized patterns. The transposed principal axis factor loading matrix is postmultiplied by a binary hypothesis matrix. The matrix product is premultiplied by the reciprocal of the eigenvalues associated with the largest principal components, and the resulting matrix normalized by columns. This matrix thus serves as an oblique transformation matrix, h, which is used to postmultiply the initial principal axis factor loading matrix, A, to yield a primary component pattern matrix, b (cf. Kaiser, 1962), representing loadings of tests on oblique axes.

$$b = Ah \quad . \tag{1}$$

A proof of the rationale on which this method is based is provided by Horst (1965, pp. 411-412).

Bentler's (1968, 1971) clustran criterion, based on a proof due to Gibson (1962), yields an orthonormal rotated factor loading matrix, B, fitting b in a least squares sense. If

$$h = P\delta^{1/2}Q' \quad , \tag{2}$$

then Gibson (1962) has proved that by removing the diagonal matrix, $\delta^{1/2}$, one obtains a matrix T, which will transform the original principal axis factor matrix, A, into B.

$$T = PQ' \quad , \tag{3}$$

and

$$B = AT \quad . \tag{4}$$

Computationally, T may be obtained by first extracting eigenvalues and associated eigenvectors from the minor product moment of h. The eigenvectors will correspond to Q. P may be obtained by premultiplying Q by h and then postmultiplying the result by the diagonal matrix comprising the reciprocal square roots of the eigenvalues.

From matrix T, component scores, Y, may be calculated for each set (cf. Kaiser, 1962), using the original principal axis factor matrix, A, its associated eigenvalues δ_A, and the original standardized data matrix, Z,

$$K = \delta_A^{-1} T \quad , \tag{5}$$

$$L = AK \quad , \tag{6}$$

$$Y = ZL \quad . \tag{7}$$

Because the component scores are based on an orthogonal normal transformation of a principal axis factor matrix, the intercorrelation of these scores within each set will confirm their orthogonality

$$I = Y'Y \quad . \tag{8}$$

Finally, component scores from each set are correlated to yield a matrix of correlations between the estimates of component scores derived from different sets of tests

$$\left[\begin{array}{c|c} I & R_{12} \\ \hline R_{21} & I \end{array}\right] = \left[\begin{array}{c|c} Y_1'Y_1 & Y_1'Y_2 \\ \hline Y_2'Y_1 & Y_2'Y_2 \end{array}\right] \quad . \tag{9}$$

If the factors are listed in the same order within each battery, the diagonal of R_{12} will contain the estimates of psychometric factor reliability. These estimates may be corrected by the Spearman-Brown formula, if an evaluation of the reliability of factors derivable from the entire set is to be made. This procedure may readily be generalized to any number of subsets of tests. Of course, as the number of subsets of tests increases, there may be increasing difficulty in defining, within each subset, tests of sufficient quality to define reliable factors.

Wrigley and Neuhaus (Harman, 1967, pp. 271-272) have defined a coefficient of congruence for measuring the degree of factorial similarity between two sets of tests for the same sample of individuals. This is calculated by dividing the sum of the crossproducts of factor scores of two factors by the geometric mean of the respective variances. If component scores in standard score form as in the present study are employed in the Wrigley and Neuhaus formula, their formula will give the identical results given by (9).

In a similar manner, a parallel analysis was undertaken on two sets of random data. Random binary digits were generated corresponding to each of 560 items answered by 196 subjects and were scored using the same scoring keys employed on the real subjects. Scale scores were divided into two subsets of scales, the same sets employed previously. The identical factor analysis and computation of component scores was undertaken on these subsets, yielding two rotated factor loading matrices and two 196 by seven arrays of component scores for each set. These arrays, when intercorrelated, could be interpreted as the degree of stability manifested by the hypothetical subjects on two independently identified sets of seven latent dimensions. Although there is no reason in factor theory to suppose that random data of this type would yield evidence of stability across independent sets, even if these sets have been rotated to reflect the same factors, and from one point of view this demonstration is trivial, this sort of analysis nevertheless might serve to dispel any lingering doubts. There are, of course, a variety of other points at which randomness might have been introduced into the analysis. Rather than scores based on random binary digits, scores derived from random normal deviates might have been employed, for example, or real data might have been assigned randomly to subjects in one of the sets. It would not be a good use of time to evaluate these alternatives, which, indeed, would demonstrate only that factor scores based on separate sets of random data have a population correlation of zero. One suggested alternative which would clearly be inappropriate would be to use real data, but to base the variable scores on random keys. To the extent that the hypothesized general factors were present in the data, or to the extent that random keys tapped common factors, factor scores derived from such keys would not correlate zero. Such a procedure would merely be a further, although unsystematic, test of our hypotheses. In any case, the analysis of random data is an adjunct to the major analysis, which focuses on the reliability of factor scores derived from real data. Although possibly trivial and gratuitous, it does serve to emphasize the independence of analyses as between the two sets.

Table 1 presents the scales and the hypothesis matrix. The scale labels are described in detail by Morf and Jackson (in press). For present purposes it is sufficient to note the following: (1) in the case of the four letter

Insert Table 1 about here

labels the first letter (E, P, S, U, or H) stands for the content reflected by the scale (E stands for Exhibition, P for Play, S for Succorance, U for Understanding, and H for heterogeneous content), the second letter (A or S) stands for attitude item or self-descriptive format, the third (P or N) stands for positive or negative wording, and the fourth for true or false (T or F) keying; (2) that the first letter of the three letter labels (F) stands for scales consisting of California F Scale items, differing in wording (A stands for absolute wording, R for relative wording) and keying (T stands for true, and F for false keying); (3) that DA and DB stand for the parallel Desirability scales of Forms A and B of the Personality Research Form (Jackson, 1967).

Results

Factor Analytic Results

The factor loadings obtained in the factor analyses of the 23 scales comprising Set A are shown in Table 2. Those for the factor analyses of the 23 scales of Set B are presented in Table 3. These results may be summarized for each set in a parallel manner.

Factor I -- Set A and Set B

All true keyed scales load positively.

All false keyed scales load negatively.

Clearly this factor is associated with the direction of keying of the scales. In both sets, the true and false keyed F scales have extreme loadings.

Factor II -- Set A and Set B

Positively-worded true keyed and negatively-worded false keyed scales load positively.

Negatively-worded true keyed and positively-worded false keyed scales load negatively.

This factor is identified as an item endorsement factor, with the positive pole marked by a tendency to endorse personality scale content, and the negative pole, to deny it. Forty-one of the 44 hypothesized loadings were in the expected direction for this factor. Only three small loadings for Set B were exceptions.

Factor III

Desirability (A)	.69	(Set A)
Desirability (B)	.84	(Set B)

This factor represents a tendency to respond desirably or undesirably.

Factors IV, V, VI, VII

Set A		Set B	
Exhibition SPF	.69	Exhibition SPT	.49
Exhibition SNT	.60	Exhibition SNF	.47
Exhibition APT	.61	Exhibition APF	.60
Exhibition ANF	.41	Exhibition ANT	.38
Play SPT	.65	Play SPF	-.10
Play SNF	.63	Play SNT	.64
Play APF	.41	Play APT	.63
Play ANT	.59	Play ANF	.53
Sentience SPF	.65	Sentience SPT	.55
Sentience SNT	.64	Sentience SNF	.47
Sentience APT	.50	Sentience APF	.60
Sentience ANF	.53	Sentience ANT	.63
Understanding SPT	.71	Understanding SPF	.70
Understanding SNF	.63	Understanding SNT	.53
Understanding APF	.59	Understanding APT	.48
Understanding ANT	.34	Understanding ANF	.59

These factors clearly represent the four content dimensions represented by the respective scale names.

Tables 2 and 3 present the orthogonal rotated factor loading matrix for Set A and for Set B, respectively. Each table first presents the real data analysis and then the random data analysis.

Insert Tables 2 and 3 about here

Evaluations of the goodness of fit of factors based on the random and real data solutions. The random and real data factor solutions can be compared with respect to a number of indices reflecting goodness of fit to the target matrix. Figure 1 compares them in terms of two indices.[4] The first,

Insert Figure 1 about here

taking into account only the direction of relevant loadings, is most appropriate for Factors I and II, which are defined by a large number of variables. The second, taking into account both the direction and relative size of relevant loadings, is more appropriate for the content factors defined by only four variables each. The criterion used here was whether or not the predicted loadings were the highest obtained for the factor.

The Morf and Jackson (in press) study was designed to permit the emergence of Factors I (true responding) and II (item endorsement). Almost all variables analyzed were, therefore, relevant to their definition in the sense that these factors were determined by almost all variables. As Table 1 shows, in the present parallel analyses, 22 of the 23 variables were hypothesized to load in a specific direction on these two factors. Except for the near zero loadings of three variables on Factor II for Set B, all 88 relevant loadings were in the specified direction for the real data. In the case of the random data, however, 59 of these 88 loadings were in the nonpredicted direction.

The single relevant variable for Factor III, the desirability scale, obtained the highest loading in the predicted direction in the two real, but not in the two random, data analyses. In the real data solutions, the four relevant variables defining each content factor obtained the largest loadings in the specified direction, while in the random set only 18 of these 32 variables obtained loadings in the predicted direction and also exceeded the largest irrelevant loading. Thus, although it is difficult to quantify the degree of goodness of fit to the target matrix, these informal comparisons suggest that a clearly better fit was obtained for the real than for the random data.

A chi square calculated on the random data solutions to test whether the surprisingly small numbers of loadings on Factors I and II in the specified direction deviated significantly from the numbers one would expect to load in the specified direction on the basis of chance proved to be significant at the .05 level. The fit to the target matrix of these two factors is thus somewhat worse than one would expect on the basis of chance. This might be surprising at first glance, but becomes clearer when one recognizes that the clustran rotation procedure had fewer constraints operating in fitting the factors defined by few relevant variables, leaving itself very little leeway to fit the two factors defined by many variables. If this interpretation has merit, this finding has a bearing on the conclusions of Humphreys _et al_. regarding the critical role of the number of defining variables for a factor.

Real and random data factor reliabilities. Although the real data solutions seem to fit the target matrix better than the random data solutions, these results alone do not establish the psychometric reliability of the factors. In order to accomplish this, the component scores of each subject on each of the seven factors of the four solutions were computed as outlined in the previous section. Table 4 presents the correlations obtained between

---

Insert Table 4 about here

---

these component scores on corresponding real data and random data factors. Also presented are factor reliabilities, obtained by applying the Spearman-Brown formula to these correlation coefficients. The real and random data are distinguished by these correlations considerably more clearly than by their respective fits to the target matrix. All correlations obtained for the real data are significant at the .0001 level, while none of those obtained for the random data are significant at the .05 level.

Table 5 presents the upper half of the supermatrix of correlations

---

Insert Table 5 about here

---

between component scores in the two sets. The intercorrelations of the variables for Set A form an identity matrix within the limits of rounding error. The same is true for the correlations of component scores within Set B, which are not presented. Correlations between component scores based on corresponding factors from the two sets are presented in the right-hand section of Table 5. It will be noted that factor reliabilities, comprising the minor diagonal, are substantially higher than off-diagonal elements of the heteroset submatrix.

The null hypothesis (cf. Block, 1971) that the Morf-Jackson factors are due to capitalization on chance in the rotation of axes may be rejected with a substantial degree of confidence.

Discussion

The papers by Horn (1967) and by Humphreys et al. (1969) raised profound questions regarding the interpretation of the results from factor analytic studies. They emphasized the point that the apparent meaningfulness of a factor structure was no guarantee that it necessarily reflected true common factor variance among variables. A casual reading of the latter papers might suggest that one might place little confidence in the results of factor analyses conducted on sample sizes of less than a very substantial number. The present investigation focuses on the problem of factor reliability, and suggests a means of interpreting the psychometric reliability of factors, by seeking evidence for stability in factor or component scores across independent sets of tests. The method proposed tends to emphasize the parameters identified by Humphreys et al. as crucial, namely, the ratio of the number of variables to the number of factors, and the number of subjects. Our method requires a sufficient number of variables to permit partition of the set of variables into two subsets, and a sufficient number of subjects to yield statistically significant psychometric factor reliabilities. If these parameters are satisfactorily large, a significance test may be undertaken on the consistency of factor scores, and a decision reached regarding the probable psychometric significance of factors. A rejection of the null hypothesis under these circumstances would imply that the set of factors is replicable across distinct batteries of tests. Thus, it might be concluded that results are not due purely to chance, because, as our analysis of random data illustrated, there is no reason to expect data wholly lacking in psychometric reliability to show stability across separate subsets of tests. Thus, chance effects, as uncovered by Humphreys and Horn, cannot operate to contribute to factor reliability as here defined.

A finding of substantial psychometric factor reliabilities would permit the inference that the factors independently identified by separate batteries tended to reflect the same processes. These two inferences are different in the same sense that evidence for merely a statistically significant reliability might be differentiated from evidence for a substantial reliability for a given test. As with individual tests, one's confidence in the psychometric reliability of a factor might be a linear function of the magnitude of the consistency of factor scores across samples of tests. Thus, statistically significant factor reliabilities might be considered a necessary but not a sufficient criterion for judging the adequacy of a factorial solution. Because of the arbitrariness of rotation, some correlation might be expected between separate solutions so long as axes were oriented in such a way as to be mutually correlated to some degree. A high psychometric factor reliability would imply that each set of tests defining the factor tends to reflect the factor univocally, and, furthermore, that the final rotated solution tends to identify independent samples of subjects' scores along a common dimension.

The rationale for psychometric reliability undertaken here can hardly be considered novel, at least in the context of classical univariate test theory. The notion that a set of items comprise a sample from a hypothetical universe has been incorporated in a number of formulations (see Bock & Wood, 1971, for a review), and implicitly, at least, seems to have been appreciated at least 60 years ago when Spearman (1910) and Brown (1910) published their classic articles on the effect of test length upon reliability. Curiously, this kind of thinking, while occasionally appearing in theoretical articles on factor analysis, has had almost no impact on the practice of factor analysis, where there is often a dearth of good reference tests, frequently insufficient to permit subsets each capable of defining a factor. A well-known kit of reference tests (French, Ekstrom, & Price, 1963), for example, lists only three basic tests per factor. Until the advent of the computer, analyses were sufficiently laborious to discourage parallel replication for the purpose of appraising psychometric generalization. Furthermore, only a minority of investigators (e.g., Horst, 1965) have focused attention on the measurement of individuals based upon factor analysis--in most cases the factor loading matrix is of considerably more interest than the matrix of factor scores. But one would have little confidence in factor analytic results if measures based on one set of reference tests were wholly independent of those based on a second set of putatively parallel tests. Furthermore, the previous objection of undue computational labor in employing factor scores is hardly relevant to the present availability of modern, high-speed computing facilities. Certainly other approaches, such as those which might derive from intraclass correlation under various assumptions, might represent viable alternatives to the one proposed here.

It should be noted that our analysis made no attempt to focus on what Kaiser and Caffrey (1965) termed the statistical reliability of factors. The latter authors suggest that a completely general solution to the problem of factor reliability will probably take into account both the psychometric and the statistical reliability of factors, but note that this would be a rather complex problem. The evidence to date suggests that evaluation of the statistical reliability of factor solutions will have to wrestle not only with the problems posed by Horn and Humphreys et al., but with more recent findings by Nesselroade and Baltes (1970) that attempts at factor matching based upon such optimal criteria as least squares provide a relatively satisfactory fit for random data. In addition, further work will have to be undertaken more clearly to define the concept of a subject population because, as has clearly been demonstrated (Tucker, 1966), the factorial structure describing different types of subjects may vary both in terms of number and nature of the obtained factors, and in terms of factor correlations.

Summary and Conclusions

1. The psychometric reliability of a factor, defined as its generalizability across the population of tests hypothesized to measure the factor, may be appraised empirically by correlating factor scores based on independently analyzed parallel subsets of tests.

2. Product-moment correlations so obtained may be tested for statistical significance and may be corrected by the Spearman-Brown formula to yield an index of reliability.

3. When psychometric factor reliability analysis was applied to factor scores generated by tests hypothesized to reflect three response style and four content factors, all reliability coefficients were significant at the .0001 level, thus failing to support a conjecture made by Block that these factors were due to capitalization on chance.

4. When a similar analysis was applied to random data, none of the psychometric factor reliabilities departed significantly from zero.

5. Comparison of the results from real and random data analyses supported the critical role of the number of tests defining each factor.

6. An evaluation of the psychometric reliability of a factor should be undertaken routinely in factor studies, particularly in those employing rotation to optimize fit to a set of hypotheses.

References

Bentler, P. M. A new matrix for the assessment of factor contributions. Multivariate Behavioral Research, 1968, 3, 489-494.

Bentler, P. M. Clustran: A program for oblique transformation. Behavioral Science, 1971, 16, 183-185.

Block, J. On further conjectures regarding acquiescence. Psychological Bulletin, 1971, 76, 205-210.

Bock, R. D., & Wood, R. Test theory. In P. Mussen and M. Rosenzweig (Eds.), Annual Review of Psychology, 1971, 22, 193-224.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Cliff, N. Orthogonal rotation to congruence. Psychometrika, 1966, 30, 33-42.

Cliff, N., & Hamburger, C. D. The study of sampling errors in factor analysis by means of artificial experiments. Psychological Bulletin, 1967, 68, 430-445.

Cliff, N., & Pennell, R. The influence of communality, factor strength, and loading size on the sampling characteristics of factor loadings. Psychometrika, 1967, 32, 309-326.

French, J. W., Ekstrom, R. B., & Price, L. A. (Eds.), Kit of Reference Tests for Cognitive Factors. Princeton, N. J.: Educational Testing Service, 1963.

Gibson, W. On the least-squares orthogonalization of an oblique transformation. Psychometrika, 1962, 27, 193-196.

Guilford, J. P., & Hoepfner, R. Comparison of varimax rotations with rotations to theoretical targets. Educational and Psychological Measurement, 1969, 29, 3-23.

Harman, H. Modern factor analysis. (2nd ed.) Chicago: University of Chicago Press, 1967.

Horn, J. L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-185.

Horn, J. L. On subjectivity in factor analysis. Educational and Psychological Measurement, 1967, 27, 811-820.

Horst, P. Factor analysis of data matrices. New York, New York: Holt, Rinehart, & Winston, 1965.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. Capitalization on chance in rotation of factors. Educational and Psychological Measurement, 1969, 29, 259-271.

Jackson, D. N. Manual for the Personality Research Form. Goshen, New York: Research Psychologists Press, 1967.

Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967, 32, 443-482.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Kaiser, H. F. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 1960, 20, 141-151.

Kaiser, H. F. Formulas for component scores. Psychometrika, 1962, 27, 83-87.

Kaiser, H. F., & Caffrey, J. Alpha factor analysis. Psychometrika, 1965, 30, 1-14.

Lawley, D. N. The application of the maximum likelihood method to factor analysis. British Journal of Psychology, 1943, 33, 172-175.

Morf, M., & Jackson, D. N. An analysis of two response styles: True responding and item endorsement. Educational and Psychological Measurement, 1972, in press.

Nesselroade, J. R., & Baltes, P. B. On a dilemma of comparative factor analysis: A study of factor matching based on random data. Educational and Psychological Measurement, 1970, 30, 935-948.

Pennell, R. The influence of communality and N on the sampling distributions of factor loadings. Psychometrika, 1968, 33, 423-440.

Rao, C. R. Estimation and tests of significance in factor analysis. Psychometrika, 1955, 20, 93-111.

Saunders, D. R. A computer program to find the best-fitting orthogonal factors for a given hypothesis. Psychometrika, 1960, 25, 199-203.

Spearman, C. Correlation calculated with faulty data. British Journal of Psychology, 1910, 3, 271-295.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-311.

Footnotes

[1]Reprint requests should be directed to Douglas N. Jackson, Department of Psychology, University of Western Ontario, London 72, Ontario, CANADA. The authors appreciate the helpful comments provided by Harry H. Harman, Walter Kristof, Ingram Olkin, and Roger Pennell regarding the reporting of this study. Thanks are also due to William Krane, who assisted in generating the random data. Supported in part by Research Grant No. 397 from the Ontario Mental Health Foundation, and a Special Research Fellowship from the National Institute of Mental Health, U.S. Public Health Service to Douglas N. Jackson and a Research Grant to Martin E. Morf from Canada Council.

[2]This study was completed while Douglas N. Jackson was a Visiting Scholar, Division of Psychological Studies, Educational Testing Service, Princeton, New Jersey.

[3]Although, strictly speaking, the rotation of axes does not occur in certain modern approaches to factor analysis, such as the one proposed by Jöreskog (1969), the analogous problem of the fitting of parameters on the basis of observed data remains.

[4]Factor III, the Desirability Factor, with only a single hypothesized high loading, is not evaluated in Figure 1.

Table 1

Hypothesis Matrices for Parallel Sets of Personality Scales

	Set A	Set B	I		II		III	IV	V	VI	VII
			Set A	Set B	Set A	Set B					
1.	ESPF	ESPT	-1	1	-1	1	0	1	0	0	0
2.	ESNT	ESNF	1	-1	-1	1	0	1	0	0	0
3.	EAPT	EAPF	1	-1	1	-1	0	1	0	0	0
4.	EANF	EANT	-1	1	1	-1	0	1	0	0	0
5.	PSPT	PSPF	1	-1	1	-1	0	0	1	0	0
6.	PSNF	PSNT	-1	1	1	-1	0	0	1	0	0
7.	PAPF	PAPT	-1	1	-1	1	0	0	1	0	0
8.	PANT	PANF	1	-1	-1	1	0	0	1	0	0
9.	SSPF	SSPT	-1	1	-1	1	0	0	0	1	0
10.	SSNT	SSNF	1	-1	-1	1	0	0	0	1	0
11.	SAPT	SAPF	1	-1	1	-1	0	0	0	1	0
12.	SANF	SANT	-1	1	1	-1	0	0	0	1	0
13.	USPT	USPF	1	-1	1	-1	0	0	0	0	1
14.	USNF	USNT	-1	1	1	-1	0	0	0	0	1
15.	UAPF	UAPT	-1	1	-1	1	0	0	0	0	1
16.	UANT	UANF	1	-1	-1	1	0	0	0	0	1
17.	HMPF	HMPT	-1	1	-1	1	0	0	0	0	0
18.	HMNT	HMNF	1	-1	-1	1	0	0	0	0	0
19.	HPPT	HPPF	1	-1	1	-1	0	0	0	0	0
20.	HPNF	HPNT	-1	1	1	-1	0	0	0	0	0
21.	DB	DA	0	0	0	0	1	0	0	0	0
22.	FRF	FRT	-1	1	-1	1	0	0	0	0	0
23.	FAT	FAF	1	-1	1	-1	0	0	0	0	0

Note:--A positive unity causes the rotation to seek to yield a positive loading for the variable in question, negative unity seeks to yield a negative loading and a zero leaves the loading for that variable unconstrained.

Table 2

Rotated Factor Loading Matrix for 23 Personality Variables: Set A

	Real Data								Random Data							
Factor	I	II	III	IV	V	VI	VII	h^2	I	II	III	IV	V	VI	VII	h^2
1. ESPF	-11	-27	23	69	02	13	00	.63	-29	18	-24	42	-04	15	-12	.39
2. ESNT	21	-25	22	60	17	26	09	.62	-07	18	01	41	14	-06	31	.32
3. EAPT	18	17	-27	61	03	-13	-19	.56	-16	-38	-19	08	-33	28	07	.40
4. EANF	-29	23	18	41	13	05	10	.37	08	-17	29	39	12	-05	-02	.29
5. PSPT	34	13	13	-10	65	23	12	.65	01	09	-25	24	41	26	-12	.38
6. PSNF	-23	27	-04	08	63	14	11	.56	57	04	01	10	33	-02	21	.49
7. PAPF	-45	-04	-25	22	41	-23	-23	.59	16	33	18	-17	34	27	10	.39
8. PANT	29	-13	-28	05	59	-12	-12	.56	-37	-14	-17	-26	43	-22	-15	.51
9. SSPF	-25	-22	02	02	05	65	19	.57	25	13	08	03	-08	63	-06	.49
10. SSNT	14	-04	21	05	05	64	00	.48	01	12	01	22	-18	-15	-49	.36
11. SAPT	35	19	-24	15	-02	50	-15	.51	-18	-08	-08	21	22	37	42	.45
12. SANF	-32	09	00	01	-02	53	-26	.46	01	-37	24	-21	27	37	-03	.45
13. USPT	19	35	-05	03	-02	-05	71	.67	20	05	10	05	-43	04	52	.51
14. USNF	-30	20	00	05	01	-19	63	.51	05	-33	-06	29	14	-27	27	.36
15. UAPF	-40	-01	-09	-06	-16	08	59	.55	-09	00	-11	-39	08	-11	34	.31
16. UANT	30	-37	-09	-06	06	-10	34	.37	-29	-06	05	27	25	04	45	.43

Table 2 (cont'd)

Rotated Factor Loading Matrix for 23 Personality Variables: Set A

Factor	Real Data I	II	III	IV	V	VI	VII	h^2	Random Data I	II	III	IV	V	VI	VII	h^2
17. HMPF	-33	-51	53	06	-11	-16	-04	.69	16	15	28	10	14	-41	-01	.32
18. HMNT	41	-28	55	-01	01	-33	02	.66	-06	60	07	07	02	-16	-16	.43
19. HPPT	50	58	29	08	01	-07	02	.68	-36	04	42	27	-30	05	-03	.47
20. HPNF	-38	55	34	-04	23	00	05	.62	39	-43	-02	-08	-08	-25	-01	.41
21. DB	-05	-12	69	09	-12	-01	-07	.52	20	-06	-08	14	-19	-14	-04	.13
22. FRF	-37	-23	-10	-13	36	21	-06	.39	00	10	-43	-30	-33	13	20	.45
23. FAT	50	04	07	-23	02	17	-31	.43	-55	08	11	-32	-05	-23	05	.48
Sum of Squares of Factor Loadings	2.37	1.77	1.76	1.57	1.63	1.87	1.75		1.49	1.25	.86	1.42	1.42	1.43	1.36	

Note:--Rotation for data on Tables 2 and 3 was by an orthogonal Procrustes criterion.

Table 3

Rotated Factor Loading Matrix for 23 Personality Variables: Set B

Factor	Real Data I	II	III	IV	V	VI	VII	h^2	Random Data I	II	III	IV	V	VI	VII	h^2
1. ESPT	17	19	-19	49	11	-08	-15	.38	-42	-35	33	-20	06	28	-01	.53
2. ESNF	-34	29	03	47	34	-01	07	.54	17	12	-02	46	-33	-08	-08	.38
3. EAPF	-41	-13	16	60	-03	27	-12	.66	02	17	12	44	02	-31	08	.34
4. EANT	47	-24	-10	38	-11	18	-01	.48	14	13	03	33	15	45	-06	.37
5. PSPF	-33	-03	-07	09	-10	-04	-25	.68	26	03	-03	20	05	15	22	.18
6. PSNT	29	12	-05	10	64	-11	-22	.58	33	33	00	-35	10	-48	-10	.59
7. PAPT	25	25	10	00	63	18	-07	.57	-40	34	-07	-04	-09	-08	25	.36
8. PANF	-33	-03	-16	12	53	25	08	.50	25	00	26	-07	37	-07	-42	.45
9. SSPT	01	29	-27	23	05	55	-25	.58	-25	-39	07	14	-31	05	06	.34
10. SSNF	-13	47	-18	09	10	47	-15	.53	41	-01	-20	07	-17	47	12	.48
11. SAPF	-38	-20	27	16	-07	60	09	.66	-51	07	-29	06	17	26	06	.45
12. SANT	25	-20	02	-06	26	63	14	.59	10	30	57	02	08	22	-23	.53
13. USPF	-25	05	-13	15	05	-09	70	.61	-09	24	44	-31	-04	-24	34	.53
14. USNT	15	-22	-02	17	-20	-17	53	.45	21	-08	05	23	-28	-16	24	.27
15. UAPT	31	19	19	-20	-27	-03	48	.51	42	35	-15	-16	11	-03	39	.51
16. UANF	-41	04	18	-23	00	16	59	.63	15	-24	35	-10	04	18	03	.25

Table 3 (cont'd)

Rotated Factor Loading Matrix for 23 Personality Variables: Set B

	Real Data								Random Data							
Factor	I	II	III	IV	V	VI	VII	h^2	I	II	III	IV	V	VI	VII	h^2
17. HMPT	33	62	-21	23	-05	04	-02	.59	01	-51	11	-04	-21	13	25	.40
18. HMNF	-40	66	-08	11	01	12	-18	.66	25	-37	-06	-19	-09	-44	-19	.48
19. HPPF	-44	-24	07	-28	-24	-08	-44	.59	31	-01	-04	-61	-22	34	09	.64
20. HPNT	31	-34	-41	-23	03	-08	-34	.56	-17	18	-51	05	-01	21	-27	.44
21 DA	-12	-08	84	-02	-03	-03	05	.73	10	-06	51	19	-04	11	13	.34
22. FRT	45	17	-13	-43	09	21	-27	.56	15	15	00	01	-29	05	-40	.29
23. FAF	-48	-01	-39	-18	-04	30	02	.51	-02	51	21	-16	-40	08	15	.52
Sum of Squares of Factor Loadings	2.48	1.78	1.48	1.67	1.99	1.73	2.04		1.60	1.61	1.57	1.41	.89	1.49	1.09	

Table 4

Factor Reliabilities for the Parallel Real Data Analysis and the Parallel Random Data Analyses

Factor	Real Data Analysis Uncorrected	Real Data Analysis Corrected[a]	Random Data Analysis
I	.74	.85	04
II	.59	.70	04
III	.49	.65	04
IV	.50	.66	01
V	.56	.72	07
VI	.53	.70	-11
VII	.56	.72	12

[a]Spearman-Brown formula.

Table 5

Correlations of Scores for Seven Factors from Each of
Two Sets of Principal Components Factor Analyses.
Minor Diagonals Are Factor Reliabilities

(N = 196)

		Set A							Set B						
	Factor	1A	2A	3A	4A	5A	6A	7A	1B	2B	3B	4B	5B	6B	7B
Set A	1A	1.00	.00	.00	-.00	.01	.01	-.00	.73	.16	.06	.08	-.05	-.07	-.16
	2A	.00	1.00	.00	.01	-.00	-.00	-.00	.09	.54	-.05	.09	-.01	.05	.18
	3A	.00	.00	1.00	.00	.00	-.01	.01	-.05	-.18	.49	-.04	.02	.06	.10
	4A	-.00	.01	.00	1.00	.00	-.00	-.00	-.02	-.05	.01	.50	.12	.03	-.08
	5A	.01	-.00	.00	.00	1.00	.00	-.00	-.09	.04	-.09	.12	.56	-.02	-.17
	6A	.01	-.00	-.01	-.00	.00	1.00	-.01	-.05	.23	-.06	.18	-.02	.53	-.13
	7A	-.00	-.00	.01	-.00	-.00	-.01	1.00	-.15	.05	.12	.15	-.07	-.05	.56

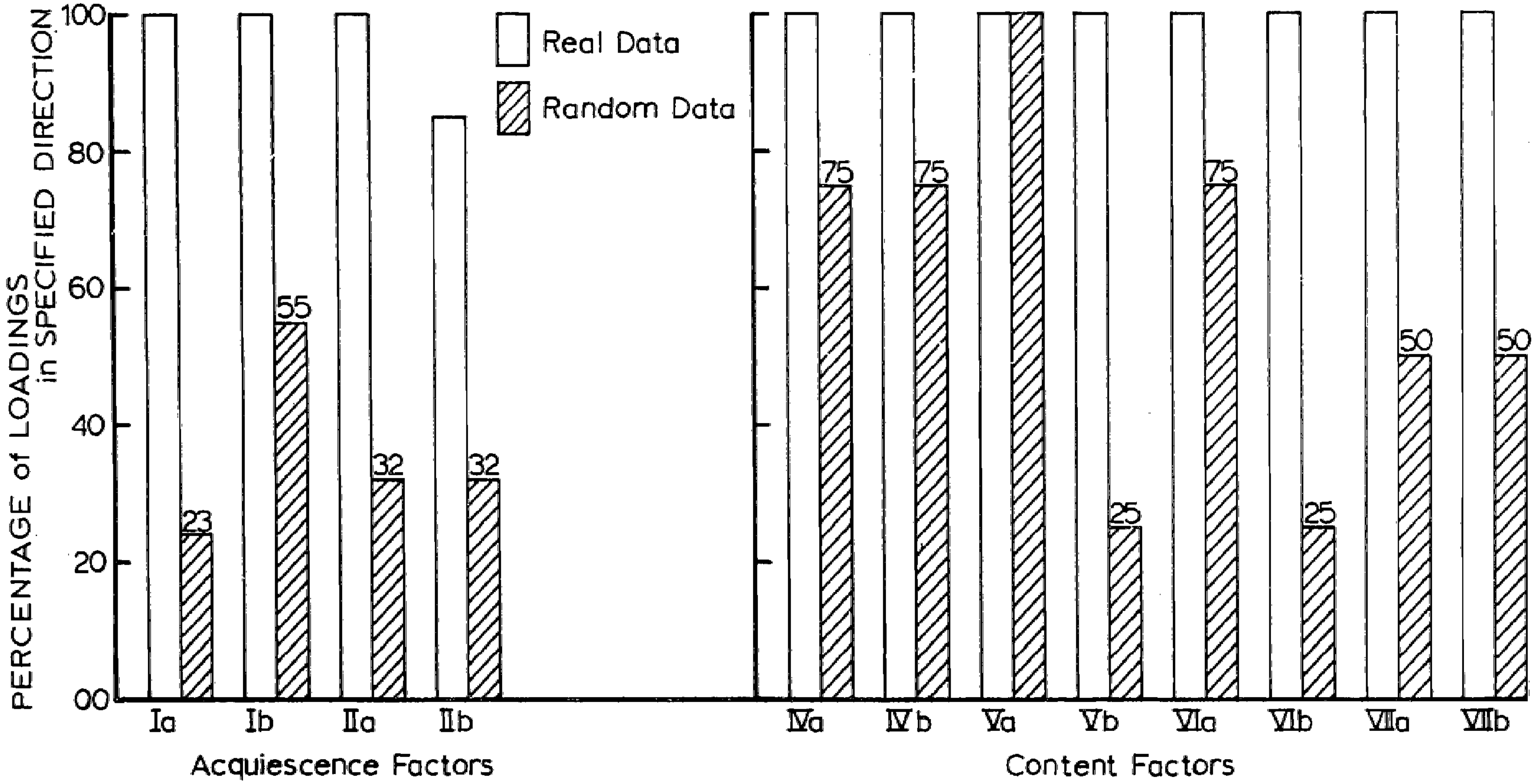

Fig. 1. Comparison of factor analyses of real and random data in terms of the percentage of factor loadings in specified direction for acquiescence and content factors.