PSYCHOMETRIKA—VOL. 38, NO. 1
MARCH, 1973

# A RELIABILITY COEFFICIENT FOR MAXIMUM LIKELIHOOD FACTOR ANALYSIS*

LEDYARD R TUCKER
AND
CHARLES LEWIS

UNIVERSITY OF ILLINOIS

Maximum likelihood factor analysis provides an effective method for estimation of factor matrices and a useful test statistic in the likelihood ratio for rejection of overly simple factor models. A reliability coefficient is proposed to indicate quality of representation of interrelations among attributes in a battery by a maximum likelihood factor analysis. Usually, for a large sample of individuals or objects, the likelihood ratio statistic could indicate that an otherwise acceptable factor model does not exactly represent the interrelations among the attributes for a population. The reliability coefficient could indicate a very close representation in this case and be a better indication as to whether to accept or reject the factor solution.

Maximum likelihood factor analysis offers effective procedures for statistical estimation of factor matrices and for statistical tests as to whether a factor analysis model represents the interrelations of attributes in a battery for a population of objects or individuals. In practical use of these methods, however, there is a problem in judging the quality of a factor analytic study. While the factor analytic approach may be quite profitable in establishing latent traits which account for essential interrelations among observations in a domain of phenomena, the factor analytic model involving a limited number of common factors almost surely will not represent exactly the phenomena for a population of objects. This proposition raises questions as to the use of the likelihood ratio test associated with maximum likelihood factor analysis. When a study is conducted with a very large sample of individuals the statistical test may indicate that the factor analytic model with a scientifically desirable number of common factors would not represent data for a population of objects. In these cases a measure of goodness of fit of the model to the phenomena is needed.

Lawley's [1940] initial solution of maximum likelihood factor analysis appeared to offer an elegant procedure for estimation of factor matrices

* This research was supported by the Personnel and Training Research Programs Office of the Office of Naval Research under contract US NAVY/00014-67-A-0305-0003. Critical review of the development and suggestions by Richard Montanelli were most helpful.

and the associated likelihood ratio statistic seemed to promise a solution for the long standing number of factors problem. Due to the extensive calculations involved these procedures were little used but were discussed in the theoretic literature. Rao [1955] derived canonical factor analysis and demonstrated the equivalence to maximum likelihood factor analysis. Lord [1956] provided the first application of maximum likelihood factor analysis to a large battery of measures using the Whirlwind computer at the Massachusetts Institute of Technology. With the further developments of high speed digital computers and of effective computer programs by Jöreskog [1967] maximum likelihood factor analysis has become quite feasible for application. Experience with maximum likelihood factor analysis has been developing with these applications. This experience indicates a dilemma in the application of the likelihood ratio statistic to decisions concerning the factor analyses.

The problem with the use of the likelihood ratio statistic, or any other similar statistic, involves the form of the decision procedure. The statistical hypothesis is that the factor analytic model with a specified number of common factors applies strictly for a population of objects. Rejection of this hypothesis most surely will occur for a very large sample of objects at any usual level of significance. This rejects the scientific hypothesis. A reversal as to role of the statistical hypothesis and alternative hypothesis has occurred from the common use of decision procedures in scientific investigations for which the scientific hypothesis is the alternative hypothesis and is to be accepted when the statistical hypothesis has been rejected.

Consider a case involving a well developed battery of attribute measures such that with an extremely large sample of objects there would be common agreement that $r$ important common factors are involved and that any further factors are trivial and uninteresting. Tucker, Koopman, and Linn [1969] proposed a system for producing correlation matrices based on the conception of a major factor domain and a minor factor domain to simulate observed correlation matrices. Browne [1969] pointed out that any correlation matrix may be perfectly reproduced from a factor matrix when a large enough number of factors was permitted and that the different methods of factoring differed in the definition of a limited number of factors accepted and in the factors not accepted. The failure of the factor analytic model with a limited number of common factors to reproduce the matrix of correlations or covariances can be transformed to the existence of additional common factors which are to be rejected. In the case being considered there is common agreement as to the number of major factors in the common factor space and that the remaining common factors derive from a minor factor space and are to be discarded. The likelihood ratio and usual decision process would be quite appropriate in rejecting fewer than $r$ factors. The problem is that this statistic and decision procedure probably would reject also $r$

common factors. This would occur with a large enough sample of objects even for very trivial and meaningless minor factors. For example, Harman [1967, see page 229] states in reference to the maximum likelihood solutions for his 8 physical variables example:

> This example illustrates the general principle that one tends to underestimate the number of factors that are statistically significant. For twenty years, two factors had been considered adequate, but statistically two factors do *not* adequately account for the observed correlations based on a random sample of 305 girls. However, the third factor (whose total contribution to the variance ranges from 2 per cent to 5 per cent for the different solutions) has little "practical significance," and certainly a fourth factor would have no practical value.

As shown in Table 1 both the 2 factor and 3 factor models would be rejected at high levels of significance, $p$ less than .001 and .01, respectively.

This situation is quite analogous to that of paired comparison scaling for which Mosteller [1951] provided a statistical test concerned with whether the model might or might not be rejected. Gulliksen and Tukey [1958] provided a reliability type coefficient for measuring goodness of fit of the model to data. They contrasted two examples: Mosteller's baseball data, for which the significance test did not reject the model while the reliability was low, with quality of handwriting data, for which the significance test indicated a decision to reject the model while the reliability was high. This contrast was due in part to quite different numbers of cases on which each proportion used was based: 22 for the baseball data versus 200 for the total sample for the handwriting data. An analogous reliability type coefficient is needed for factor analysis.

In developing a reliability coefficient for maximum likelihood factor analysis an asymptotic identity developed by Lawley [1940] for large $N$ and several analogies are used. ($N$ is the number of objects in a sample.) Jöreskog [1967] utilizes a derived function, $F_m$ for $m$ common factors, which is minimized to maximize the likelihood function. He indicates that $(N - 1)F_m$ is $(-2)$ times the logarithm of the likelihood ratio and that $n_m' F_m$ is ap-

TABLE 1

Maximum Likelihood Factoring Reliability
Harman's 8 Physical Measures Example
N = 305 , n = 8

| Number of Factors | F | df | p | M | $M^{1/2}$ | ρ |
|---|---|---|---|---|---|---|
| 0 | 6.941 | 28 | *** | .2479 | .498 | --- |
| 1 | 2.039 | 20 | *** | .1020 | .319 | .597 |
| 2 | .253 | 13 | *** | .0195 | .140 | .934 |
| 3 | .076 | 8 | ** | .0095 | .097 | .975 |
| 4 | .015 | 3 | .23 | .0048 | .069 | .994 |

** p < .01
*** p < .001

proximately a $\chi^2$ variable, where

$$n'_m = N - 1 - \tfrac{1}{6}(2n + 5) - \tfrac{2}{3}m.$$

Let $C$ be the observed covariance matrix for a battery of $n$ attributes, $\hat{A}_m$ be the estimated factor matrix for $m$ common factors, $\hat{U}_m$ be the estimated unique factor loadings for an $m$ common factor model, and $G_m$ be an $n \times n$, symmetric matrix defined by

$$G_m = \hat{U}_m^{-1}(C - \hat{A}_m\hat{A}'_m)\hat{U}_m^{-1}. \tag{1}$$

In the maximum likelihood solution

$$\hat{U}_m^2 = \text{Diag }(C - \hat{A}_m\hat{A}'_m). \tag{2}$$

Consequently

$$\text{Diag }(G_m) = I. \tag{3}$$

The matrix $G_m$ may be considered to contain the partial intercorrelations of the attributes partialling out the estimated common factors. Lawley's identity may be combined with Jöreskog's function $F_m$ to yield

$$F_m \doteq \sum_{j=1}^{n-1} \sum_{j'=j+1}^{n} g_{mjj'}^2 \tag{4}$$

where $g_{mjj'}$ are the entries in $G_m$. Thus, $F_m$ may be considered as approximately the sum of squares of the partial correlations on one side of the diagonal in $G_m$.

The preceding suggests an analogy with components of variation in analysis of variance. In this interpretation let $M_m$ be a mean square corresponding to $F_m$.

$$M_m = F_m/df_m \tag{5}$$

where $df_m$ is the degrees of freedom associated with $F_m$ in the maximum likelihood solution. For variance components let $\alpha_m$ be a variance associated with a model having $m$ common factors, $\delta_m$ be a variance representing the deviation of the model from actuality, and $\epsilon_m$ be a variance associated with sampling. For this component of variance model consider

$$\mathcal{E}(M_0) = \alpha_m + \delta_m + \epsilon_m, \tag{6}$$

$$\mathcal{E}(M_m) = \delta_m + \epsilon_m, \tag{7}$$

where $M_0$ is the mean square for a model having zero common factors. A value for $\epsilon_m$ may be obtained for the case when $\delta_m$ is zero, that is when the model fits exactly for a population of objects. Then $n'_mF_m$ is approximately a $\chi^2$ random variable with $df_m$ degrees of freedom and has an expected value of $df_m$. From this and (5) the expected value of $n'_mM_m$ is unity and the expected value of $M_m$ is $1/n'_m$ . Using this result as a value for $\epsilon_m$ (6) and (7) become

$$\mathcal{E}(M_0) = \alpha_m + \delta_m + \frac{1}{n'_m}, \tag{8}$$

$$\mathcal{E}(M_m) = \delta_m + \frac{1}{n'_m}. \tag{9}$$

A reliability coefficient may be defined by

$$\rho_m = \frac{\alpha_m}{\alpha_m + \delta_m}. \tag{10}$$

This is analogous to an intraclass correlation. It represents a ratio of the amount of variance associated with the model to total variance. An estimate may be obtained by substitution of observed values of $M_0$ and $M_m$ for the expected values in (8) and (9). Then

$$\rho_m \doteq \frac{M_0 - M_m}{M_0 - 1/n'_m}. \tag{11}$$

This reliability coefficient may be interpreted as indicating how well a factor model with $m$ common factors represents the covariances among the attributes for a population of objects. Lack of fit would indicate that the relations among the attributes are more complex than can be represented by $m$ common factors.

Several examples are given in the tables for application of the reliability coefficient to correlation matrices taken from the literature. Table 1 presents results for Harman's [1967] 8 physical measures example. These measures were selected from a battery of 17 measures used by Mullen [1939] and the correlations based on an $N$ of 305 were taken from her study. As indicated previously, a two common factor model is rejected by the likelihood ratio statistic at a significance level of .001. The reliability, $\rho$, was .934 for the two factor solution which has been accepted for years. A three common factor model may be rejected according to the likelihood statistic for which the $p$ was less than .01. The reliability had risen to .975. This three common factor structure has two very highly correlated factors after rotation; one for height and length of lower leg, and one for arm span and length of forearm. These four measures loaded on a single rotated factor in the two factor solution; thus, the three factor solution is providing a differentiation between length of leg bones and length of arm bones which may be of scientific interest. The two and three factor solutions had similar factors for the last four measures involving weight and girths. A four common factor model cannot be rejected by the likelihood ratio statistic and the reliability has risen to .994. However, the four factor solution does not add a meaningful factor in our judgment to the rotated solution for the three factor solution. We suggest that the three factor solution should be accepted in that the reliability is high and in that the four factor solution does not add a meaningful factor beyond the three in the three factor solution. This suggestion disregards the likelihood ratio result which indicates that the three factor solution would not be exact for a population of girls.

The square roots of the $M$'s for the various numbers of factors are listed also in Table 1. These values may be interpreted as root mean squares of the partial correlations among the attributes after the given number of factors have been extracted. One point to remember is that the approximate sum of squares, $F$, has been divided by the number of degrees of freedom remaining rather than by the number of partial correlations. Thus, even though $F$ dropped from .076 for 3 factors to .015 for 4 factors, $M$ dropped only from .0095 to .0048 since the degrees of freedom decreased markedly from 8 to 3. The corresponding decrease in the root mean square was only from .097 to .069. With consideration of this point, the values of $M^{1/2}$ may be considered as measures of the sizes of the partial correlations. Both .097 and .069 are quite small.

A second example is presented in Table 2. This is the nine test combined battery selected by Tucker [1958] from the larger battery studied by Thurstone and Thurstone [1941]. Tucker used this battery to illustrate his interbattery factor analytic method and selected it to have two common factors. The two factor solution may be rejected at a high level of significance, $p < .001$; however, the two factor solution has a reliability of .982 indicating a very good fit of the model to the interrelations among the scores on these eleven tests. The three factor solution, which may not be rejected and for which the reliability is 1.000, does not add a third meaningful factor. Loadings on the third dimension are moderately small. In consequence, the two factor solution which does not fit the data by the likelihood ratio test appears justified to represent the relations among the scores on these nine tests.

Results for Harman's 24 psychological test example, which he obtained

TABLE 2

Maximum Likelihood Factoring Reliability
Selected Battery from Thurstone & Thurstone
N = 710 , n = 9

| Number of Factors | F | df | p | M | $M^{1/2}$ | ρ |
|---|---|---|---|---|---|---|
| 0 | 4.490 | 36 | *** | .1247 | .353 | --- |
| 1 | 1.308 | 27 | *** | .0484 | .220 | .619 |
| 2 | .071 | 19 | *** | .0037 | .061 | .982 |
| 3 | .017 | 12 | .46 | .0014 | .037 | 1.000 |

*** p < .001

TABLE 3

Maximum Likelihood Factoring Reliability
Harman's 24 Psychological Tests Example
N = 145 , n = 24

| Number of Factors | F | df | p | M | $M^{1/2}$ | ρ |
|---|---|---|---|---|---|---|
| 0 | 11.437 | 276 | *** | .0414 | .204 | --- |
| 1 | 4.631 | 252 | *** | .0184 | .136 | .678 |
| 2 | 3.140 | 229 | *** | .0137 | .117 | .816 |
| 3 | 2.220 | 207 | *** | .0107 | .104 | .905 |
| 4 | 1.711 | 186 | .02 | .0092 | .096 | .951 |
| 5 | 1.417 | 166 | .13 | .0085 | .092 | .973 |

*** $p < .001$

from a study by Holzinger and Swineford [1939], are shown in Table 3. Four factor solutions have been used in past analyses. This size model may be rejected by the likelihood ratio test at a value of $p = .02$. Reliability of the four factor model is relatively high at .951. Rotation of axes for the five factor solution presents some problems whereas the four factor solution has four rather nice rotated factors. Consequently, the four factor solution appears to be appropriate.

The number of individuals in the 24 psychological test example is less than the numbers of individuals in the preceding two examples. A conjecture may be made that if the study were repeated on a larger sample, the four factor model could be rejected by the likelihood ratio statistic at a higher level of significance. If our development of the reliability coefficient is justified, it should not change in a systematic fashion.

Table 4 presents results for the eighteen special tests in Lord's [1956] study of speed factors. Six tests were constructed in each of three ability factors. Two of the tests for each factor were power tests, one test was moderately speeded, and three were speed tests. A three factor model may be rejected at a very extreme level of significance but this model has a moderately high reliability of .958 and the three rotated factors represent the three ability factors. Some psychologists might wish to accept this representation of the relations among the scores on these tests. A four factor solution may also be rejected at an extreme level of significance but it has a quite high reliability of .988. This solution adds a small general speed factor to the three ability factors. Again, some psychologists might wish to accept this solution. Rejection of a five factor model on the basis of the likelihood ratio is problematic. By this number of factors the reliability has become extremely high. However, the five factor solution adds only an indication of some differentiation among the types of speeded tests. Otherwise the results appear very similar to the four factor solution.

The preceding examples utilized data from studies involving measures

TABLE 4

**Maximum Likelihood Factoring Reliability Example From Lord's Speed Factor Study**
**N = 649 , n = 18**

| Number of Factors | F | df | p | M | $M^{1/2}$ | ρ |
|---|---|---|---|---|---|---|
| 0 | 14.133 | 153 | *** | .0924 | .304 | --- |
| 1 | 7.576 | 135 | *** | .0561 | .237 | .399 |
| 2 | 2.248 | 118 | *** | .0191 | .138 | .807 |
| 3 | .548 | 102 | *** | .0054 | .073 | .958 |
| 4 | .234 | 87 | *** | .0027 | .052 | .988 |
| 5 | .143 | 73 | .08 | .0020 | .044 | .996 |
| 6 | .100 | 61 | .39 | .0016 | .040 | .999 |

*** $p < .001$

and performances of real people. To gain further experience with the reliability coefficient, maximum likelihood solutions were obtained for twelve of the correlation matrices in the study by Tucker, Koopman, and Linn [1969] on simulated correlation matrices. These matrices were constructed for 20 attributes and for populations of individuals. A domain of major common factors was combined with minor common factors and unique factors. The minor common factors numbered 180 and had random factor loadings in decreasing magnitude with progression of factors. These twelve correlation matrices differ on three experimental design variables: number of factors in the major domain, 3 or 7; proportion of variance in the measures deriving from the major domain (range of $B$), high (.6 to .8) or low (.2 to .4); and form of derivation model: "formal" involving only major domain common factors and unique factors, "middle" involving all three types of factors; and "simulation" involving only major domain and minor domain common factors. A point to be considered is that maximum likelihood factor solutions are quite feasible for these matrices but that the likelihood ratio statistics are not appropriate. There is no sampling of individuals problem. Differences between the matrices represent differences in how well the theoretic factor model with a limited number of common factors represents the interrelations of the attributes.

Results for these twelve matrices are given in Table 5 which presents the reliabilities for factor models having numbers of factors equal to the number of factors in the major domain. In all four cases the fit was exact for the formal model and the reliabilities were unity. Results for the middle model were higher than for the simulation model. Reliabilities, except for the formal model were higher for three factors in the major domain than for seven factors in the major domain. The combination of high range of $B$ and middle model yielded quite acceptable reliabilities in the middle nineties while combinations of low range of $B$ and simulation model yielded quite

TABLE 5
Maximum Likelihood Factoring Reliability
Simulated Correlation Matrices
$N = \infty$ , $n = 20$

1. Three factors in major domain,
reliabilities for three common factor models

| Range of B | Derivation Model Formal | Middle | Simulation |
|---|---|---|---|
| High | 1.000 | .961 | .831 |
| Low | 1.000 | .851 | .548 |

2. Seven factors in major domain,
reliabilities for seven common factor models

| Range of B | Derivation Model Formal | Middle | Simulation |
|---|---|---|---|
| High | 1.000 | .941 | .712 |
| Low | 1.000 | .739 | .481 |

unacceptable reliabilities around .5. These results indicate the relation of the reliability coefficient to the quality of the data entered into a factor analysis. For high reliability, the minor common factors should be held to a low level of influence. Higher reliabilities are obtained when higher proportions of the variances of measures on attributes are derived from the major factor domains. It is better to have a higher ratio of number of attributes to number of factors in the major domain.

The proposed reliability coefficient for maximum likelihood factoring appears to summarize the quality of representation of the interrelation of attributes in a battery by a factor analytic model having a limited number of common factors. It does not appear to provide a criterion as to how many common factors to accept. However, as pointed out previously, the likelihood ratio test also does not provide such a criterion. The number of factors to accept appears to depend on size of loadings and meaningfulness of factoring results. In conducting a factor analytic study, a large enough sample of individuals or objects should be used to yield stable results. The likelihood ratio statistic should indicate that all models with fewer common factors than acceptable on other grounds should be rejected at an extreme level of confidence. This statistic might indicate that the accepted model would be rejected as not exactly representing the interrelations for a population. Any accepted solution should have a high coefficient of reliability.

## REFERENCES

Browne, M. W. Fitting the factor analysis model. *Psychometrika*, 1969, **34**, 375–394.

Gulliksen, H. & Tukey, J. W. Reliability for the law of comparative judgment. *Psychometrika*. 1958, **23**, 95–110.

Harman, H. H. *Modern factor analysis*, 2nd edition. Chicago: University of Chicago Press, 1967.

Holzinger, K. J. & Swineford, F. A study in factor analysis: the stability of a bi-factor solution. *Supplementary Educational Monographs, No.* **48**. Chicago: Dept. of Education, University of Chicago, 1939.

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

Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 1940, **60**, 64–82.

Lord, F. M. A study of speed factors in tests and academic grades. *Psychometrika*, 1956, **21**, 31–50.

Mosteller, F. Remarks on the method of paired comparisons, III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, **16**, 207–218.

Mullen, F. Factors in the growth of girls seven to seventeen years of age. Unpublished Ph.D. Thesis, Dept. of Education, University of Chicago, 1939.

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

Thurstone, L. L. & Thurstone, T. G. Factorial studies of intelligence. *Psychometric Monograph No.* **2**, 1941. Chicago: University of Chicago Press.

Tucker, L. R. An inter-battery method of factor analysis. *Psychometrika*, 1958, **23**, 111–136.

Tucker, L. R, Koopman, R. F., & Linn, R. L. Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 1969, **34**, 421–459.

*Manuscript received 6/30/71*
*Revised manuscript received 2/2/72*