ACCURACY OF AN INDIRECT FACTOR ANALYTIC SOLUTION AS A FUNCTION OF THE METHOD OF ASSIGNING VARIABLES TO TOTALS[1]

Barbara M. Barker, Harry R. Barker, Betty B. Carlton
University of Alabama

Introduction

In research, dimensions within any set of measurements are frequently determined subjectively by the investigator. What these dimensions actually are is testable through an approach known as factor analysis. Measured variables in most research typically are many, but limitations on numbers of variables have been set by computer capability. Computers at most installations cannot process more than 80 to 100 variables at one time, and this many variables consume much computer time, space and money. Because of computer limitations, researchers have usually factored subsets of large data sets, but this method of coping is generally inaccurate and unsatisfactory.

An example of a data set previously unfactorable in its entirety is the MMPI. This instrument represents 566 variables--challenging the capability of even the largest computer. Subjectively, past research has indicated or proposed 12 dimensions or scales in this instrument. To test the theories, Comrey (1958) factored subsets of the MMPI that represented the scales. Comrey found that these scales each were impure and that each contained several factors (rather than the postulated single dimension).

Barker, Fowler and Peterson (1971) with the aid of a special system (Matlan, 1968) factored a 373 item short form of the MMPI. Using an IBM 360/50 computer, it required 50 hours of computer time over a period of two years. Their factoring indicated nine factors or dimensions in these 373 variables rather than the postulated twelve. They attempted to factor the original 511 item MMPI, but because of computer limitations, the attempt was not successful. What science needs is either a larger computer capacity or a method of handling a large number of variables more efficiently.

Horst (1965) proposed a novel, indirect factor analysis method which is capable of handling a large number of variables and which is based on three steps:

(1) The set of variables is reduced to a limited number of totals by grouping individual variables in some unspecified order.
(2) Factor analysis is applied to the matrix of totals.
(3) Matrix operations are used to estimate the factor loads for individual variables.

Barker translated Horst's method into a Fortran Program which has been refined and evaluated along several parameters (Stallings, 1973; Sloan, 1973; Barker and Barker, 1974, 1975A, 1975B). Potential number of variables is maximal; computer time and space required are minimal. Accuracy, however has continued to be a problem. Sequential assignment of variables to totals has proven quite unsatisfactory. Assignment of variables to totals based on prior knowledge of factor composition was more accurate. A rationale or method for assigning variables to totals is needed.

Methodology

Using the Univac 1110 and a computer program, CORR99 (Barker, 1973), the accuracy of the indirect factor analytic method was examined as a funtion of the method of clustering variables used for totals. A data set previously factored by Barker, Fowler and Peterson (1971) consisting of 1575 Veterans Hospital Patients was used. Because of Comrey's previous factorization of the scales of the MMPI, the F scale (64 items) of the MMPI was chosen as a subset small enough to factor by the conventional method. Comrey had used Thurstone's technique of factoring; consequently, the principal axes method used in this research did not replicate precisely Comrey's results. A program devised by Barker, CORR12 (1973), was used for the principal axes solution and for the varimax rotation. It was noted, using a Scree Test (Cattell, 1966), that the F scale was composed of four dimensions. These four dimensions were identified in terms of the variables that loaded greater than or equal to .3 on only one factor.

To establish that it was possible to replicate the conventional solution through the indirect method, the variables defining each factor were clustered together into separate totals, and factored by the indirect method. Items not identifying a factor were included in the factor analysis, but did not form a total that was included in the factoring itself. The estimated factor loadings of all 64 variables were then rotated to the conventional solution variable factor loads by a computer program, CORR22. Results were satisfactory (see Table 1).

Since it was established that the accuracy of the indirect method depends upon having the variables that define a factor composing a total, the next step appeared to be to find a rationale for assigning variables to totals. A clustering technique, Ward's Hierarchical Grouping Technique, seemed appropriate. This technique clusters variables into homogeneous groups. Our computer program (CORR23) is presently limited to 150 subjects and 150 variables. This accommodates the number of variables of the F scale, but larger data sets could not be clustered at one time. Because our data set was composed of 1575 subjects, too many for the program to handle at one time, three random samples of 150 subjects were selected and analyzed. Three overlapping sets of 42 sequential variables were also used for assigning variables to totals in the indirect method (items 1-42; 22-64; and 1-21 and 43-64). Rationale for using overlapping sets of variables in the Ward's technique was that large numbers of variables might also exceed the limitations imposed by the computer for this program.

Based on a prior knowledge of the factor structure four factors were rotated to a varimax criterion. This was then compared with the conventional solution in terms of identifying variables to define factors. Using the factor loads of the conventional solution and the indirect solution, the factors were rotated in factor space to determine the extent to which variables could be aligned.

Results

The results of the clustering indicated about six groups of variables before the error rate increased prohibitively. The resulting variable groupings were compared across the three overlapping sets and across the three random samples. When a variable clustered with a particular group in two of the three overlapping sets, it was included in a total. All variables that did not appear together on two of the overlapping sets were put into a seventh total. The seventh total was not factored but factor loads were estimated for all variables involved. Four factors were rotated to a varimax criterion. The three factor analyses were examined again for overlapping sets of variables to use in an iteration. Those items that appeared in a factor on at least two of the three factorings from the overlapping sets were regrouped into a total and factored. This improved the identification of variables loading significantly on the first factor, but no improvement was made in identifying variables in factors II, III, and IV. All 26 variables on factor I were identified. Nine of the thirteen from factor II, and one each from factors III and IV were also identified. Five items were also correctly identified as loading less than .3 on all factors. A second iteration produced no further improvement in identification of items on any factor. The rotation of the indirect solution to the conventional solution showed much agreement between the two factor methods. The results are in Table 2.

Summary

In summation, it would appear that a rationale for assigning variables to totals for use in the indirect factor analytic method is not easily attainable. Factor identification has been progressively better through successive use of Wards' Hierarchical Grouping Technique. Fifty-nine of the 64 variables from the F scale of the MMPI have been rotated to agree with the conventional solution at .8 or better.

An important discovery evolved from the use of the Wards' Technique. It was found that scoring the items according to a directionality key increased the accuracy of prediction. Further progress is expected in providing the researcher with a method of factor analyzing very large numbers of variables. Current research is underway that uses the method as an hypothesis testing procedure for factor structure.

Footnote

[1]Paper presented at American Statistical Association, August 28, 1975. This research was supported by funds from the Research Grants Committee of the Univeristy of Alabama and Roche Laboratories, Division of Hoffmann-La Roche, Inc.

References

Barker, B.M., and Barker, H.R. Search for a method of increasing the accuracy of a novel, indirect method of factor analysis. (Paper presented at Southeastern Psychological Association, Atlanta, Georgia, May 28, 1975).

Barker, H.R. Behavioral Sciences Statistics Program Library. Tuscaloosa, Alabama: Reproduction Services, 1973.

Barker, H.R., and Barker, B.M. Accuracy of an indirect factor analytic solution as a function of the number of totals used. (Paper presented at Southeastern Psychological Association, Atlanta, Georgia, May 28, 1975).

Barker, H. R., Fowler, R.D., and Peterson, L.P. Factor analytic structure of the short form MMPI items in a VA hospital population. Journal of Clinical Psychology. 1971, Vol. XXVII, No. 2, 228-233.

Comrey, A.L. A factor analysis of items on the F scale of the MMPI. Educational and Psychological Measurement, 1958, 18, 632.

Horst, P. Factor Analysis of Data Matrices. New York: Holt, 1965.

Hunnicutt, (Barker), B.M., and Barker, H.R. Application and evaluation of a novel and indirect factor method to a very large data matrix. (Paper presented at Southeastern Psychological Association, Hollywood, Florida, May 3, 1974.

IBM system/360 matrix language (Matlan 360A-CM-05X). Program Description Manual, 1st Ed., 1968, IBM Technical Publications Department.

Sloan, H.C. A simplified procedure for estimating the factor structure of large data matrices. Unpublished doctoral dissertation. University of Alabama, 1973.

Stallings, N.A. Evaluation of an indirect method of estimating the factor structure of large data matrices. Unpublished doctoral dissertation, University of Alabama, 1973.

Table 1

Cosines of Angles Between Original and Rotated Location of Variables in Factor Space*

VARIABLES	COSINES								
1-10	.95	.92	.87	.97	.98	.98	.97	.96	.97	.96
11-20	.96	.95	.99	.94	.92	.95	.98	.93	1.00	.93
21-30	.90	.97	.95	.96	.94	1.00	.93	.95	.92	.91
31-40	.96	.91	1.00	.98	.97	.95	.95	.92	.95	1.00
41-50	.99	.96	.98	.99	.97	.98	.99	.99	.99	.99
51-60	.98	.99	.98	.99	.95	.99	.98	.99	.98	.99
61-64	1.00	.93	.99	.99						

*Assignment of variables to totals based upon conventional factoring.

Table 2

Cosines of Angles Between Original and Rotated Location of Variables in Factor Space**

VARIABLES	COSINES									
1-10	.87	.87	.13	.81	.92	.93	.92	.96	.94	.86
11-20	.94	.86	.65	.94	.83	.93	.93	.91	.99	.84
21-30	.69	.86	.86	.93	.94	.98	.91	.95	.81	.89
31-40	.85	.82	.98	.92	.91	.88	.98	.76	.94	1.00
41-50	.96	.89	.94	.94	.95	.76	.93	.98	.98	.94
51-60	.97	.94	.97	.92	.93	.97	.90	.90	.95	1.00
61-64	1.00	.95	.96	.98						

**Assignment of variables to totals based upon Wards' Hierarchical Grouping Technique.