AN INDIRECT METHOD FOR TESTING THE DIMENSIONALITY OF LARGE DATA SETS

Harry R. Barker and Barbara M. Barker, University of Alabama

The modern computer imposes severe limitations on the number of variables which can be factored by conventional methods. Various alternative factor solutions are discussed and evaluated by Barker, Fowler, and Peterson (1971). Each of these alternatives proves to be grossly inadequate from a theoretical standpoint.

An alternative factor solution proposed by Horst (1965) seems theoretically sound and offers a method of factoring huge data matrices with minimum time requirements on a medium size computer. Horst's method essentially entails combining the variables into a relatively small number of subsets, factoring the subsets and then estimating the factor loads of the individual variables. In this way, the large matrix of variables is bypassed. (See Barker and Barker (1974) for a more extended illustration of the method.)

Horst's indirect factor method has been systematically studied and refined over an extended period of time. See Stallings (1973), Sloan (1973), Barker and Barker (1974), Barker and Barker (1975-A), Barker and Barker (1975-B), Barker, Barker, and Carlton (1975), Barker and Barker (1975-C), Barker and Barker (1976). A summary review of research on the method through 1975 is given by Barker and Barker (1975-C). Two major findings resulted from the series of studies:

(1) In order for Horst's method to produce accurate results it is essential that the variables comprising each subset belong to the same factor.
(2) Combination of the variables into a subset, by totaling, must take into account the directionality of the measurement, positive or negative.

These findings appeared, at first, to impose drastic limitations on the use of Horst's method. For it is essential to know the correct solution before the method can produce the correct solution. On further reflection it was noted that theory serves just this purpose; providing an assertion as to what the correct answer or solution is. Horst's method then seems to be an ideal method for testing the accuracy of theory. The method has proved useful in evaluating the accuracy of competing theories as to the factor structure of large matrices. (See Barker and Barker (1976) and Hamlett (1976). The manner of using the procedure is as follows. A theory is used to cluster variables into subsets. Horst's factor method is then applied and one simply notes the extent to which variables which were previously clustered into the same subset emerge together on a single factor.

The purpose of this study is to demonstrate the use of Horst's indirect factor method in evaluating four competing theories relating to the factor structure of the widely used Minnesota Multiphasic Personality Inventory (MMPI).

The four competing theories to be evaluated are as follows:

(1) Lushene (1967) A theory derived from the entire 566 items MMPI which was administered to 189 male undergraduate college students. 18 factors were postulated.
(2) Barker, Fowler, and Peterson (1971) A 373 item short form of the MMPI administered to 1575 VA male hospital patients comprised the source of the 9 factor theory.
(3) Tryon (1966) The entire 566 item MMPI, administered to 310 adult subjects, sex not specified, was the origin of a 7 factor theory. Tryon performed an oblique factor rotation. The first three factors were only slightly correlated, whereas the remaining four are moderately correlated with the first three; therefore the present treatment of his factors as orthogonal should be regarded as suggestive only.
(4) Conventional theory The entire MMPI (511 items at the time) was presumably involved in selecting items for the 9 standard clinical scales, which represent Kraepelinian nosological groupings in vogue at the time of construction of the scales. The remaining items are treated as filler items.

METHOD

DATA

Item answers of 225 male subjects to the original 511 items of the MMPI comprised the data for the study. This is the original sample of males used to determine the standard male norms for the MMPI scales.

COMPUTER PROGRAM

Horst's indirect factor method was translated into a Fortran Computer Program by Barker (1973). The program has been subsequently refined by Stallings (1973) and by Sloan (1973). The computer program clusters variables into a priori subsets, totals the variables in the subsets, determines principal components of the totals and then estimates the principal axes loads of individual variables. The principal axis solution is finally rotated to a varimax criterion. A factor load of + or - .3 or greater on only one factor is required to identify a variable with a factor.

COMPUTER

The computer program was originally developed for an IBM 360 Mod 50 Computer, with 120 K core storage allotted. It was necessary to make extensive alterations in the program to enable it to run accurately on a Univac 1110 system with 128 K core storage allotment. The program capacity is approximately 2000 variables which can be factored.

D MEASURE

The information measure D (relative uncertainty reduction) is employed to relate the degree of association between theorized factor structure (specified subsets of items) and actually obtained factor structure.

RESULTS

For each of the competing theories respectively, Tables 1, 2, 3, and 4 provide a visual display of the association between theorized and obtained factor structure. If the agreement between theory and obtained factor structure were perfect, the diagonal cells of the table would be filled and the off-diagonal cells would be empty. The D measure would then reflect perfect agreement with a value of 1.00. To the extent that cell entries appear in off-diagonal cells, a lack of agreement between theorized and obtained factor structure is indicated. The limit of disagreement results in an equal number of entries in each cell and the resulting D measure is 0.0.

The D values computed for each of the theories are rank ordered from highest to lowest as follows:

(1) Lushene D = .56
(2) Barker, Fowler, and Peterson D = .52
(3) Tryon D = .42
(4) Conventional D = .42

It is clear that none of the four theories provides a satisfactory theory of the factor structure of the MMPI in this population of subjects. The Lushene and Barker et al theories are virtually identical in D value whereas the Tryon and conventional theories are clearly inferior. The Barker et al theory is more parsimonious than the Lushene theory in that only 9 factors are specified whereas Lushene's theory specifies 18. The results with respect to Tryon's theory should be regarded as tentative. Tryon theorized oblique factors whereas the method presently employed treats all factors as orthogonal.

DISCUSSION

The size of the obtained D values relating the theorized and obtained factor structures indicates that none of the four theories is adequate in representing the dimensionality of this data set. Several lines of conjecture merit consideration.

First, none of the existing theories may be truly adequate. This could conceivably result from defects in methodology and/or procedure in arriving at the competing theories.

Second, the population of males, used to derive standard male norms for the MMPI, may represent an unusual cross section of individuals, which serves to conceal some factors and high light other or different factor dimensions. Perhaps it should be noted here that the male subjects were relatives of neuropsychiatric patients, whom they were visiting in the hospital.

Third, the size of the sample used for norming purposes must be viewed as grossly inadequate by current multivariate research standards. For example, Nunnaly (1967) argues for at least ten subjects per variable while Cattell (1966) views a desirable sample size as 100 + the number of variables. The number of subjects in this study falls far short of either Nunnaly's or Cattell's criterion. Thus, instability of findings due to the unfavorable ratio of number of variables to number of subjects may have produced inaccurate results.

Another interesting matter relates to the concept of false positives. The term accurately relates to three of the theories Lushene, Tryon, and conventional theory. Each of these three theories was developed from the entire MMPI. Thus, the appearance of items on factors which were not predicted by the theory are accurately labeled false positives. But in the case of the Barker, Fowler, and Peterson theory, the theory was derived from a subset of the entire MMPI (Viz. 373 items). Therefore the 9 factor theory did not make prediction for items not included in the 373 item set. Identification of these items as false positives is actually inappropriate. Nevertheless the items are so regarded for this study.

A possible solution of this issue is to extend the nine factor theory to encompass the new items (false positive items not included in the original 373 items on which the theory was derived). The new items thus identified with the theory, the indirect factor method would again be employed and the resulting solution evaluated by the D measure. Of course the altered 9 factor theory would now be viewed as a revision of the earlier theory. This general procedure of revising theorized factor structure on the basis of false positives (and perhaps also false negatives) is currently under investigation as a procedure to:

(1) improve the accuracy of theories regarding factor structure.
(2) develop factor theory in areas where none presently exists.

SUMMARY

An indirect method of factor analysis was used to evaluate four competing theories relating to the factor structure of the MMPI for the male population on whom the original MMPI standardization was performed.

Results clearly indicated the inadequacy of each of the four theories and provided a rank ordering of their relative accuracy.

REFERENCES

Barker (Hunnicutt), 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 the Southeastern Psychological Association, Hollywood, Florida, May 3, 1974.

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, March 28, 1975B.

Barker, B. M., Barker, H. R., and Carlton, B. B. Accuracy of an indirect factor analytic solution as a function of the method of assigning variables to totals. Proceedings of the Social Statistics Section, American Statistical Association, 1975, 295-297.

Barker, H. R. Behavioral Sciences Statistics Program Library. University of Alabama, 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 the Southeastern Psychological Association, Atlanta, Georgia March 28, 1975A.

Barker, H. R., and Barker, B. M. An indirect factor method for testing the dimensionality of large data sets. Paper presented at Southeastern Psychological Association, New Orleans, Louisiana, March 19, 1976.

Barker, H. R., and Barker, B. M. Summary of Research on a novel, indirect factor analytic solution. Proceedings of the Social Statistics Section, American Statistical Association, 1975C, 298-301.

Barker, H. R., Fowler, R. D., and Peterson, L. D. 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.

Cattell, R. B. (Ed.). Handbook of multivariate experimental psychology. Chicago: Rand-McNally, 1966.

Hamlett, C. C. Validation of the theorized factor structure of the personal orientation inventory. Paper presented at American Statistical Association, Boston, Massachusetts, August 23, 1976.

Hathaway, S. R., and McKinley, J. C. MMPI manual, Revised Ed. New York: Psychology Corporation, 1951.

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

Lushene, R. E. Factor structure of the MMPI item pool. Unpublished master's thesis, Florida State University, 1967.

Nunnally, J. <u>Psychometric theory</u>. New York: McGraw-Hill, 1967.

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

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

Tryon, R. C. Unrestricted cluster and factor analysis, with applications to the MMPI and Holzinger-Harman problems. <u>Multivariate Behavioral Research</u>, 1966, Vol. 1, 229-244.

TABLE 1

ASSOCIATION BETWEEN ITEM SUBSETS AND VARIMAX FACTORS (CONVENTIONAL THEORY)

ITEM SUBSETS X	Factors (Y) I	II	III	IV	V	VI	VII	VIII	IX	False -	Sum
1	6									2	8
2		4			1					18	23
3		5	2							9	16
4		2		4		1	1			13	21
5		4			3		2			12	21
6		7				6				4	17
7		10					1			7	18
8		15		1						22	38
9										19	19
False +	8	53					6				67
Sum	14	100	2	5	4	7	10	0	0	106	248

H(X) = 1.884 H(X,Y) = 4.184
H(Y) = 3.085 HT = .785 D = .42

TABLE 2

ASSOCIATION BETWEEN ITEM SUBSETS AND VARIMAX FACTORS (TRYON THEORY)

ITEM SUBSETS X	Factors (Y) I	II	III	IV	V	VI	VII	False -	Sum
1	19							6	25
2		20						12	32
3			12					13	25
4			1	10				16	27
5			2		1			17	20
6			5	1		6		9	21
7			1	9			1	23	34
False+		5	6	1		2	1		15
Sum	19	25	27	21	1	8	2	96	199

H(X) = 2.232 H(X,Y) = 4.255
H(Y) = 2.959 HT = .936 D = .42

TABLE 3

ASSOCIATION BETWEEN ITEM SUBSETS AND VARIMAX FACTORS (BARKER, FOWLER AND PETERSON THEORY)

ITEM SUBSETS X	Factors (Y) I	II	III	IV	V	VI	VII	VIII	IX	False –	*Sum
1	*20						8			14	* 42
2	*	23					1			25	* 49
3	* 2		14							11	* 27
4	*			11						11	* 22
5	*		1		9					4	* 14
6	*10						1			7	* 18
7	*						22			11	* 33
8	*							11		1	* 12
9	*						2		4	2	* 8
False+	30	3	3	1	0	0	22	2			* 61
Sum	*62	26	18	12	9		56	13	4	86	*286

H(X) = 2.664 H(X,Y) = 4.363

H(Y) = 3.084 HT = 1.385 D = .52

TABLE 4

ASSOCIATION BETWEEN ITEM SUBSETS AND VARIMAX FACTORS (LUSHENE THEORY)

ITEM SUBSETS X	Factors (Y) I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	False –	Sum
1	31																		17	48
2	8																		15	23
3	1		7								1								10	19
4	3			10									1						10	24
5	2																		14	16
6						2													8	10
7							1													1
8								8	1										4	13
9	1					4													6	11
10						6													4	10
11						2					3								5	10
12												1							5	6
13										2			2					1	2	7
14														2					6	8
15													1		5				2	8
16											2					3			3	8
17								1									4		2	7
18	1																	3		4
False +	54										3	1	16	1				1		76
Sum	101		7	10		14	1	9	1	2	9	2	20	3	5	3	4	5	113	309

H(X) = 2.648 H(X,Y) = 4.832

H(Y) = 3.668 HT = 1.484 D = .56