SOME CRITICISMS OF FACTOR ANALYSIS WITH SUGGESTIONS FOR ALTERNATIVE METHODS

By

MONICA A. CREASY, B.Sc.

Lecturer in Statistics
Psychology Department, Bedford College, University of London

I.—Introduction

It is well known that a factor analysis may be carried out in various ways; the subjective elements involved make it by no means clear that two people beginning with the same data will arrive at the same conclusions. Furthermore, not enough attention has been paid to the confirmation of factors revealed in one analysis by repetition of the tests on another sample from a similar or different population.

Besides this, very often the data which are factor analysed are not suitable for this type of treatment, and the value of a piece of research is often obscured by the inappropriate use of the technique, and by a failure to clarify the purpose of the statistical analysis. In a recent paper Storms (1958) has shown how the technique of discriminant functions may reveal more clearly than factor analysis differences in personality structure in different psychoneurotic groups.

The purpose of this paper is twofold. In the first place an attempt is made to elucidate some of the sources of confusion and circularities of argument which arise when a factor analysis is carried out. Attention is also drawn to the need for a detailed description of the sample before an interpretation of the factors extracted can be made. Alternative approaches are then discussed. In some circumstances a more satisfactory procedure is to make a hypothesis regarding the presence of certain factors, and to test this hypothesis; in others, other techniques may be appropriate, and these are discussed and compared. Lastly the use of a new technique is exemplified in relation to a particular piece of research.

II.—Naming and Orthogonality of Factors

It has frequently been pointed out (for example by Burt (1940) and Thomson (1950)) that a factor is a weighted mean of the scores of the tests included in the battery. In spite of this, in naming a factor it is often implied that the measure obtained has some absolute meaning. This cannot be the case unless the tests included in the battery are really representative of all the possible tests of the ability or trait which gives the factor its name. The name given to a factor may, in fact, conceal its true nature, and little is gained in being told that "the factor X has been identified", if, on inspection of the tests loaded with this factor, it appears that the author has greatly extended or limited or even changed the attributes usually ascribed to the quality X.

Furthermore, the definition of a factor may be considerably strained owing to a prior assumption that any factors which are present are orthogonal, or alternatively, that any orthogonal factors found with high loadings are psychologically meaningful. Now, if an examination of the tests *after* the factor

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

analysis suggests that they measure certain characteristics in common, then a similar examination *before* factor analysis should reveal the same possibility. A hypothesis can therefore be made to this effect which can be tested in the manner outlined in Section IV. Other factors can also be postulated and their significance and mutual orthogonality tested.

It is interesting to examine Eysenck's (1954, chap. 4) study of political affiliation in the light of this criticism. In this study, which is based on the answers to an opinion survey, the author claims to have discovered two fundamental dimensions on which opinions are based; one he names Radicalism-Conservatism, and the other, which is orthogonal to the first, he names Toughminded-Tenderminded. Now, examination of those opinions which have high loadings on the first factor suggest that it has been correctly named, and it seems unlikely that many people will disagree with what is there regarded as "radical" and "conservative". However, the axis orthogonal to it is not so easily recognizable and it is not at all obvious that it measures a "nameable" characteristic. Because a few statements at one end of the scale appear to fit in with William James' (1907) concept of "Toughmindedness" and a few at the other end fit in with the opposite idea of "Tendermindedness", the whole scale has been labelled with these names. As a result of this, some of the statements, particularly those relating to morality and religion, seem to be classified as "tenderminded" whereas in the everyday use of these terms they might well be regarded either as unclassifiable as they stand, or else as "toughminded". It seems worth noting that the qualities listed by James under the two headings were ascribed to the two main schools of philosophers existing at that time, and their meanings can only be interpreted in this context. For one group we have Rationalistic, Intellectualistic, Idealistic, Optimistic, Religious, etc. (tenderminded), and for the other, Empiricist, Sensationalistic, Pessimistic, Irreligious, etc. (toughminded); in the context of James' writing it becomes clear that these adjectives have a considerably narrower meaning than is usual. Whether or not one agrees with the meaning ascribed by Eysenck to the new dimension is clearly a matter of opinion, but the method of arriving at the factor cannot be regarded as satisfactory. An examination of the meaning of each statement should enable one to make a hypothesis regarding the extent to which it measures "tendermindedness" or "toughmindedness", and the significance of the dimension could then be tested, and also its orthogonality to "Radical-Conservatism".

III.—Population

A fact which is often ignored is that a change in the variation of the population by selection affects the correlation coefficients, which in turn affect the factor loadings. It is unlikely that a "g" factor will be discovered in a highly intelligent group of subjects, for the emergence of a factor implies variation on that factor in the group tested. Thus if a "neurosis" factor is found in a group it implies that some members of the group are more neurotic than others, and, on the other hand, if this factor is not revealed it may mean that either the whole group is neurotic or that none of them is.

Secondly, the "significance" of a factor is often interpreted in terms of the proportion of the overall variance accounted for by the factor. If a large number of irrelevant tests are included in the battery the proportion contributed by a particular factor is reduced; conversely, if only tests which are known to be highly correlated are included, the factors which emerge will account for a

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

high proportion of the variance. This has been pointed out by Babington Smith (1950, section 9). Thus, every time a factor is revealed, or not revealed in an analysis an interpretation can only be made in terms of (*a*) the variation in the population sampled, which should be reported, and (*b*) the actual tests used.

There is another source of trouble which arises when instead of dealing with a random sample from one population we have several groups whose means differ significantly on some of the test scores. If this is the case, a correlation coefficient calculated over all the people is a mixture of "between groups" and "within groups" correlations, and is impossible to interpret as it stands. Figures 1 and 2 show how the inclusion of groups which differ may obscure the meaning of the correlation coefficient.

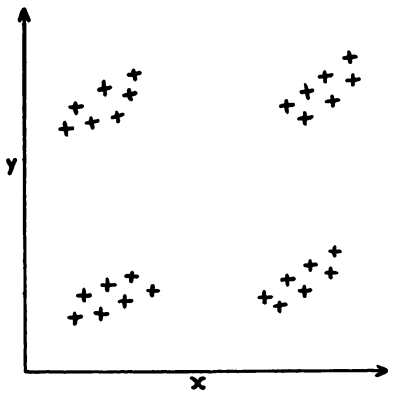


FIG. 1

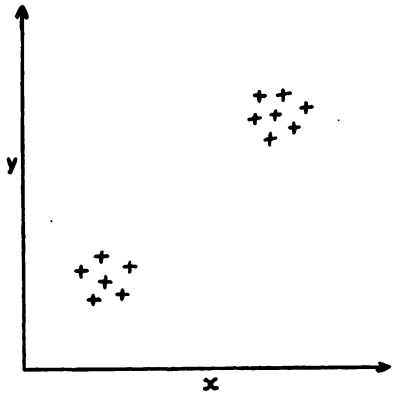


FIG. 2

In Figure 1 there is a positive correlation between the two variables within each group, but the group means are so placed that the overall correlation would be very low, and no one factor could account for most of the variance. In Figure 2, on the other hand, there is no correlation within the groups, but since the means of the two groups differ significantly an overall positive correlation would be found. In this case a factor would emerge which would probably be named by the quality distinguishing the two groups. This factor would account for nearly all the variation, since variation along the line joining the groups is large compared with that perpendicular to it. But the loadings cannot be interpreted in the usual way, nor a test of significance applied. For one of the assumptions underlying a factor analysis is that in the population studied each variable is normally distributed, and this is clearly not the case if several differing groups are included. This type of situation can probably be dealt with best by the discriminant function technique discussed in the next section. The loadings obtained from a factor analysis may be roughly related to those obtained from the use of a discriminant function, but there is no justification for carrying out the factor analysis.

Reference may again be made here to Eysenck's (1954) study. The sample investigated consisted of groups from the four main political parties, and many of the statements were such that the proportions agreeing with them differed significantly between the groups. This is therefore comparable with the situation illustrated in Figure 2, and care should be taken in attempting to interpret the factor loadings except in a very rough way. This could perhaps account for the fact that correlations calculated between the two factors *within* groups are consistently and significantly negative (Eysenck, op. cit., p. 135).

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

IV.—Alternative Approaches

In this section three possible alternative approaches are discussed, which may be appropriate in various circumstances. Each one appears to the author more objective than factor analysis, and more susceptible to tests of significance.

(a) *Analysis of Variance*

The existence of a factor implies that people vary significantly in their amounts of this factor; the method outlined here is based on this argument. A more detailed discussion is given in a paper entitled "Analysis of variance as an alternative to factor analysis" (Creasy, 1957).

This approach is essentially different from factor analysis, as it requires the research worker to postulate initially which factors are present in the tests used, and to what degree. The need for careful construction of a test battery has, indeed, been stressed by Guilford (1952), who points out that the initial planning should emphasize the formation of hypotheses as to what factors are likely to be found. It should be possible from examination of the tests, or from the results of a previous factor analysis, to postulate the "loading" of each test on each factor on either a 3- or 5-point scale, or merely as "present" or "absent". A measure of each person's "amount" of this factor is then the weighted sum of the test scores, weighted with the postulated factor loadings. This can be obtained for each factor and the variation between people's scores calculated.

The statistical analysis entails breaking down the overall variation between people on all the tests into components due to variation between people on the different factors postulated and residual variation. It is then possible to test the significance of these components compared with the residual variation as in an analysis of variance. If it is required to test whether the factors are orthogonal to each other this can also be done. More details can be obtained from the above mentioned paper where some examples are given and comparisons made with factor analysis. The algebra given there reveals an interesting relationship between the analysis of variance components and the correlation coefficients between the tests. (This is illustrated in the example in the following section.) It is also shown that there is very little loss of accuracy if "crude" loadings are used in place of loadings with two significant figures. If the battery of tests is balanced so that each factor occurs equally often with every other factor, the amount of computation is considerably less than that involved in a factor analysis; in any case the labour is to some extent reduced.

The advantage of this method is that it avoids some of the unsatisfactory features of factor analysis. Any elements of subjectivity involved in naming the factors can be clearly stated at the outset when the factors are postulated; furthermore, the significance and orthogonality of the factors can be tested in a satisfactory manner.

Lawley (1958) has described a method of estimation of loadings in a factor analysis where it can be initially postulated that certain loadings are zero. This method gives "best" loadings under these assumptions and is in a sense halfway between the analysis of variance approach, where all loadings are postulated, and factor analysis where all loadings are estimated. The disadvantages of the method are that as estimates are obtained by iteration the method becomes very unwieldy if more than three factors are postulated and no test of significance is available.

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

(b) *Discriminant Function*

Suppose we have two groups of people that are known from general and perhaps subjective considerations to differ in respect to a quality. Then the construction of a battery of objective tests related to this quality may enable us to assign a further individual to one or other of the groups. The problem then arises of the best way to make use of the test scores to achieve this purpose. For example, suppose the quality is "neuroticism" and we have a group of "normals" and a group of "neurotics" diagnosed on the basis of a psychiatric interview. It may be that a series of tests could be devised which differentiated to some extent between the two groups. We require a method of combining the test scores in order to give a criterion for deciding from which category a new individual probably comes. The combination which gives the minimum number of errors of allocation is that obtained by the use of Fisher's (1948) technique of "discriminant function". If the requirements of normality and equality of variance in the groups have been satisfied, this will result in a weighted sum of the test scores, which gives the best discrimination (in a mathematical sense) between the groups. In the above example this weighted sum could be regarded as a measure of neuroticism. Such an approach helps to emphasize the fact that we are not necessarily dealing with a "factor" in any absolute sense, but are merely making the best use of the data available; measures of the main characteristics of neuroticism may, in fact, have been omitted from the battery. This situation is similar to that illustrated in Figure 2, where with only two tests, it can be seen to be reasonable to measure the distance along the line joining the two groups in order to decide to which group an individual is closer. In his study of the "neurotic dimension" Eysenck (1947) appears to recognize this situation when he "rotates to a criterion", and his result is probably fairly close to that which would be obtained using a discriminant function. However, the latter is more justifiable mathematically, more objective, and incidentally involves less computation. In a re-analysis of some data first analysed by Hildebrand (1958), Storms (1958) has clearly shown that a discriminant function can pick out the tests which discriminate groups of psychoneurotics more successfully than a factor analysis.

(c) *Arbitrary Scoring*

An individual's measure of a factor is essentially a weighted sum of his test scores. If these weights are postulated, as in (*a*) above, we can obtain for each person a "postulated" measure which, depending on the weights chosen, will be more or less reasonable. If the purpose of a piece of research is simply to show that two groups differ in respect of this factor it is sufficient to show that the means of the postulated measures differ significantly by a t-test. This method of arbitrary weighting will not give the most sensitive test of difference between the groups, which is that discussed in (*b*) above, but it may be the most appropriate test in the circumstances and will certainly be easier to calculate.

V.—Analysis of Variance (Example)

In a recent (unpublished) study by Dr. G. A. Foulds and Mr. T. M. Caine an attempt has been made to group questions on the M.M.P.I. which measure hostility into five classes, three of which refer to "Extrapunitive" and two to "Intropunitive" hostility. There are thus five measures for each person, which will be expected to intercorrelate to some extent. The question arises whether the correlations are sufficiently large to suggest there is one underlying factor

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

of hostility, and secondly whether there is another "factor" which gives a measure on a possible "Extrapunitive-Intropunitive" dimension.

If a general factor is present, it can be shown that a very close approximation to its value for each person is given by his average score on all the tests. The significance of the factor is tested by seeing if variation between these averages is significantly large compared with variation between the individual test scores, within each person. This is equivalent (Creasy, 1957) to testing the significance of the average of the correlation coefficients, i.e. to testing $F=\{1+(n-1)\bar{r}\}/(1-\bar{r})$ as an F with $\{(p-1), (n-1)(p-1)\}$ degrees of freedom where n is the number of tests, p the number of people tested and $\bar{r}$ the average of the $n(n-1)/2$ correlations. The test is only approximate if $p<40$, and tends to over-estimate significance.

Now suppose we postulate an Extrapunitive hostility factor, present in the first three tests, and absent in the last two. The difference in the means of the two sets will give a measure of a person's "Extrapunitive" score, over and above his average hostility. If variation in these differences is significant, it suggests this second classification of the tests "means something" and two measures or dimensions are necessary to represent a person's "hostility position" relative to the others. The test of significance turns out to be expressible in terms of correlation coefficients, and compares the average of the correlations *within* the two sets with the correlations *between* the two sets. This is illustrated below. For details of the formulae the reader is referred to Creasy (op. cit.).

Scores were obtained from several groups of people, of which only two will be discussed here, a group of 20 "Normal" women and a group of 20 "Normal" men. (The tests of significance are therefore only approximate.) The average correlations are given in Table I. The analysis of variance takes the form shown in Table II (see op. cit.).

TABLE I

		Women	Men	
Average (of 10)	$\bar{r}$	·286	·519	
Average extrapunitive	$\bar{r}_1$	·476	·427	
Intropunitive	$\bar{r}_2$	·470	·484	
Average extrapunitive × intropunitive	..	·160	·571	
$F=\frac{1+4\bar{r}}{1-\bar{r}}$	.. ..	3·00	6·40	(both significant at 1%)

TABLE II

Source	d.f.	Mean Square	Women Mean Square	F	Men Mean Square	F
General factor	19	$1+4\bar{r}$	2·14	4·06	3·08	5·60
Second factor ..	19	$1-4\bar{r}+2\bar{r}_1+\bar{r}_2$	1·28	2·42	·26	·47
Residual ..	57	$\frac{1}{3}(3-2\bar{r}_1-\bar{r}_2)$	·53		·55	

It can be seen that the first and second factors for women are significant; this means that two measures are necessary to represent a person's hostility "position" relative to the group. These may be, for example, the overall average score and the difference "Extrapunitive-Intropunitive", or simply the Extrapunitive and the Intropunitive scores separately. For the men, on the other hand, the second factor is not significant, the Extrapunitive and Intropunitive scores in fact correlate fairly highly with each other (·57) which suggests that

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press

they do not measure different things and that the overall average score is sufficient to describe a person's hostility measure. These conclusions refer, of course, to variation within the group, and say nothing about the measures to use in a comparison of groups.

VI.—Conclusions

The technique of factor analysis involves so many subjective elements that a piece of research summarized by a table of factor loadings is almost impossible to appreciate, even when details of tests, population, etc., are given, and these are frequently omitted. Alternative methods are available and are discussed in Section IV of this paper.

References

Bartlett, M. S., "Tests of significance in factor analysis", *Brit. J. Psychol.*, Statist. Sect., 1950, **3**, 77–85.

Burt, C., *The Factors of the Mind*, 1940. University of London Press.

Creasy, Monica A., "Analysis of variance as an alternative to factor analysis", *J. Roy. Statist. Soc.*, 1957, **B, 19**, 318–325.

Eysenck, H. J., *Dimensions of Personality*, 1947. London: Kegan Paul.

Idem, *The Psychology of Politics*, 1954. London: Routledge & Kegan Paul.

Fisher, R. A., *Statistical Methods for Research Workers*, 1948. London: Oliver and Boyd. 10th edition.

Guilford, J. P., "When not to factor-analyse", *Psychol. Bull.*, 1952, **49**, 26–37.

Hildebrand, H. P., "A factorial study of introversion-extraversion", *Brit. J. Psychol.*, 1958, **49**, 1–11.

James, William, *Pragmatism*, 1907. New York: Longmans Green.

Lawley, D. N., "Estimation in factor analysis under various initial assumptions", *Brit. J. Statist. Psych.*, 1958, **11**, 1–12.

Smith, B. Babington, "Evaluation of factor analysis from the point of view of a psychologist", *J. Roy. Statist. Soc.*, 1950, **B, 12**, 73–85.

Storms, L. H., "Discrepancies between factor analysis and multivariate discrimination among groups as applied to personality theory", *J. Ment. Sci.*, 1958, **104**, 713–721.

Thomson, G. H., *The Factorial Analysis of Human Ability*, 1950. London: University Press. 5th edition.

Thurstone, L. L., *Multiple Factor Analysis*, 1947. Chicago: University of Chicago Press.

https://doi.org/10.1192/bjp.105.440.755 Published online by Cambridge University Press