# CRITERION ANALYSIS—AN APPLICATION OF THE HYPOTHETICO–DEDUCTIVE METHOD TO FACTOR ANALYSIS

### BY H. J. EYSENCK

#### Psychology Department, Institute of Psychiatry, University of London

### I. INTRODUCTION

Among statistical methods now in common use in psychology, factor analysis presents certain interesting and paradoxical features. Although more and more psychologists, sociologists, and recently even physicists and chemists, have used factor analysis as their preferred statistical research tool, and although many experts, among whom we may mention Thurstone, Burt, Cattell, Guilford, and Vernon, have expressed their faith in its adequacy to solve some of our most pressing taxonomic problems, vet much criticism of this technique has been advanced by two schools of thought usually on opposite sides of the fence. Factor analysis has been criticized severely by those who, like Allport, Murray, and other adherents of the psychiatric, individualistic, idiopathic point of view believe its atomistic assumptions violate the holistic nature of human personality. It has equally suffered the onslaughts of professional statisticians who point out its formal deficiencies, and prefer the more rigorous methods of discriminant function analysis, analysis of variance, and regression equations. No other method in statistical psychology has suffered such a multiplicity of criticisms, and it behooves those of us who make use of it to look carefully at the various aspects of factor analysis which may be considered most vulnerable to such attacks.

Truman L. Kelley has pointed out that statistics has three main functions. "The first function of statistics is to be

purely descriptive, and its second function is to enable analysis in harmony with hypothesis, and its third function to suggest by the force of its virgin data analyses not earlier thought of" (9, pp. 22, 23). While most statisticians would agree to the descriptive purposes of many statistical constants, there is less agreement regarding the other two functions mentioned by Kelley. "We may say that there are two occasions for resort to statistical procedures, the one dominated by a desire to prove a hypothesis, and the other by a desire to This has led to distinct invent one. schools of statisticians, both lying within the general field of scientific endeavor" (9, p. 12).

Reference to this dichotomy of functions and purposes within the field of statistics gives us one hint as to the reasons for the general statistical criticisms of factorial methods. Unlike analysis of variance and covariance, discriminant function analysis, or even the humble C.R. or t form of analysis, factor analysis does not in general attempt to prove or disprove a hypothesis; it does not set out to disprove any form of null hypothesis at some critical level of significance. Its function appears to be far more dominated by the desire to "invent" a theory, and in spite of Kelley's words this function of statistics is not generally recognized by statisticians as being truly within the purview of this particular branch of science.

Holzinger describes the nature of factor analysis thus: "Factor analysis is a branch of statistical theory concerned with the resolution of a set of descrip-

tive variables in terms of a small number of categories or factors. . . . The chief aim is . . . to attain scientific parsimony or economy of description" (7, p. 1). Similarly Kelley: "There is no search for timeless, spaceless, populationless truth in factor analysis; rather, it represents a simple, straightforward problem of description in several dimensions of a definite group functioning in definite manners, and he who assumes to read more remote verities into the factorial outcome is certainly doomed to disappointment" (10, p. 120). It is clear that these authors regard the task of factor analysis as essentially one of arriving at a convenient small set of fictitious variables which can be used to describe the interrelations of a larger set of (real) variables: these fictitious variables or factors are the hypotheses which the statistical method of factor analysis helps us to "invent." There can be very little doubt that factor analysis has played an important part in aiding in this task of "invention," and the writer does not wish to underrate the contribution to psychological theory which has been made in this way.

Yet when one looks over the list of factors found in the cognitive field (12), or in the field of conation and affection (1), one may be pardoned for concluding that most of these hypotheses which factor analysis has led modern writers to espouse are far from original, and may be found in writings free from any contamination with statistical procedures. If the true contribution of factor analysis has really been the "invention of hypotheses." then it would seem that other less laborious methods would frequently have given us hypotheses not essentially different from those emerging from this modern concourse of calculating machines. And if the rejoinder be made that hypotheses derived from careful statistical investigations of large-scale experimental materials are more valuable than hypotheses of the Schreibtisch-experiment type, the answer surely must be that the derivation of an hypothesis does not guarantee its value, but that what is required is some method of indisputable proof of the correctness of the conclusion. In other words, we are led back to the alternate function of statistics mentioned by Kelley, the proving of an hypothesis. Can factorial methods be used to prove, as well as to suggest, scientific hypotheses?

This question is taken up directly by Cattell. He points out that the psychologist, as a scientist, will want to find the set of factors which corresponds to a set of psychologically real influences because he is interested in understanding the psychological meaning of his predictions and because he is curious to gain truth for its own sake. "In that case he may (1) devise possible ways of overdetermining the analysis of the given correlation matrix so that only the one set of true factors will emerge, or (2) start from the opposite shore and propound, on psychological grounds alone, a hypothesis about what source traits are operative in the variables. Then he will see if these factors correspond to any of the possible mathematical factors found in the matrix" (1, p. 273). Cattell rejects the second, or hypothetico-deductive, type of method on two grounds. "... in the first place, personality study has so few other reliable avenues for arriving at, or even suspecting, the basic source traits, that hypotheses are likely to be erratic. In the second place, the mathematical solutions to any set of correlations are so numerous and varied that unless the hypothesis can be stated in very precise quantitative terms the 'proof' of it is easy-so easy as to be worthless" (1, p. 274). Cattell repeats his position in slightly different terms: ". . . we have

rejected one of the two major approaches normally approved by scientific method—namely, that of inventing a hypothesis about the particular factors expected and attempting to discover a factorization to match it—because in this field almost any hypothesis could be so 'confirmed.' Instead, we seek general guiding principles for the mathematical analysis itself which will lead to a unique solution" (1, p. 281).

It is the purpose of this paper to suggest a method of rotation of factor axes which will give a unique, invariant solution along the lines of the hypothetico-deductive method; in other words, we believe that Cattell dismisses too easily the most powerful instrument of scientific methodology so far devised, and advocates instead methods which we shall try to show to be in no way adequate substitutes for it. First, we shall turn to an examination of these "methods of overdetermining the analysis of the given correlation matrix," and to a review of the results which may be expected from the use of these methods; then we shall describe the principles on which the method of criterion analysis was devised, in an attempt to get over the difficulties pointed out by Cattell; the principles will be discussed by reference to a worked example to show the application of this new method to a concrete problem.

# II. PRINCIPLES OF FACTOR ROTATION

Cattell lists seven principles for determining the choice of factors: (1) Rotation to agree with clinical and general psychological findings; (2) Rotation to agree with factors from past factor analyses; (3) Rotation to put axes through the centre of clusters; (4) The principle of orthogonal additions; rotation to agree with successively established factors; (5) The principle of expected profiles; rotation to produce loading profiles congruent with general psychological expectation; (6) The principle of "simple structure" relative to the given correlation matrix; (7) The principle of proportional profiles or "simultaneous simple structure." These seven principles may in our discussion be reduced to two: rotation where there is an outside criterion, and rotation where reliance is placed exclusively on statistical properties of the correlation matrix.

When there is an outside criterion, there are many different ways of making use of the criterion; these all reduce ultimately to the most simple and direct-inclusion of the criterion in the correlation matrix, and subsequent rotation of factors in such a way that all the common-factor variance of the criterion score is taken up by one factor, which is then identified with the principle of classification underlying the criterion. As an example of this approach, we may quote the correlational analysis by Cox of Rorschach scores, taken on 60 normal and 60 neurotic children matched for age, sex, and I.Q.<sup>1</sup> Each score was also correlated with the normal-neurotic dichotomy, which thus became the criterion score in the matrix of intercorrelations. After factorization, axes were rotated in such a way that all the common-factor variance of the criterion score was taken up by one factor, thus identified as "neuroticism," leaving only zero projections for this item on the remaining factors. While this method has certain advantages, it suffers from two great and fundamental drawbacks. In the first place, if the correlations between the individual scores and the criterion score are significant (if, in other words, several test scores discriminate significantly between the criterion groups), then it follows in-

<sup>1</sup> This study was carried out under the writer's direction, and will be published in due course. It should be noted that our comments on this research are not intended as criticisms, as Cox's purpose was not identical with that which we are discussing.

evitably that a factor should be produced from the intercorrelations of the scores which, when rotated in conformity with the principle outlined, would have high loadings for the criterion score. In other words, it is doubtful if the factorial approach adds anything of fundamental scientific interest that would not be given equally well, and more quickly, by some form of multiple regression or discriminant function analysis.

The other objection is even more fundamental; the procedure outlined begs the question which is really the ultimate justification for the factorial quest. We assume that the criterion groups are situated along a continuum which constitutes what Cattell calls a "source trait"; this fundamental assumption cannot be proved by means of the procedure described here. The fundamental assumption that neuroticism is a source trait remains an assumption, and if, as the writer believes, it be true that the main raison d'être of factorial methods lies in their ability to prove or disprove fundamental taxonomic questions of this kind, then clearly the method of rotation through an outside criterion is not of great general importance. It assumes, as do all the orthodox statistical methods dealing with the significance of differences, or the maximizing of such differences (analysis of variance and covariance, discriminant function analysis, multiple regression, etc.), that the main dimensions, or source traits, have already been located, and that contrasting groups, representing extremes along these dimensions, have already been lo-Factor analysis alone sets out cated. to discover which are these main dimensions, and it is precisely this feature which constitutes its claim to serious consideration.

Having rejected the method of external criteria, we must now turn to the method of internal criteria, *i.e.*, the

methods of simple structure and of proportional profiles. There is such a large . body of discussion dealing with these principles that we shall merely indicate with extreme briefness why we consider that they also fail to solve the problem which factor analysis sets out to attack. We shall not enter into such points as the question of invariance or uniqueness of the solutions offered; it is realized fairly widely that simple structure solutions are not unique (i.e., different psychologists analyzing the same matrix would not emerge with identical solutions), and they have not been proved to be invariant. These matters are vitally important, of course, but we would lay stress rather on a different aspect of these proposals which appears even more open to criticism. Thurstone and Cattell make the assumption that if some "general guiding principles" could be arrived at from an analysis of the matrix of intercorrelations itself, then we would be ensured of finding "factors corresponding to realities." Thurstone phrases this point rather differently by saving that only when factors are rotated in conformity with his principles do they become "psychologically meaningful." It is this fundamental assumption underlying the work of both Thurstone and Cattell which appears doubtful to the present writer. At the very least it would appear to require some form of proof; clearly as a principle it is not self-evident, as is shown by the fact that many experts, Burt and Thomson among them, have expressed views seemingly in contradiction to it. Yet no such proof is attempted, nor is it at all easy to see precisely how it could be given. It would almost seem as if the principle of simple structure, and that of proportional profiles, were engaged on a gigantic game of tautological huntthe-slipper, in which artificial statistical rules applied to a matrix or a set of matrices are supposed to give reliable

and valid information about real psychological influences.

### **III.** CRITERION ANALYSIS

Any statistical method of analysis is appropriate only to certain types of problems. The type of problem to which criterion analysis is appropriate may be described most easily by reference to an actual investigation. On the basis of a number of experimental and statistical investigations described elsewhere (2), the writer has advanced the heuristic hypothesis that there exists "a general factor of neuroticism, similar in mode of derivation and general interpretation on the orectic side to the general factor of intelligence on the cognitive side." This hypothesis assumes the existence of some strong, innate tendency predisposing individuals towards definite degrees of emotional adjustment or maladjustment, maturity or immaturity, neurotic or non-neurotic reactivity to environmental stress. It also assumes that the amount of environmental stress suffered by any given individual will affect the likelihood of his actual breakdown. We are not concerned here with the relative contribution of heredity and environment to neurotic maladjustment; what does concern us here is the hypothesis that this putative factor of "neuroticism" forms a quantitative continuum at one extreme of which are to be found hospitalized neurotics, while so-called normals are to be found all the way from the near-neurotic and neurotic to the conspicuously non-neurotic, mature, stable and integrated type of personality.

A second heuristic hypothesis was also advanced, again on the basis of various empirical investigations, to the effect that within the general field of temperament, a general factor of extraversion-introversion could be found which was orthogonal to the factors of intelligence and neuroticism, and

which found its prototypes in the neurotic disorders known as hysteria and dysthymia (psychasthenia, neurasthenia, anxiety neurosis) respectively. While this scheme of organization was based on factorial studies, it was recognized that the researches reported did not contain any definite proof regarding the feasibility of the assumptions made. The possibility could not be ruled out that certain qualitative differences existed between normal and neurotic groups, for instance, which gave rise to differences in test scores between these groups; if this were so, the assumption of a quantitative continuum would clearly be untenable.

A similar problem arises in conjunction with a much more widely held theory, namely, that associated with the name of Kretschmer (11). This author believes that there exists a normality-abnormality continuum whose one extreme is not the neurotic, but the psychotic; he also posits that the main factor in the temperamental field finds its prototypes in the main functional psychoses (schizophrenia and manic-depressive insanity), rather than in the neuroses. Instead of a taxonomic system based on neuroticism and extraversion-introversion, Kretschmer therefore has an entirely different system based on psychoticism and cyclothymiaschizothymia. The present writer has outlined Kretschmer's position at length elsewhere (3), and is publishing experimental evidence regarding the adequacy of his theoretical position; here he only desires to draw attention to the fact that both theories (Kretschmer's and the writer's) cannot be right, although they may well both be wrong, and that consequently some form of proof becomes indispensable. Kretschmer has attempted such a proof, which has much methodological interest; the writer has examined it elsewhere, and does not wish to repeat his arguments here; the conclusion arrived at was that this alleged proof really leaves the issue indeterminate. Clearly what is wanted is a deduction from the hypotheses presented which can be tested by means of statistical procedures; a deduction, needless to say, which is sufficiently precise to avoid an equivocal answer.

Deduction 1: The type of deduction on which our method relies may be illustrated by reference to a hypothetical example in which we are dealing with the "neuroticism" factor, and two tests,  $T_1$  and  $T_2$ , which discriminate significantly between a normal and a neurotic group. (We shall leave aside for the time being a consideration of the question of how these groups are chosen, or of problems of sampling which arise.) On the hypothesis that neuroticism is a continuous variable, a "normal" group would include persons differing in degree of "neuroticism." Now clearly the more highly "neurotic" subjects in the normal group should have higher "neurotic" scores on the two tests than the less highly "neurotic" subjects (we are assuming here that the tests have a threshold and a ceiling sufficiently far apart to allow differentiation at all levels of "neuroticism," an assumption which will be discussed below). It would follow from this argument that on the average  $T_1$  and  $T_2$  should be correlated in the normal population, a deduction which can easily be verified. (This correlation should of course be purified of the effects of irrelevant factors, such as intelligence, etc.)

This suggested proof could of course be extended to any number of tests; if a battery of n tests discriminates between normal and neurotic subjects, then on the basis of our hypothesis we should expect all the intercorrelations between these tests within the normal group to be positive on the average. We may, however, go further than this and add another specification which also follows directly from our hypothesis. This specification takes into account, not only the fact that our n tests discriminate between normals and neurotics, but also the additional fact that they do so with widely differing success.

Deduction 2: Let us correlate each of our n tests with the criterion, *i.e.*, the normal-neurotic dichotomy, by means of biserial or tetrachoric correlations; we thus obtain a criterion column  $(C_N)$ consisting of the correlations of tests  $T_1, T_2, T_3, \ldots T_n$  with the criterion. Let us next take the table of intercorrelations between the n tests for the normal population only, and submit it to a process of factorization, using either Burt's summation method or Thurstone's centroid method. This will result in the reduction of the large original table of correlations to a small number of factors in terms of which the original correlations can be reconstructed.<sup>2</sup> The actual factors found are purely arbitrary, as their position in the factor space depends on the original selection of tests, conditions of univariate and multivariate selection of the population, and other considerations of a similar kind. Our suggestion for deriving a unique, invariant, and psychologically meaningful solution of the problem posed by this fortuitous structure of factor positions is to rotate the first summation or centroid factor into a position of maximum correlation with the criterion column, C<sub>N</sub>.

The reasoning behind this suggestion follows directly from the two-test example given earlier. If the fact that two tests discriminate between the criterion groups results in a correlation between the two tests, then clearly the greater the discrimination effected by a test, the higher (*ceteris paribus*) the correlations

 $^{2}$  Only significant factors should be included in the analysis, of course, using one of the twenty or so available approximate criteria discussed by Vernon (13).

that test will show with other discriminant tests. Similarly, the lower the discriminative ability of a test, the lower will be its correlations with other discriminant tests. Now a factor is the expression of a pattern of intercorrelations existing between a set of tests: if a pattern of intercorrelations such as the one posited here exists in the matrix which is being analyzed, then it should be possible to arrive at a corresponding factor by suitable rotation of the arbitrary factors which emerge from the original analysis. The factor which embodies, as it were, our hypothesis, is of such a kind that the test which best discriminates normals from neurotics would have the highest loadings, while the test which least discriminated between normals and neurotics would have the lowest loadings; the other tests would be intermediate between these two extremes, having factor loadings proportional to the criterion column values. Rotation according to the principle of maximizing the correlation between factor and criterion column would enable us to discover to what extent the hypothesis was borne out; in this sense the principle suggested enables us to use factorial methods as part of the general hypothetico-deductive procedure,

It should be noted that if the hypothesis which is being tested is not borne out by the data, no amount of rotation would succeed in giving us any but a chance correlation between the criterion column and the rotated factors. In other words, the appearance of a high correlation between factor and criterion column may be interpreted as definite support of the correctness of the hypothesis; failure of such correlation to appear is proof of the incorrectness of the hypothesis. These statements are subject to a number of qualifications, which will be discussed below; also the single maximization principle outlined so far must be supplemented and extended by a double maximization principle to which we will turn later. At this stage we shall first give an example to illustrate our method before entering into further theoretical discussion.

# IV. SAMPLE STUDY

In Table 1 are given product-moment correlations between 16 tests which we had reason to believe measured the general factor of neuroticism hypothecated by us, to varying degrees of accuracy; the number of subjects is 64, all of whom were normal in the sense that they were not under psychiatric treatment at the moment, or had been under psychiatric treatment previously as far as could be ascertained.<sup>8</sup> Table 2 gives the first two factors extracted from this matrix by means of Burt's "summation method," grouping of tests being obtained from a preliminary application of Burt's simple summation. Reflection of signs was carried out using an external criterion, namely, the signs of the correlations in the criterion column described below. Reflection of signs was also determined by more usual methods from the matrix itself, using Burt's "inspection technique," and it was found that the two methods agreed in every case. The second factor is significant according to Burt's chi square method at the one per cent level of significance. (Residuals after the extraction of these two factors were insignificant.) Also given in Table 2 are the values for the criterion column, derived from 93 controls and 105 neu-

<sup>8</sup> I am indebted to my colleagues, Drs. Himmelweit, Petrie, and Desai, for permission to re-analyze data collected by them in their work on neuroticism at Dartford; additional data regarding the nature of the groups tested, and the tests employed, will be found in their original publication (6). I am also indebted to Mr. A. Jonckheere, of the Statistics Section of the Psychology Department, for carrying out the statistical work reported below.

IABLE 1	
---------	--

	A	в	с	D	E	F	G	н	I	J	к	L	м	N	о	Р
Α		.009	.043	.118	.157	197	019	049	008	042	017	.061	.229	.299	.095	.196
В	.009	—	.358	.418	051	.313	.344	.049	.104	.241	.131	.070	.220	.270	001	080
C	.043	.358		.680	.036	.375	.202	.378	.260	.248	.288	.178	.191	.128	.262	071
D	.118	.418	.680		004	.451	.271	.286	.173	.134	.301	114	256	238	221	008
E	.157	051	.036	004		.387	177	.007	.159	.010	.131	.062	.039	.025	072	058
F	197	.313	.375	.451	.387		.380	.327	.084	.222	.140	.045	.072	.060	.209	027
G	019	.344	.202	.271	177	.380		.062	122	.085	.035	040	.213	.208	.197	055
н	049	.049	.378	.286	.007	.327	.062	i —	.356	.031	.345	.261	.007	013	.154	.316
I	008	.104	.260	.173	.159	.084	122	.356	-	.252	.315	.346	.292	.294	.147	.029
J	042	.241	.248	.134	.010	.222	.085	.031	.252		073	.105	.253	.417	062	.059
к	017	.131	.288	.301	.131	.140	.035	.345	.315	073		.709	.394	.198	.186	.103
L	.061	.070	.178	.114	.062	.045	040	.261	.346	.105	.709	—	.450	.410	.167	.273
Μ	.229	.220	.191	.256	.039	.072	.213	.007	.292	.253	.394	.450	—	.562	.130	.201
Ν	.299	.270	.128	.238	.025	.060	.208	013	.294	.417	.198	.410	.562	-	.178	.327
0	.095	001	.262	.221	072	.209	.197	.154	.147	062	.186	.167	.130	.178	— I	.048
Р	.196	080	071	.008	058	027	.055	.316	.029	.059	.103	.273	.201	.327	.048	

Intercorrelations of 16 tests of neuroticism; for description of tests see text.

-

CRITERION ANALYSIS

\*

rotics, by means of tetrachoric correlation coefficients. (This normal group of 93 includes the group of 64 on whom our correlations in Table 1 are based; the reason why not all 93 were used for the intercorrelations is that a number of incomplete scores had to be eliminated.) In Fig. 1 are plotted the positions of the 16 tests with reference to the two factor axes.

Below is given a list of the tests used; the scoring in each case is indicated by emphasizing the direction in which the normal group scores as compared with the neurotics. Thus normals are more persistent, less suggesti-

	CN	$F_1$	F2	h²	D	Dι
A	.23	.143	.211	.065	,080	.127
B	.27	.392	220	.202	.256	.407
C	.51	.620	416	.557	.409	.650
D	.54	.044	438	.007	.425	.075
E	.06	.100	.089	.018	.059	.094
F	.10	.497	455	.454	.333	.529
G H I	.05 .57 30	.275 .405 438	397 078	.233 .170 222	.191 .258 267	.303 .410
J	.03	.300	.018	.090	.188	.299
K	.27	.523		.284	.324	1515
L	.17	.565	.461	.532	.333	.529
M	.46	.607	.384	.516	.363	.576
N	.26	.632	.430	.584	.377	.599
O	.24	.294	103	.097	.189	.300
P	.21	.207	.241	.101	.119	.189
		.203	.093	.296		

TUDUDE	TABLE	2
--------	-------	---

Explanation	of	Column	Heading:
-------------	----	--------	----------

- $C_N$  = Criterion Column, *i.e.*, correlation of each test with normal-neurotic dichotomy.
- $F_1$  and  $F_2$  = First and second unrotated factors from analysis of intercorrelations of normal group only.

 $h^2 =$ Communality.

 $\vec{D} = F_1$  rotated into maximum correlation with criterion column.

 $\hat{D}_{l} = \hat{D}$  with vector extended to unity

$$r_{\hat{D}\hat{D}_{l}} = .574$$
  
 $\rho \, \hat{D}\hat{D}_{l} = .587$ 

ble, more speedy, averagely perseverative, more fluent ideationally, have a higher personal tempo, are more flexible, have better motor control, better dark-adaptive powers, fewer neurotic symptoms on a questionnaire, and show less autism on level of aspiration tests. This attempt to present the tests from the uniform point of view of *normality* has forced us into certain awkward ways of phrasing which are difficult to avoid; in the case of intelligence tests this difficulty does not arise as the directionality of such tests is obvious, and recognized in common speech.

### List of Tests

- A. Maudsley Medical Inventory—40 item neuroticism questionnaire. Score = number of questions answered "No" (non-neurotic).
- B. Dark Adaptation—U. S. Navy Radium Plaque Adaptometer. Score = goodness of dark vision.
- C. Non-Suggestibility—body-sway test. Ability to resist suggestion to sway forward.
- D. Motor Control—absence of static ataxia; given as preliminary test to C.
- E. Goal Discrepancy Score--smallness of level of aspiration scores on O'Connor tweezers test.
- F. Judgment Discrepancy Scores—smallness of judgment discrepancies on O'Connor tweezers test.
- G. Index of Flexibility—number of shifts in aspiration scores on O'Connor tweezers test, irrespective of size or direction.
- H. Manual Dexterity-best score of nine trials on tweezers test.
- I. Personal Tempo—speed of writing 2, 3, 4, repeatedly for two trials of 15 seconds each.
- J. Fluency—number of *round things* and of *things to eat* mentioned during 30-second periods.
- K. Speed Test (1)—speed of tracing when instructed to be both quick and accurate. (Choice conditions.)
- L. Speed Test (2)—speed of tracing prescribed path on track tracer under instruction to be quick.

- M. Persistence Test I—length of time during which leg is held in uncomfortable and fatiguing position.
- N. Persistence Test B—holding breath as long as possible, without inhaling or exhaling.
- O. Stress Test—ability of S to recover previous scoring rate on pursuitmeter type of test after special stress period.
- P. Non-Perseveration—extremes of perseveration (SZ test), either very high or very low, are scored low, while scores nearer the average are scored high.

When the correlation matrix in Table 1 is considered, it will be seen that the coefficients arrange themselves in such a way that every single column total is positive; 4 in other words, we find that sixteen tests which differentiate between normals and neurotics intercorrelate (when correlations are run over the normal range alone) almost entirely positively; there is not a single negative correlation which is significant at the 5 per cent level in the table, while 39 positive correlations exceed this level, out of 120. (22 of these are significant at the 1 per cent level.) This result may be taken as a confirmation of our hypothesis, as such a large number of positive, and such a dearth of negative, coefficients is unlikely to have arisen by chance. No strict (non-parametric) test of significance is possible, unfortunately, as the correlations are not strictly speaking independent of each other.

It will be seen from Table 2 that the two factors extracted account for 30 per cent of the variance; the contributions being 20.3 per cent and 9.3 per cent respectively for the first and second factors. When we consider that the tests were chosen largely on a pri-

<sup>4</sup> The fact that every column total is positive would, of course, not be surprising if the signs had been determined by inspection of the matrix, but it will be remembered that the signs were determined independently by means of the criterion column.

ori grounds, and that almost half of them fail to correlate even at the  $r_t =$ .25 level with the criterion, these figures appear very promising. Thev would certainly be considerably higher if only the most discriminating eight tests had been used in the factor analysis; under those conditions, the percentage of variance accounted for would have risen above the 40 per cent level. Such an exclusion would be perfectly admissible, as it would take place on the basis of an extrinsic criterion, not of one intrinsic to the factor matrix; we have preferred, however, to retain all sixteen tests in our analysis. It is interesting to compare these results with those of a factorial study of ratings carried out on 700 neurotics by the present writer; the contribution of the first two factors was markedly lower than in the present case (2). In view of the fact that inter-personal variability is very likely much greater in a neurotic group than in a normal group, it might be concluded from this comparison that objective tests have a validity which may already be superior to that of psychiatric ratings of the kind used.

We now turn to the most crucial part of this analysis, the interpretation of the factors (Deduction 2). As explained above, the first step in this process involves a rotation of the first factor extracted into maximum correlation with the criterion column.<sup>5</sup> The rotation required is very small, amounting to only 5 degrees. This is understandable, in the present experiment, as the tests were specially selected to define one variable (neuroticism), so that

<sup>5</sup> More precisely, the two factor matrix was rotated in such a manner that the square of the differences between the loadings on the first factor and the criterion column values were made into a minimum. The resulting first factor of the new matrix was then extended to unit length.



.

the centroid or average of their intercorrelations would not be likely to fall far from the true value. When a larger number of factors is involved, or when a process of univariate or multivariate selection has entered into the sampling of the population tested, the first centroid axis may bear no such close relation to the rotated axis. This finding bears out the writer's previously expressed opinion that the first centroid factor is not necessarily lacking in invariance and psychological meaning, but that interpretation depends in each case on the circumstances surrounding it, and that no general rules of this kind can be laid down.

The correlation between the rotated factor and the criterion column is + 0.574. This value is in the expected direction, and while it is not large enough to be considered definitive proof of our hypothesis, particularly in view of the lack of a statistical criterion regarding its significance, it suggests that further work along these lines should ultimately lead us to a definite conclusion regarding the value of our hypothesis. When we take into account the fact that the great majority of the intercorrelation between the tests are positive, as pointed out above, and that the four tests which from previous large-scale work were known to discriminate particularly well between normals and neurotics (C, D, M, N) have the highest factor saturations in this factor, then we may feel a certain amount of confidence in the ultimate value of the hypothesis under investigation.

The study was not designed to test the hypothesis that an extravert-introvert factor would emerge in addition to the general neuroticism factor, there being no criterion column showing the correlations of the tests with this dichotomy. (Diagnoses of the neurotics enabling us to calculate such a column were not available.) It is instructive, however, to note that the grouping of the tests on the second factor is in conformity with what in previous investigations had been shown to be characteristic of the introvert-extravert (dysthymic-hysteric) dichotomy. Thus introverts (dysthymics) have been shown to be more persistent, extraverts (hysterics) to show less judgment discrepancy, and somewhat less suggestibility, as well as better dark-adaptation. However, not all the tests are in agreement with this hypothesis, and little emphasis is laid on the possible identification of this second factor.

A Thurstone-type rotation into a positive manifold is possible with our two factors, only one insignificant negative saturation higher than -.10 being required.<sup>6</sup> It is our considered opinion that this rotation is psychologically meaningless, and entirely lacks the uniqueness of our suggested solution. In addition, it is irrelevant to the hypothesis which is being tested. It would not be permissible to conclude that "simple structure" solutions would in all circumstances show these characteristics; it is for the adherents of the Thurstonian scheme to indicate more precisely when this scheme is applicable and when not.

# V. DISCUSSION

The example quoted above will have brought out certain problems which require discussion. The first of these relates to the question of significance; how can we determine the statistical significance of a maximized correlation between factor and criterion? The answer must be, at the moment, that a strict statistical criterion of significance exists here just as little as it does for the significance of a single factor saturation, or for the residuals on which a factor is based. In most cases it will

<sup>6</sup> This rotation is indicated in the figure, Axes I' and II'. be quite clear whether the original hypothesis is borne out or not; however, the absence of a proper method of evaluating significance must of course be regarded as a serious limitation of our method. It is possible, and indeed quite easy, to obtain empirical values which may serve to give an assessment of the chance values to be expected from maximizing correlations between factors and arbitrary criterion columns, and while such determinations cannot take the place of proper statistical derivations, they may help to tide us over until the more rigorous methods are available.

The second problem to be discussed relates to the presence of more than one hypothecated factor in a matrix. Let us assume that we had given a battery of tests purporting to measure both neuroticism and extraversion-introversion to a group of normal subjects: let us also suppose that the same tests had been given to a neurotic group equally divided into hysterics (prototypes of the extravert according to our hypothesis) and dysthymics (prototypes of the introvert). We would then be able to derive two criterion columns,  $C_N$  and  $C_{I-E}$ , containing respectively the correlations of each test with the dichotomy normal-neurotic, and with the dichotomy hysteric-dysthymic. We would then proceed to carry out a factor analysis of the intercorrelations of all the tests for the normal group, and rotate the first factor found into maximum correlation with the  $C_N$  column. The second factor would then be rotated into maximum correlation with the  $C_{I-B}$  column, giving either an orthogonal or an oblique angle with the first factor. If we insist on orthogonal relations between factors, we might prescribe that the two correlations between the two factors and the two criterion columns should be maximized simultaneously, giving equal weight to each. If we admit oblique relations, we might make the amount of obliqueness observed a test of the original hypothesis that neuroticism and extraversion-introversion are in fact unrelated. It is impossible to anticipate results in this matter, or in the even more complex problem arising with a larger number of criterion columns; one's decision will be determined by one's purpose, as well as by the exact nature of the data.

The two problems dealt with so far are problems of detail; the third one to be discussed now is one of principle. and of much greater importance. Let us assume that we have selected our neurotic population in such a way that it contains equal numbers of hysterics and dysthymics. Another experimenter might object that obsessional states ought to be included in the typical neurotic group, while yet another might wish to include psychopaths. It must be clear that if the inclusion of different syndromes is based entirely on the whim of the experimenter, we are still far removed from the essential objectivity of a proper scientific procedure. It is here that we may make use of the double maximizing principle. Stated briefly this principle demands that any addition to the main criterion group be permitted only if it increases, or leaves unaltered, the correlation between factor and C column. For example, if it were found that including a number of obsessionals in the criterion group significantly raised the correlation between the neuroticism factor and the  $C_N$  column, then we would have objective evidence that obsessional disorders belonged functionally with the hysteric and dysthymic disorders which made up the original criterion group. If psychopaths, when added similarly to the hysteric-dysthymic criterion group, failed to produce a higher correlation between the neuroticism factor and  $C_N$ . or even to maintain the existing correlation, then we would have proof that psychopathy did not belong functionally with the original criterion group.

This double maximizing principle, which is of course widely used in the physical sciences, should enable us to purify the criterion while still following the dictates of the hypothetico-deductive method. The hypothesis that a given clinical syndrome was a neurosis gives rise to the deduction that the addition of persons suffering from this particular syndrome to the existing (imperfect) criterion would raise the correlation between the criterion and the factor; this deduction is capable of being submitted to a crucial test along the lines discussed above. In this way we should be able, starting from a correlation between factor and criterion as low as the one observed in our experiment (r = .574), to improve our criterion successively through the addition of new groups until a substantially higher value was reached. This method is not open to objections on the score of subjectivism, and is an important and indeed indispensable corollary to our main principle of criterion analysis: clearly criterion analysis stands and falls with the adequacy of the criterion.

It may be asked whether the maximized correlation could and should reach the value of unity when assumptions are made about a perfect criterion and very large numbers of subjects and tests. It is here that we encounter our fourth problem. The correlation should reach the value of unity if all the tests used had a threshold below, and a ceiling above, the performance of the least and the most neurotic person respectively in the combined populations (assuming for the moment that we are dealing with the trait of neuroticism). If we take as an example the body-sway test of neuroticism, which has been shown to be highly discriminative in comparing normal and neurotic groups (2), we can see from published reports

that its threshold is well above the level of the least neurotic person in a combined group, while its ceiling is well below the level of the most neurotic (5). In other words, this test measures only over part of the range. Expressed in purely hypothetical units. we might find that if degree of neuroticism is expressed in terms of the numbers from 1 to 100, suggestibility tests might only be effective measuring devices in the region from 50 to 90, all subjects below 50 falling at or below the threshold of the test, and thus scoring the same mark, and all subjects above 90 falling above the ceiling, also all scoring the same mark. If, now, the dividing line between our normal and our neurotic population happened to be drawn at the 50 mark, it would follow that the discrimination value of this test would be extremely high, but its correlation with the other tests for the normal group alone would be rather low, as its variance would be extremely small and entirely due to error. While this example is of course exaggerated, it does indicate a very real difficulty, and certainly the regressions of each test used on neuroticism should be examined very carefully for linearity in order to safeguard against misinterpretation. There is little doubt in the writer's mind that in part at least the observed correlation's falling short of unity is due to the fact that many of the tests used have thresholds and ceilings which fall within the range of neuroticism measured.

The fifth and last problem to be discussed briefly will be the question of extending the principle of criterion analysis to slightly different types of investigation. As an example, let us take the question of the inheritance of "neuroticism." Along the orthodox lines one method of assessing the influence of hereditary factors would be the administration of a battery of neuroticism tests to monozygotic and dyzygotic twins, and a determination of the intertwin correlations within the two groups. These correlations would either be calculated for each test separately, or for some form of summed score. It is suggested that this traditional method does not enable us to extract all the valuable information inherent in the data, and that the use of a modified form of criterion analysis, with particular reference to the hypothetico-deductive method of reasoning, would give us additional information.

In brief, the hypothesis set up would be of the form: A general factor of "neuroticism" exists within the population tested, and is inherited as a multifactorial unit; in other words, "neuroticism" is not a statistical artifact, or an arbitrary "principle of classification," but a biological reality. Our method of proof would rest on the maximization of the correlation between the first factor extracted from the intercorrelations between the tests for our subjects, with a criterion column made up of values indicating the apparent influence of heredity on each test separately. Thus if Test 1 showed inter-twin correlations of .5 and .9 for dyzygotic and monozygotic twins respectively, while Test 2 showed correlations of .6 and .8, the values entered in the criterion column would be .56  $(=.9^2 - .5^2)$  and .28  $(= .8^2 - .6^2)$ . To express this method of proof in other words, the higher the saturation of a test with the factor under investigation, the higher (ceteris paribus) should be the relative influence of heredity on this test, provided the original hypothesis is correct. Thus again the correlation between factor and criterion column would serve as a crucial test of the hypothesis.

Many other uses of the principles suggested will be apparent, and there appears little use in discussing them in detail. While in this paper we have dealt exclusively with problems in the field of temperament, there is no reason why criterion analysis should be restricted in this way. The method has already been applied to problems in the field of social attitude measurement (4), and it seems likely that it might be used with advantage in the field of cognitive testing also. Thus, to mention but one example, the hypothesis that certain types of brain injuries, or certain types of therapy (E.C.T.), or certain types of disorder (Korsakoff's syndrome) lead to memory defects could be tested by matching patients falling into one of these groups with a control group equated for intelligence; testing both groups with a battery of memory tests of the kind developed by Ingham (8), and shown to define a factor additional to "g"; developing a criterion column in terms of the differences on these tests between controls and patients; intercorrelating the tests within the normal population; and rotating the first factor extracted into maximum conformity with the criterion column. Again, failure to find an acceptably high correlation would disprove the hypothesis; success in finding such a correlation would support it. Also it should be possible to compare the three criterion columns derived from three different experimental groups (brain injuries, E.C.T., Korsakoff patients), in order to show whether the "memory" defect shown in these cases could be considered to cover the same mental function. One virtue of such a triple comparison would be that only one control group would be needed.

# VI. SUMMARY AND CONCLUSIONS

The writer has tried to show in this paper that orthodox methods of factorial analysis are inadequate for genuinely scientific research because they reject the hypothetico-deductive method which is fundamental to all scientific work. The method of criterion analysis has been developed in an attempt to imbricate factorial analysis and the hypothetico-deductive method, and an example is given of the use of this new method in relation to the heuristic hypothesis outlined in the writer's book on Dimensions of Personality (2). Certain problems raised by the use of criterion analysis are discussed, and suggestions are made regarding the use of this method in a variety of circumstances. It is not claimed that criterion analysis would be a method useful for all types of problems, but it is maintained that for those taxonomic purposes which constitute the primary justification of factorial methods criterion analysis provides a scientifically acceptable and worth-while tool.

### BIBLIOGRAPHY

- 1. CATTELL, R. B. Description and measurement of personality. London: George Harrap & Co., 1946.
- 2. EYSENCK, H. J. Dimensions of personality. London: Kegan Paul, 1947.
- Cyclothymia-schizothymia as a dimension of personality. (To appear.)

- 4. ——. Primary social attitudes (II). (To appear.)
- S. —, & FURNEAUX, W. D. Primary and secondary suggestibility: an experimental and statistical study. J. exp. Psychol., 1945, 35, 485-503.
  HIMMELWEIT, H. T., DESAI, M., & PETRIE,
- HIMMELWEIT, H. T., DESAI, M., & PETRIE, A. An experimental investigation of neuroticism. J. Personality, 1946, 15, 173-196.
- HOLZINGER, K. J., & HARMAN, H. H. Factor analysis. A synthesis of factorial methods. Chicago: Univ. Chicago Press, 1941.
- INGHAM, J. An investigation into the relationship between memory and intelligence. Unpublished thesis, University of London, 1949.
- 9. KELLEY, T. L. Fundamentals of statistics. Cambridge: Harvard Univ. Press, 1947.
- Comment on Wilson and Worcester's 'Note on factor analysis.' Psychometrika, 1940, 5, 117-120.
- 11. KRETSCHMER, E. Körperbau und Charakter. Berlin: Springer-Verlag, 1948.
- THURSTONE, L. L. Primary mental abilities. Chicago: Univ. Chicago Press, 1938.
- 13. VERNON, P. E. How many factors? Admiralty Report, London, 1949.

[MS. received July 6, 1949]