Brit. J. Psychiat. (1978), **133**, 53–67

Multivariate Statistical Methods and Problems of Classification in Psychiatry

By R. F. GARSIDE and MARTIN ROTH

SUMMARY This paper is concerned with the use of multivariate statistical methods in relation to classification in psychiatry. There are two aspects of classification: the classification of symptoms to form syndromes and of patients into diagnostic groups. Critical papers are reviewed, and proposed severe limitations on the methods answered. The use of multivariate methods has increased objectivity and agreement between investigators, but differences remain, largely on issues in statistical logic, which must be resolved.

Introduction

It is a chastening thought that virtually all the diagnostic concepts in current use have resulted from the observations of clinical observers who were professionally active more than half a century ago. All were working within the traditional framework of medical diagnosis which had taken centuries to evolve. But one has to read only a few paragraphs of one of the classical descriptions to be made sharply aware of qualities of freshness, immediacy and originality that have since become more rare. These descriptions were no pious tributes to some dead and outworn technique for delineating states of mind. The great clinicians were the masters of the clinical and ethological methods; their descriptions survive, as all good and original descriptions survive, because they had arisen from clear hypotheses. No new description is likely to make its mark unless there is in the observer's head a new concept which corresponds to how things are in the world.

It is salutary to recall that schizophrenia, manic-depressive illness, paranoid psychoses, obsessional states, anxiety neurosis, hysteria and senile dementia were all described with the aid of clinical observations alone. There were no computers with which to establish the shape of distributions or search for symptom-sign clusters. Nor were there statistical methods to assist Kraepelin, Bleuler, Freud, Alzheimer; nor, in our own era, did these methods assist Leonhard, when he advanced the current influential view that bipolar affective disorders were distinct from unipolar ones.

These clinical descriptions, which have so robustly withstood the test of time, were ostensibly derived in the first instance from a few examples, or an unstipulated number, of the diagnostic categories. This was true of agoraphobia, of anorexia nervosa, and even of anxiety neurosis. But it is almost certain that the great figures of descriptive psychiatry were making use of a large body of previous, perhaps unconscious or undisclosed observations. They moved by inferences from the particular to the general. And although it is fashionable to deride such inductive forms of reasoning as unscientific, it seems highly probable that it is one of the ways in which important new hypotheses have come to be formulated.

Such hypotheses, however, are necessarily complex in that they involve numerous symptoms or other features. Moreover, the physiopathological causes of most psychiatric conditions have not been isolated. Thus, nearly all classificatory hypotheses in psychiatry are based entirely on symptoms (and perhaps other features). It follows that these hypotheses must be tested by multivariate statistical methods

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

which are capable of dealing with more than one symptom at a time.

Since Moore's (1930) classical paper, there have been many attempts to do this using multivariate methods, especially since general purpose digital computers have become available. But more recently there has been a tendency to suggest that multivariate statistical methods have severe limitations (Everitt, 1975; Maxwell, 1971), particularly with regard to the classification of patients. Such papers have had, we consider, an unfortunate and exaggerated effect. For example, Kraupl Taylor (1976) holds that 'statistical procedures are not well suited for the classification of objects', and he quotes Maxwell (1971) in support of this rather gloomy view. The purpose of this paper is to consider the use of multivariate statistical methods in relation to both (a) the establishment of psychiatric syndromes and (b) the classification of patients into diagnostic groups. Although this paper is concerned with these general topics, most of the actual discussion will be in relation to the classification of depressive illness. This is because it is in this area that multivariate methods have been used most often, and thus it provides the best psychiatric examples of the use of these methods.

Symptoms, Syndromes and Patients

In clinical medicine it is generally assumed that the isolation of a distinct syndrome means that there are patients whose symptoms in greater or lesser degree correspond with the symptom profile in question, and who consequently can be regarded as suffering from a particular 'disease' or syndrome (Scadding, 1967). A similar classification of psychiatric patients has been so predictive of prognosis and of the effect of relevant treatments that one wonders whether a categorical or typological model is not more relevant to psychiatry than a dimensional one. But the fact remains that in psychiatry the identification of a symptom cluster, or syndrome, does not necessarily imply that there is a corresponding group of patients distinct from other patients, or from people in general. Kendell (1968a), for example, isolated two syndromes of depression (neurotic and psychotic), but (in our view unjustifiably) considered that there were not two distinct corresponding groups of patients, but a continuous unimodal distribution—or dimension—of depressed patients from one extreme (neurotic depression) to the other (psychotic depression).

There are, therefore, two aspects of psychiatric classification: first, the classification of *symptoms* (and perhaps other features) into groups to form syndromes, a syndrome being defined here as a group of symptoms (and other features) which are correlated amongst themselves, but are relatively uncorrelated with other symptoms (or features).

The second aspect of psychiatric classification is the classification of *patients* to form diagnostic groups. By a diagnostic group we mean a category of patients whose symptoms (and other features) correspond with a particular syndrome. The existence of such a diagnostic group is demonstrated when the number of patients whose symptoms correspond to a particular syndrome is greater than the number of patients whose pattern of symptoms is ambiguous.

These two aspects of classification must be considered separately; the failure to do so has caused much confusion, as has been pointed out by Eysenck (1970) and Garside *et al* (1971).

The Classification of Symptoms

The classification of symptoms is relatively simple, and can be achieved by the well-known methods of factor analysis (Gorsuch, 1974). By using factor analysis it can be decided whether the correlations between the various symptoms are consistent with the hypothesis that there is one general factor, or dimension, underlying all the symptoms and other features in question. If the correlations are consistent with such an hypothesis, then there is no evidence of any grouping of symptoms; they are all indicative, to a greater or lesser extent, of a single psychiatric syndrome. But if the correlation data are inconsistent with such an hypothesis there must be an additional dimension, or dimensions, which are not general, since that which is general has already been accounted for by the general factor or dimension. The additional factor must, therefore, be differentiating or bipolar, and thus classifies the

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

symptoms and other features into groups corresponding, perhaps, to recognized psychiatric syndromes.

In this case, factor analysis is used as a *classificatory* procedure. But when a classification has been achieved, factor analysis can be used in a *descriptive* way by rotating the factors or dimensions so that the rotated factors describe, or measure, the two (or more) syndromes. Kendell (1968a), for example, rotated his factors so that one factor was descriptive of psychotic and the other of neurotic depression. There are thus two functions of factor analysis, one of *classification* and the other of *description.* There has been a regrettable tendency recently to publish only rotated, descriptive factors. But until a classification has been established, this seems to be begging the question. Moreover, as Mendels and Cochrane (1968) and Everitt (1975) point out, different methods of rotation may lead to different results. Unrotated factors are more stable, but are sometimes difficult to interpret.

The Classification of Patients

Cluster Analysis

There is a technique, namely cluster analysis, which provides a direct attack upon the problem of the classification of patients into groups. Cluster analysis attempts to divide a supposedly heterogeneous set of data (individuals) into clusters. It is a method of numerical taxonomy and is discussed at length by Sokal and Sneath (1963). The simplest method of cluster analysis consists of calculating measures of similarities between pairs of individuals and then grouping the individuals into groups according to levels of their similarity coefficients.

There are several computer programmes now available, and thus there is a danger that cluster analysis will be used without its limitations being realized. These limitations have been considered by Everitt (1972), who concluded that 'different techniques of cluster analysis applied to the same set of data may give very different results'. This conclusion is reiterated by Everitt (1974) and supported by Gower (1966, p 326). Thus, any classification generated by a single method of cluster analysis *must be validated.* As Everitt remarks (1972), 'It is obviously not enough to run out data on a single method of cluster analysis, however good this method may be thought to be, and however reasonable the clusters found seem to the experimenter'. More than one method should therefore be used, and the cluster only taken to be established if found by all, or the majority of, methods. This is the technique used by Everitt *et al* (1971). Other validation procedures suggested by Everitt (1974) are to carry out the analysis on two separate sub-samples of the data and to 'make predictions about the effect which the omission of some of the variables would have on the group structure and then to check that the predictions are verified' (Everitt, 1974, p 67). But the validation of a cluster analysis is best achieved by the finding of clusters in a further sample, using coefficients derived from the first sample.

There is, however, a fundamental weakness in cluster analysis which none of the validation methods proposed by Everitt can deal with. This is that 'the problem of deciding the 'best' number of groups into which to partition a set of data must really be considered unsolved and, in general, user evaluation is the criterion used. Hopefully some method will be developed, but until then users must keep in mind that 'groups' found may only be the extremes from a single distribution' (Everitt, 1972). This last point is of decisive importance. It means that such groups are not groups at all; there may be only one homogeneous group whose members form a continuous distribution. Moreover, Cormack (1971), in his review of classification, states that the ability of this procedure 'to find non-existent clusters is well established' and he condemns the growing tendency to regard numerical taxonomy as a satisfactory alternative to clear thinking.

It is quite clear that cluster analysis is limited to the generation of, as opposed to the establishment of, hypotheses regarding heterogeneity—at least until the present methodological problems are solved. Moreover, if one method of cluster analysis can generate spurious clusters, other methods may generate the same spurious clusters, particularly if such spurious clusters are merely the ends of an important dimension. Thus, the use of two or more cluster analyses on

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

the same data, as suggested by Everitt (1972), is no real guarantee that the 'clusters' found are genuine and distinct groups. Hence, before any conclusion can be drawn from a cluster analysis about heterogeneity, the distributions of scores along the dimensions which discriminate between pairs of clusters must be examined to see whether such distributions are unimodal or bimodal.

The dimensions used may be dimensions derived from a factor analysis, if the factors provide appropriate discriminating dimensions. If the factors do not, then a discriminant function analysis (see below) may be carried out by using pairs of clusters obtained, and the distribution of patients' scores examined to see whether the distributions are bimodal. If they are bimodal, then there is evidence that the groups are distinct. It should be noted, however, that the usual tests of significance for the discriminant function do not apply, as Paykel (1971, p 283) points out. This is because the discriminant functions which display the separation are based on the same scores that generated the groups.

Paykel (1971), in his important paper, argues that it is desirable that future studies concerned with psychiatric classification and grouping by mathematical methods should make increased use of cluster analysis. While not disagreeing with this, the limitations of cluster analysis indicated by Everitt (1972, 1974), Gower (1966) and Cormack (1971) must be borne in mind.

Discriminant Function Analysis

Discriminant function analysis, as Maxwell (1971) points out, is not a direct attack on the classification problem, for it requires an existing two (or more) group classification as a starting point. Given two groups, weights can be calculated so that the total weighted score ($T = \sum ax$) maximises the distance between the two groups. When there are more than two groups or clusters, multiple discriminant function analysis, sometimes called canonical variate analysis, can be used to find those canonical variates which maximise the ratio of between group to within group variance of three or more existing groups.

Discriminant function, and canonical variate, analyses involve the assumptions of normality of variables. But, nevertheless, as Maxwell (1971) states, they work reasonably well with Yes–No (0, 1) data, provided samples are large and item incidences are within the 20–80 per cent range. With 0, 1 variables, it is also necessary that their numbers should not be very small; one such variable, for example, must necessarily produce a bimodal distribution. Where no heterogeneity exists, however, the addition of a number of different (0, 1) variables will, provided that they are not highly correlated (Guilford, 1956, p 451), produce a unimodal distribution by the central limit theorem. Thus, if bimodality occurs, heterogeneity is established.

Winkel and Juhl (1971) have pointed out that in a clinical situation, the assumptions underlying some of the requirements of discriminant function analysis are usually not fulfilled, and they occasionally may 'fail to bring about optimal solutions of differential diagnostic problems'. This is clearly the case; but Winkel and Juhl do not suggest a better alternative. If a bimodal distribution of discriminating scores ($T = \sum ax$) *is* obtained, however, then it can be concluded that heterogeneity exists, regardless of whether the assumptions underlying the calculation of coefficients are met or not; the proof of the pudding is in the eating.

It is arguable that a discriminant function analysis is necessarily circular, in that it merely refines that which is already present and is therefore just a description of the diagnostic process used. This argument has a certain plausibility. Those symptoms which form the basis of the differential diagnosis will, of course, carry most weight. Moreover, the weighting coefficients obtained by using discriminant function analysis are derived by maximising the desired discrimination. The discrimination is reduced when the weights are applied to fresh data. Thus, before any firm conclusion can be drawn the weights should be applied to a new sample.

If possible, to avoid observer bias, these fresh data should be collected by other, presumably unbiased, investigators. An example is given later, when we consider the question of observer bias. If a significantly non-unimodal distribution

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

is then obtained, it seems safe to reject the hypothesis that the population distribution is unimodal, and thus to conclude that distinct groups of patients exist.

Classification of patients by factor analysis

Maxwell (1971) discussed the use of factor analysis in the classification of patients. He states that the factor analysis model is linear and additive. There is no question about this. But he goes on to say (Maxwell, 1971, pp 122–3) 'The variables themselves, taken in pairs, are assumed to be bivariate normal or, in the case of binary variables, to have incidences approaching 50 per cent (the 20 per cent–80 per cent range is generally satisfactory). Fulfilment of these assumptions is the experimenter's guard against spurious or biased correlations.

'If the observed variables are normally distributed then the factors obtained from them, being weighted sums of them, may confidently be expected also to be normally distributed. Indeed the factors as weighted sums will tend strongly to normality, even if the variables themselves are not normally distributed, in view of the *central limit theorem*. Consequently, if an experimenter claims to have found a bimodal distribution of factor scores, his analysis is doubly suspect—in the first instance because of the central limit theorem, and in the second instance because of the assumptions under which correlations are estimated reliably'.

Maxwell (1971, p 122) also states that the population sampled must be homogeneous with respect to the items, or symptoms, in question. If the population sample is heterogeneous (that is, if there are two or more populations) then, he claims, such a situation would lead to quite spurious results. We disagree with these statements. But spurious heterogeneity arising, say, from different investigators may generate spurious results.

It is clear that if a bimodal distribution of factor scores is obtained Maxwell would *ipso facto* regard the factor analysis as suspect, because such bimodality indicates that (a) the variables, taken in pairs, are not bivariate normal, or that (b) the population sampled is heterogeneous, or, of course, both.

This view is far from being universal. A number of eminent authorities adopt an exactly opposite position. For example, Cattell (1965, p 420) states 'There is no objection to the use of factor analysis as such with populations that are heterogeneous and remote from normal distributions on the variables'. Rummel (1970) takes a similar view; he states . . . 'Factor analysis, as a general technique, has been criticised on three grounds. The first is that the data must have an underlying multinormal frequency distribution, or at least must be measured on an interval scale. Both of these objections are based on misconceptions. Again, factor analysis can be applied to the data of any matrix. A multinormal, or near multinormal, distribution is required only when tests of statistical significance are applied to the factor results. As to measurement, factor analysis can be meaningfully applied even to nominally scaled data of a yes-no, or presence-absence type, the lowest and least demanding rung on the measurement ladder'. Moreover, Gorsuch (1974, p 14) states . . . 'While various procedures for finding the factors make additional assumptions, *the basic factor model assumes only that the variables are additive components of the weighted factors*" (his italics).

The clue to the difference between Maxwell, on the one hand, and Cattell, Rummel and Gorsuch on the other, lies in the fact that 'when questions of statistical inference arise—regarding the number of common factors or the significance of factor loadings—specific assumptions on the distribution functions of the factors and the observed variable are introduced' (Harman, 1967, p 16). But this is clearly insufficient reason for demanding normality and homogeneity before a factor analysis can be carried out at all. It is true that there are one or two methods of factor analysis, for example Lawley's maximum likelihood method (Lawley and Maxwell, 1963), which involve the assumption of normality. But, as Cattell (1965) points out, this is not, in general, the case. If, however, the importance of any factor, crucial to the point at issue, is in doubt, then tests of significance become essential and assumptions of normality must be made, whatever method of factor analysis is used.

But in neither of the Newcastle studies on

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

depression which Maxwell (1971) cites (Carney *et al*, 1965; and Kiloh and Garside, 1963), nor in Kendell's (1968a) study, was there any doubt as to the importance of the crucial bipolar factors.

As Eysenck (1970, p 242) remarks in relation to these studies . . . 'Thus factors of one kind or the other might be on the borderline of significance; there might be arguments about the respective size of the contributions different factors made to the variance; or the actual pattern of the items and their loadings might only partly correspond to the theoretical pattern expected on the binary hypothesis. In actual fact, these difficulties have not arisen: all the protagonists have found a clear-cut positive answer to the question put—the unitary hypothesis is wrong, and the binary hypothesis is supported very strongly indeed'. In this kind of clear-cut situation there is no need to carry out tests of significance. But if they are carried out, the results are bound to be so striking that they will support the obvious conclusions, even if assumptions regarding normality are not met. For example, the level of significance obtained by Kiloh and Garside (1963), that there was more than one kind of depression, was less than one in several thousands.

Neither Kiloh and Garside (1963), nor Carney *et al* (1965) calculated distributions of factor scores. Hamilton (1959), who is also quoted by Maxwell, did not calculate such a distribution. Kendell (1968a) did, but obtained a unimodal distribution. What Carney *et al* (1965) did do was to calculate a discriminant function based upon diagnostic groups, having shown by principal component analysis that the differential diagnosis was meaningful. So, in any case, Maxwell's criticism does not apply to the few papers he cites.

Principal component analysis

If the main object of the analysis is to begin with the observations and look for components in the hope that we may be able to reduce the dimensions of variation, and also that the obtained components may be given psychological meaning (Kendall, 1961, p 37), then tests of significance are not so important. Indeed, if Kaiser's (1960) criterion is adopted, such tests can be dispensed with altogether. This procedure of working from the data towards a hypothetical model, however, is usually called *component* analysis, and not factor analysis.

In factor analysis, on the other hand, 'we work the other way round; that is to say, we begin with a model and require to see whether it agrees with the data and, if so, to estimate its parameters' (Kendall, 1961, p 37). As Kendall goes on to remark, the distinction between component and factor analysis is often blurred in practice, and it is not always easy to classify examples into either component or factor analysis. Moreover, if the number of variables is large (say 15 or more), the numerical difference between a component analysis (with unity in the diagonals of the correlation matrix) and factor analysis (with less than unity on the diagonal) is trivial (c.f. Hope, 1968, p 61). Gorsuch (1974, p 32) takes the view that the component model and factor model 'are only variants of the general linear model. The basic characteristics of factor analysis derive from the general model and the objective of factoring, not from differences in the variants of the general linear model'. Gower (1967) takes a similar view.

Thus, to maintain a rigid distinction between principal component and factor analysis is rather academic. Yet in component analysis (as even Maxwell agrees) no hypothesis need be made regarding the distribution of the original variables (Lawley and Maxwell, 1963, p 2), whereas in factor analysis, according to Maxwell (1971), the variables, taken in pairs, are assumed to be bivariate normal.

Having argued that a bimodal distribution of factors scores is self-negating and that a normal distribution of such scores is the only acceptable one, Maxwell naturally concludes (1971, p 124) 'a proper use of factor analysis does not lead to a useful classification of patients'. On the other hand, as it is valid to use *component* analysis to study the grouping of people and symptoms alike, it follows that it can lead to a useful classification of patients. For, as Moran (1966, pp 1168–9) points out, 'The basic theoretical problem is to find a discriminating function $T = \sum a_i x_i$, whose frequency distribution will demonstrate the existence of two different groups, if such groups exist. No mathematical theory for the construction of such a function

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

exists, and we are forced into using hit or miss methods. Since the general position of the n points in the N-dimensional space can be more or less described by the variance-covariance matrix, we start from this (or the correlation matrix) and use the method of component analysis to see how the points are distributed along the directions of the principal axis'.

Component analyses may, therefore, not *only* provide components which may have a psychological or psychiatric meaning in terms of symptom clusters or syndromes, but the same *meaningful* components may also be used to describe or classify patients. The distribution of patients' components scores may be unimodal (Kendell, 1968a) or may be bimodal (Garside *et al*, 1971; Fahy *et al*, 1969). If the distribution is bimodal, then a diagnostic grouping of patients has been legitimately achieved, but needs to be replicated on a fresh sample (see below).

As clinical psychiatric data are often best expressed in binary (0, 1) form, such variables assume special importance. There is, therefore, a particular significance in Gower's demonstration (Gower, 1966) that with (0, 1) data a principal component analysis is the direct statistical opposite of numerical taxonomy. For it is exactly equivalent to assuming that individuals are represented by points whose distance apart are proportioned to $(1 - S_{ij})^{\frac{1}{2}}$, where S_{ij} is a matching coefficient between people suggested by Sokal and Sneath (1963, p 125). It follows that a principal component analysis carried out on (0, 1) data may be used to classify people. Should any doubt remain, Slater's (1958) important paper should be consulted. He shows that there is a simple relationship between *variables* and *person* components, provided that the data are treated consistently—for example, by using standardized variable scores throughout. This is automatically achieved if the analysis is based upon correlations between variables; this is the method normally used.

There is, therefore, a great deal to be said for using component analysis in preference to factor analysis in psychiatric inquiries into classification, even though the results obtained are numerically much the same. First, the kind of statement Maxwell (1971) makes would not be made; no assumptions are necessary in component analysis about the distribution of the variables. Second, with (0, 1) data the relationship between component analysis and at least one method of numerical taxonomy is simple. Third, principal component analysis avoids difficulties associated with what figures to put in the diagonal cells of the correlation matrix; unity is used.

The danger of Maxwell's (1971) paper is that some people may not realise that he is discussing factor analysis and not component analysis. Maxwell might argue that if component analysis were interpreted as a factor analysis, his objections would still apply to component analysis. But this standpoint is at variance with that adopted by most authorities in the field.

It is becoming fashionable to rotate factors or components towards 'simple structure'. This procedure, it is true, generates descriptive dimensions, which pass through the origin and the centres of symptom clusters or syndromes. But it also may destroy the discriminating power of any of the bipolar axes, which usually do not pass through the origin and the centre of any symptom cluster. Thus it is essential that axes should not be rotated willy-nilly to simple structure by using, say, 'Varimax' or 'Oblimax' methods if differentiation between disorders is the objective of the exercise. On the other hand, if a graph has been drawn, it may become clear how the axes should be rotated to arrive at an optimum discrimination (Garside, 1976). But this rotation is quite distinct from simple structure rotation.

Bimodality of the discriminating dimension

Before leaving the discussion of Maxwell's (1971) paper there is one crucial point which must be made. This is, as Moran (1966) has pointed out, that if the frequency distribution of patients' scores T ($T = \sum ax$) is clearly bimodal, this provides very strong evidence that the patients in question consist of two separate groups. Moran (1966, p 1168) goes on to say 'It is worth emphasizing that the evidence for such heterogeneity rests on the bimodality of the frequency distribution of the score and

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

not on the manner in which the latter has been found' (our italics). Thus, no matter how the T scores are calculated, that is, regardless of how the coefficients have been found, if a bimodal distribution is obtained, this is, *in itself*, evidence for the existence of distinct and discontinuous patient groups. It is therefore immaterial from the point of view of patient classification that such a bimodal distribution may render a factor analysis invalid, as Maxwell claims. On the other hand, Day (1969) has shown that dimensions can be found in reasonably sized samples, from multivariate spherical normal distribution, for which the distribution of scores appears to be bimodal. Thus, although Moran's statement is correct in so far as the evidence of bimodality in the scores is not dependent on the particular method by which the scores were derived, it is not unlikely that bimodality will be obtained when the differentiating scores are calculated from the same data as the score coefficients. Obtained bimodality from one sample should therefore be validated by using the coefficients to generate differentiating scores on another set of patients.

When, however, a differentiating score is generated by using data derived on the basis of *a priori* considerations, and a bimodal distribution is obtained, this does not need validating on a second sample. For example, Sandifer *et al* (1966) summed the scores of patients on fifteen items, commonly described as differentiating between psychotic and neurotic depression, and obtained a clearly bimodal distribution on the resulting scale, as did Garside *et al* (1971), using the same scale and merely adding the unweighted item scores (they also used weighted principal component scores and obtained a very similar resulting distribution). This simple summation procedure is not necessarily free from observer bias; this matter is discussed below.

Bimodality and Syndromes

In his important paper, Eysenck (1970) considers the classification of depressive illnesses, but the present paper only deals with those of his arguments which are of general methodological importance. Eysenck argues that bimodality is irrelevant to the question whether depression is unitary or binary, i.e. whether there are one or two syndromes of depression. He considers that bimodality is relevant only to the question whether depression or depressions are either continuous dimensions or categories (disease entities).

This argument would be sound if an assumption unjustifiably made by Eysenck were valid; namely that, in a representative sample of the total population at risk, endogenous and neurotic depressions would prove to be independent, that is, uncorrelated with each other. But such an assumption is for the present unwarranted. Eysenck makes it because he thinks that the investigators concerned have 'demonstrated that there are two independent axes, factors of causal principles involved in the distribution of depressive symptoms' (Eysenck, 1970, p 244).

It is true that Kendell (1968a, p 37), after a Promax rotation, which allows an oblique solution, found his two depressive factors to be 'virtually orthogonal' or independent. But this analysis was carried out in a population of depressed in-patients, that is, upon a very highly selected group. Data derived from such a highly selected sample of patients do not justify conclusions about the independence of two kinds of depression (or other illness) in the general population. Eysenck's assumption, therefore, begs the very question at issue—that is, whether the available data can be reconciled with a unitary concept of depression, or whether they favour the existence of at least two distinct syndromes. The interpretations that may be placed upon different distributions of clinical symptom scores will now be further considered by examining the implications of three different models (1) A categorical model (2) A dimensional model (3) A mixed model.

(1) *Bimodality and the distinction between types of illness: categorical model*

In the case of two illnesses (illnesses because we have hypothetically adopted a categorical model), if a sixth or more of patients are misclassified, the distribution of discriminating scores will be unimodal (Hope, 1969, p 733). Fig 1a makes this clear; the two shaded circles represent people with (say) endogenous or neurotic depression; people with neither are

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

shown as 'well'. If the axes are near together (if patients tend to be wrongly diagnosed) (Fig 1a), the distribution of patients along the discriminating dashed line will be unimodal. But if there is no overlap and the axes are orthogonal, or independent (Fig 1b), the distribution along the discriminating dimension will be bimodal.

Thus, given a categorical model, a bimodal distribution (on a discriminating dimension) signifies that the two syndromes can be distinguished one from another. On the other hand, a unimodal distribution has an ambiguous meaning. It cannot be construed as inconsistent with the presence of categorical entities. In the case of two hypothetical disorders, it may signify only that in a substantial number of individuals the two conditions cannot be distinguished from each other. A unimodal distribution would also be obtained if there were one, not two, illnesses in question.

(2) *The distinction between types of illness: dimensional model*

The case of the dimensional model will be considered next. If it is assumed, as Eysenck does, that severity of illness (or, more accurately, the probability of becoming a patient) is directly proportional to the number of symptoms, then the distribution along the discriminating dimension will be unimodal, whatever the correlation between the two underlying dimensions. Fig 2 makes this clear; the greatest frequency along either of the discriminating lines, cutting off the shaded areas (Figs 2a (independent) and 2b (correlated) are at their respective centres.

Thus, making Eysenck's assumption about the severity of illness, a dimensional model must produce a unimodal discriminating dimension, whether the two dimensions are independent (Fig 2a) or correlated (Fig 2b). But such a unimodal dimension is also produced, as we have seen above, in a categorical model, if the unitary hypothesis of depression (or other condition) is correct. Hence a unimodal distribution may be derived in a number of different situations, being consistent with a categorical and unitary, or a dimensional and unitary, or a dimensional and binary hypothesis. On the other hand, a bimodal distribution has an unequivocal significance. It is consistent only with the binary hypothesis and can be derived only when distinct entities are present.

(3) *Bimodality: mixed model*

A third model requiring consideration is one

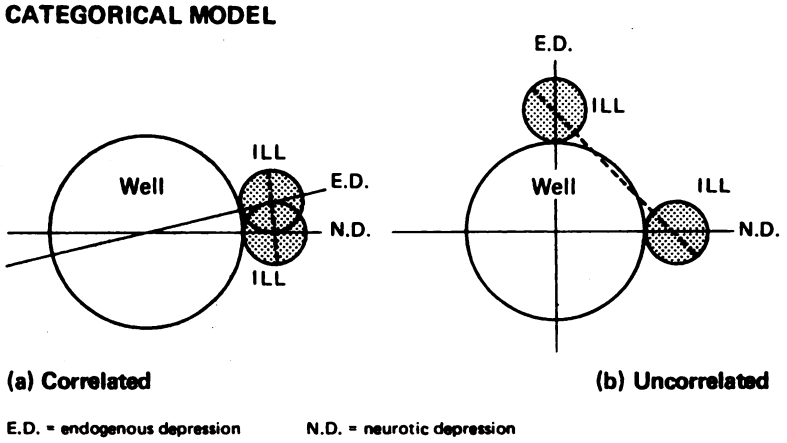


FIG 1.—The axis of the discriminating dimension would lie along the broken line in (a) and (b). The distribution of patients' summated scores would be unimodal in (a) and bimodal in (b).

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

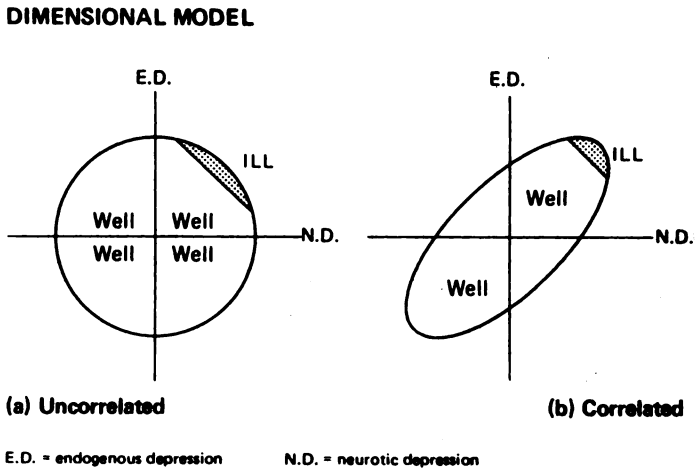


FIG 2.—The axis of the discriminating dimension is indicated by the diagnonal line both in (a) and (b). It is clear that the distribution of patients' summated scores plotted along this line would be unimodal.

in which one of the disorders under investigation is categorical while the other is continuous and dimensional. Thus, Fahy, Brandon and Garside (1969) concluded that depressions with 'endogenous' features probably constituted a categorical entity, while 'responsive, reactive' neurotic depression could be dimensional in character.

Everitt, Gourlay and Kendell (1971) recently drew a similar conclusion that 'perhaps the most useful form of classification will eventually prove to be a combination of a dimensional system with a typology'. Everitt *et al* (1971) reached this conclusion because they found separate clusters of patients with psychoses, whereas 'patients with depressive and other neuroses, with personality disorders and with alcoholism showed no tendency to form distinct clusters'.

Bimodal distributions could certainly materialise in the presence of such a blend of disorders under certain circumstances. But the precise form of the distribution would depend upon a variety of factors, such as the number of patients in each group and the presence of any correlation between the dimensional and categorical disorders. It is difficult, however, to visualise how bimodality of the discriminating dimension could possibly be consistent with a unitary hypothesis of depression or other illness, or with a wholly dimensional concept; a bimodal distribution indicates a categorical distribution of one or more axes.

(4) *The discriminating dimension*

Having argued that the shape of the distribution of the discriminating dimension is irrelevant to the unitary/binary question, Eysenck (1970, p 249) maintains that this dimension—running from reactive to endogenous depression—is 'inadequate statistically, and irrelevant psychologically'. So far as describing people are concerned, Eysenck is clearly correct; in a two dimensional space, one dimension must be statistically inadequate. One such dimension is also psychologically inadequate (but not, surely, irrelevant). It is necessary to know not only whether a patient is psychotic (endogenous) or neurotic, but also whether he is depressed or not and if so how severe his depression is. But *patients* have already been selected as being ill. Hence, if a bimodal distribution has been obtained on a discriminating dimension, it signifies that it is possible to arrive at a differential diagnosis in the

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

majority of patients. Hence, the shape of the discriminating dimension is clinically important, whatever its statistical and psychological status may be.

Moreover, having judged an individual to be suffering from a depressive illness it is not meaningless to ask whether he is psychotic or neurotic (categorical model) or more psychotic or more neurotic or both about equally (dimensional model). However, given a dimensional model it is more clinically descriptive to describe patients with reference to each of the two underlying descriptive dimensions, rather than in relation to the general and discriminating dimensions (cf. Garside, 1976).

The above discussion has been limited to two dimensions or categories. But as Fahy (1971) has pointed out, commenting upon the paper of Paykel (1971), factor analysis is by no means limited to dual dimensions or classifications. Such duality represents no more than a necessary first step in the analysis. For example, having classified affective disorders into two broad main groups—depressive illnesses and anxiety states—the next step is to investigate at a lower level in the hierarchy whether or not depressive illness and anxiety state can each be subdivided into distinct groups (Roth *et al*, 1972).

It should be noted that Eysenck (1970) does *not* argue that, if the discriminating dimension turns out to be bimodal, the original factor analysis is suspect. Although he regards the shape of this dimension, and therefore, the shape of underlying dimensions, as irrelevant for the unitary/binary question, he considers the results of factor analysis to have a close bearing upon it. But Maxwell (1971) takes a directly contrary view; he maintains that if an experimenter claims to have found a bimodal distribution of factor scores, his analysis is suspect.

Number of Subjects and Variables

Everitt (1975) states that any factor analysis in which the number of subjects is less than five times the number of variables 'should be viewed with a certain scepticism'. Burt (1952), however, related the number of subjects to the number of factors. He considered that 'with batteries of between 10–20 tests of the usual type, and with a summational method of analysis, at least 20 persons are needed to establish one factor, 50 to establish two, 100 to establish three, and between 200–400 to establish four'. Moreover, Burt did not consider that increasing the number of tests (or variables) decreased the number of significant factors.

Psychiatric data may well be less reliable than those derived from psychological tests. Hence, Burt's figures do not necessarily apply to factor analyses used to solve problems of psychiatric classification. But it seems reasonable to conclude, adopting Burt's views, that not less than 100 subjects are needed to establish two meaningful psychiatric factors or dimensions. This conclusion is based upon the correlations obtained in psychiatric and psychological research and upon Harman's (1967) Table A (p 434).

'Spurious' Correlations

Maxwell (1971) maintains that if heterogeneity exists this heterogeneity will (absolutely) increase the size of correlation coefficients obtained if these are calculated over the whole heterogeneous material. This is incontestable. But if the correlations so obtained are increased by heterogeneity, this increase is bound to be clearly reflected in any component (or factor) analysis carried out. As Maxwell himself argues, this is the reason why factor analyses should not be carried out on heterogeneous material. For these 'spuriously' large correlations will generate components or factors that do not exist within groups—that is, if the correlations within groups are zero. On the other hand, components or factors, which exist within groups, will be exaggerated in importance by the heterogeneity.

However, the use of correlation coefficients calculated over the whole heterogeneous material may, as Cattell (1965, p 421) points out, lead to the appearance of discriminating components, factors or dimensions, discriminating between and within groups, even though these correlation coefficients are 'spurious'. It is their very 'spuriousness' which may lead to the calculation of a discriminating score, $T = \sum ax$, that may prove clinically meaningful. On the other hand, when correlations are calculated

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

within diagnostic groups, and factor analyses are carried out on these correlations, rather strange results, from a clinical point of view, may be obtained.

For example, Maxwell (1972) carried out such 'pure' analyses within two diagnostic groups—schizophrenics and affective psychoses. He obtained three 'genuine' factors in the schizophrenic group—'retarded depression', 'general anxiety' and 'self-depreciation'. He also obtained the same three factors in the affective psychotic group, together with 'signs of anxiety'. He also found various factors which he regarded as largely 'artefactual in nature', being based upon 'spurious' correlations. These factors for the schizophrenic group included 'delusion of control' and catatonic symptoms, which were not found in the affective psychotic group. It seems that Maxwell's 'artefactual' factors, based upon 'spurious' correlations, may be more closely related to things as they are and more meaningful clinically than his more mathematically respectable factors.

Within group factor analyses can, however, indicate '*qualitative* differences in symptom and personality constructs across various nosological groups', as Derogatis *et al* (1972, p 398) have pointed out. They quote an example of different patterns of factor loadings obtained by two groups of patients:—anxious neurotics and depressed neurotics. This result, incidentally, lends support to the view that anxiety and depression are qualitatively distinct.

Random Error and Bias

Most psychiatric data are fallible, and thus nearly all psychiatric research involves random error. Fortunately, however, random errors rarely generate positive results; their effect is to attenuate rather than exaggerate. Moreover, tests of statistical significance are not in any way invalidated by the unreliability of psychiatric assessments. Bias, however, is a very different matter. Data which are biased can give rise to false positive conclusions. For this reason it can be argued that research should be replicated by investigators who are known not to be biased in the same direction as those who originally carried out the research in question. This has been done, for example, by Kendell (1968b), who used the diagnostic depressive scale of Carney *et al* (1965) and obtained a distribution which he described as 'perfectly unimodal'.

This conclusion illustrates a common logical error. The actual distribution (as given by Kendell and Post, 1973) is as follows: 1, 1, 3, 1, 10, 14, 14, 10, 18, 14, 8, 11, 10, 9, 5, 1. As the largest mode (18) is preceded by a frequency of 10, which is preceded by a frequency of 14, the claim that the distribution of the obtained sample is unimodal cannot be upheld. Even though it does not depart significantly from normality, the distribution is, in fact, non-unimodal, and as such is consistent with either a unimodal or bimodal hypothesis regarding the classification of depressive illness. The view that the forms of depression merge insensibly with one another constitutes a null hypothesis. As such it is not susceptible of proof, as Dalen (1972) points out in discussing the work of Popper.

This null hypothesis is inconsistent with the data of those who advance it: Garside (1973) added Kendell's (1968b) data to those of Post (1972) and obtained a significantly non-normal and non-unimodal distribution of the diagnostic scores of the resulting 222 patients. Kendell and Post (1973) then added a further 49 scores and obtained a trimodal distribution; accordingly, they declared it to be devoid of clinical meaning. This conclusion can hardly be accepted. When a non-unimodal distribution is obtained from such unbiased data, the view that the clinical features of depressive states are distributed in a unimodal manner stands rejected (Roth and Garside, 1973). Yet Kendell (1975, p 128, and 1976, p 18) still maintains that the study of Carney *et al* has not been replicated, and that the dimensional view of depression has advantages over the categorical one.

In a further attempt to avoid bias, one of us, together with Professor Kiloh (Kiloh and Garside, 1977), carried out a multivariate study of Sir Aubrey Lewis's data on melancholia (Lewis, 1934 and 1936). The conclusion from this multivariate analysis was that Lewis's data indicated there are at least two qualitatively different depressive illnesses and that one of

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

them, endogenous depression, is a categorical condition. These conclusions can hardly be due to observer bias on the part of Lewis.

Conclusion

Ever since Moore's (1930) paper, multivariate statistical methods, particularly factor or component analysis, have played an important role in sharpening the symptom profile of psychiatric syndromes. These methods have also made possible the differential weighting of symptoms and other features so that diagnostic categories, and clinical diagnosis itself, can be made more precise. The application of these methods has validated and refined the pioneering work of Kraepelin and the other great clinicians.

In using these multivariate statistical methods, investigators have been obliged to adopt uniform and objective analytical procedures resulting in a reduction of subjectivity. This objectivity has led to a considerable measure of concordance, for example the agreement between Kendell and the Newcastle workers that 'there are important and fundamental differences between different depressions that are not simply differences in severity or chronicity, as Mapother and Lewis had implied' (Kendell, 1968a, p 71). This kind of agreement, we venture to suggest, would never have been achieved without the use of appropriate statistical methods, in this case of factor analysis. Thus, psychiatric nosology has been placed upon a more solid foundation by the use of multivariate methods.

It is true that multivariate methods have not had much success in identifying new syndromes. But the dialogue between clinical classification and statistical investigations has brought to light new facets of old problems, for example the work of Paykel in relation to neurotic depression. Moreover, important issues have been raised, for example the role of personality features in psychiatric classification; such features figure (and, in our opinion, should figure) in the diagnostic profiles which have been delineated by statistical methods. Thus, the testing and refining of psychiatric diagnostic concepts and hypotheses, and the related matter of the prediction of outcome and treatment response by the use of multivariate statistical method, has been extremely fruitful.

Apart from the definition of syndromes, multivariate methods, such as discriminant function and component analyses, are clearly necessary in the testing of hypotheses regarding the differences between groups of patients and, when appropriate, in establishing adequate lines of demarcation between the groups in question. In this connection, the refutation of continuity hypotheses is of particular importance, as the recent controversy regarding the distribution of depression has illustrated.

The whole subject of the use of multivariate methods in psychiatric research is sufficiently important to warrant a large effort being made to use Monte Carlo methods to study the effectiveness of such techniques on artificial samples whose true population structure is known.

To return finally to the theme of this paper, there have been some unfortunate differences in interpretation of research findings which have arisen from differences concerning issues in statistical logic. It is important that these differences should be resolved.

References

BURT, C. (1952) Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section*, **5**, 109–33.

CARNEY, M. W. P., ROTH, M. & GARSIDE, R. F. (1965) The diagnosis of depressive syndromes and the prediction of ECT response. *British Journal of Psychiatry*, **111**, 659–74.

CATTELL, R. B. (1965) The role of factor analysis in research. *Biometrics*, **21**, 405–35.

CORMACK, R. M. (1971) A review of classification. *Journal of the Royal Statistical Society* (A) **134**, 321–67.

DALÉN, P. (1972) 'One, two or many?', in *Genetic Factors in Schizophrenia* (ed A. R. Kaplan). Springfield: Thomas.

DAY, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–75.

DEROGATIS, L. R., KLERMAN, G. L. & LIPMAN, R. S. (1972) Anxiety states and depressive neurosis. *Journal of Nervous and Mental Disease*, **155**, 392–403.

EVERITT, B. S. (1972) Cluster analysis: a brief discussion of some of the problems. *British Journal of Psychiatry*, **120**, 143–5.

—— (1974) *Cluster Analysis*. Heinemann Educational Books.

—— (1975) Multivariate analysis: the need for data, and other problems. *British Journal of Psychiatry*, **126**, 237–40.

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

—— GOURLAY, A. J. & KENDELL, R. E. (1971) An attempt at validation of traditional psychiatric syndromes by cluster analysis. *British Journal of Psychiatry*, **119**, 399–412.

EYSENCK, H. J. (1970) The classification of depressive illness. *British Journal of Psychiatry*, **117**, 241–50.

FAHY, T. J. (1971) Classification of depressed patients: a cluster-analysis-derived grouping (Correspondence). *British Journal of Psychiatry*, **119**, 116–7.

—— BRANDON, S. & GARSIDE, R. F. (1969) Clinical syndromes in a sample of depressed patients: a general practice material. *Proceedings of the Royal Society of Medicine*, **62**, 331–5.

GARSIDE, R. F. (1973) Depressive illness in late life (Correspondence). *British Journal of Psychiatry*, **122**, 118–9.

—— (1976) The comparative value of types of rating scale. *British Journal of Clinical Pharmacology*, Supplement, 61–7.

—— KAY, D. W. K., WILSON, I. C., DEATON, I. D. & ROTH, M. (1971) Depressive syndromes and the classification of patients. *Psychological Medicine*, **1**, 333–8.

GORSUCH, R. L. (1974) *Factor Analysis*. Philadelphia: Saunders.

GOWER, J. G. (1966) Some distance properties of latent root and factor methods used in multivariate analysis. *Biometrika*, **53**, 325–38.

—— (1967) Multivariate analysis and multidimensional geometry. *The Statistician*, Vol. 17, no. 1, 13–28.

GUILFORD, J. P. (1956) *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill.

HAMILTON, M. (1959) The assessment of anxiety states by rating. *British Journal of Medical Psychology*, **32**, 50–5.

HARMAN, H. H. (1967) *Modern Factor Analysis*. Chicago: University of Chicago Press.

HOPE, K. (1968) *Methods of Multivariate Analysis*. University of London Press.

—— (1969) Review of 'The Classification of Depressive Illnesses' by R. E. Kendell. *British Journal of Psychiatry*, **115**, 731–4.

KAISER, H. F. (1960) The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, **20**, 141–51.

KENDALL, M. G. (1961) *A Course in Multivariate Analysis*. London: Griffin.

KENDELL, R. E. (1968a) *The Classification of Depressive Illness*. Oxford University Press.

—— (1968b) The problem of classification. In *Recent Developments in Affective Disorders* (edited by A. Coppen and A. Walk). British Journal of Psychiatry Special Publication No. 2.

—— (1975) *The Role of Diagnosis in Psychiatry*. Oxford: Blackwell.

—— (1976) The classification of depressions: a review of contemporary confusion. *British Journal of Psychiatry*, **129**, 15–28.

—— & POST, F. (1973) Depressive illnesses in late life (Correspondence). *British Journal of Psychiatry*, **122**, 615–7.

KILOH, L. G. & GARSIDE, R. F. (1963) The independence of neurotic depression and endogenous depression. *British Journal of Psychiatry*, **109**, 451–63.

—— —— (1977) Depression: a multivariate study of Sir Aubrey Lewis's data on melancholia. *Australian and New Zealand Journal of Psychiatry*, **11**, 149–56.

LAWLEY, D. N. & MAXWELL, A. F. (1963) *Factor Analysis as a Statistical Method*. London: Butterworths.

LEWIS, A. J. (1934) Melancholia: a clinical survey of depressive states. *Journal of Mental Science*, **80**, 277–378.

—— (1936) Melancholia: prognostic studies and case material. *Journal of Mental Science*, **82**, 488–558.

MAXWELL, A. E. (1971) Multivariate statistical methods and classification problems. *British Journal of Psychiatry*, **119**, 121–7.

—— (1972) Difficulties in a dimensional description of symptomatology. *British Journal of Psychiatry*, **121**, 19–26.

MENDELLS, J. & COCHRANE, C. (1968) The nosology of depression: the endogenous-reactive concept. *American Journal of Psychiatry*, **124**, Supplement 1–11.

MOORE, T. V. (1930) The empirical determination of certain syndromes underlying praecox and manic-depressive psychoses. *American Journal of Psychiatry*, **9**, 719–38.

MORAN, P. A. P. (1966) The establishment of a psychiatric syndrome. *British Journal of Psychiatry*, **112**, 1165–71.

PAYKEL, E. S. (1971) Classification of depressed patients: a cluster-analysis-derived grouping. *British Journal of Psychiatry*, **118**, 275–88.

POST, F. (1972) The management and nature of depressive illness in late life: a follow-through study. *British Journal of Psychiatry*, **121**, 393–404.

ROTH, M. & GARSIDE, R. F. (1973) Depressive illness in late life (Correspondence). *British Journal of Psychiatry*, **123**, 373–5.

—— GURNEY, C., GARSIDE, R. F. & KERR, T. A. (1972) Studies in the classification of affective disorders: the relationship between anxiety states and depressive illnesses. *British Journal of Psychiatry*, **121**, 147–61.

RUMMEL, R. J. (1970) *Applied Factor Analysis*. Evanston: Northwestern University Press.

SANDIFER, M. G., WILSON, I. C. & GREEN, L. (1966) The two-type thesis of depressive disorders. *American Journal of Psychiatry*, **123**, 93–97.

SCADDING, J. G. (1967) Diagnosis: the clinician and the computer. *Lancet*, *ii*, 877–82.

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press

SLATER, P. (1958) The general relationship between test factors and person factors: application to preference matrices. *Nature*, **181**, 1225.

SOKAL, R. R. & SNEATH, P. H. A. (1963) *Principles of Numerical Taxonomy*. San Francisco: Freeman.

TAYLOR, F. K. (1976) The medical model of disease concept. *British Journal of Psychiatry*, **128**, 588–94.

WINKEL, P. & JUHL, E. (1971) Assumptions in linear discriminant analysis. *Lancet*, *ii*, 435–6.

Roger F. Garside, B.Sc., Ph.D., F.B.Ps.S., *Senior Lecturer in Applied Psychology, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU*

Sir Martin Roth, M.D., F.R.C.P., P.P.R.C.Psych., *Professor of Psychiatry, University of Cambridge (formerly Professor of Psychological Medicine, University of Newcastle upon Tyne)*

(*Received 19 January; revised 27 October 1977*)

https://doi.org/10.1192/bjp.133.1.53 Published online by Cambridge University Press