Multivariate Behavioral Research, 1980, 15, 475-496 #119

DETECTION OF BIOLOGICAL SEX: AN EMPIRICAL TEST OF CLUSTER METHODS

ROBERT R. GOLDEN
Columbia University
and
PAUL E. MEEHL
University of Minnesota

ABSTRACT

Six cluster methods were subjected to empirical trials evaluating their ability to solve a "dummy" problem, that of detecting an underlying taxonomy of biological sex. Each cluster method was applied to four sets of sex-discriminant self report personality-interest questionnaire items. The validity of each item-indicator was empirically determined using biological sex as a criterion variable. The four sets of item-indicators of biological sex were constructed so as to vary in their average item-indicator validity. Three of the cluster methods detected accurately the biological sex taxonomy in most of the trials; the other three methods were seldom accurate. It is argued that for detecting real empirical classes, the cluster methods are of questionable value since we typically lack assurance that the clusters are likely to be accurate and not spurious. It is suggested that use of internal validity or 'consistency' tests could eliminate this shortcoming.

INTRODUCTION

Suppose we have a substantive theory which is of a taxonomic nature and which requires an appropriate quantitative method in order to test it. Cluster analysis is the best known taxonomic method and, therefore, would often be considered for this purpose. Of concern to us in this paper is the obvious danger that while a cluster method or any other taxonomic method may normally produce "accurate" clusters, it will sometimes produce totally inaccurate or "spurious" ones. For our purposes, a set of clusters is accurate or has sufficient verisimilitude (Popper, 1962) if it corresponds closely enough to an actual underlying taxonomy of real empirical classes; if the set of clusters does not have such a degree of verisimilitude it will be said to be spurious. The purpose of the present study was to obtain an idea of the accuracy of some of the cluster methods when used to detect real empirical classes such as those considered when testing a typical taxonomic theory in the social sciences. The rationale for choosing the bio-

This research was supported in part by grants from the Psychiatry Research Unit and the Computer Center at the University of Minnesota. Requests for reprints should be sent to Robert R. Golden, Division of Biostatistics, School of Public Health, Columbia University, 600 West 168th Street, New York, New York 10032.

logical sexes is that they are known taxa with relatively discriminant indicators (MMPI items) and a clear criterion. The assumption is made that if a method fails to detect these known taxa, it is unlikely to be useful in detecting less understood taxa such as those encountered in psychological and personality theories.

This study evolved from an attempt to test a particular theory. Eighteen years ago Meehl (1962) proposed a theory of schizophrenia in which he hypothesized that only a certain class of people—those with a particular genetic constitution—have *any* liability for schizophrenia. The hypothetical class of individuals was referred to as the "schizoid taxon." While Meehl's theory has generated considerable interest, there has been little, if any, empirical evidence to either support or refute it until recently Golden and Meehl, 1979). In retrospect, the time lag was probably because the taxonomic nature of this theory engenders methodological and statistical problems which cannot be handled readily by standard taxonomic techniques including the cluster methods. While the study of schizoidia is not the topic of this paper we give a brief description of this research problem to illustrate the kind of real, empirical taxonomies we desire to detect with a taxonomic method.

Meehl (1973, also see Notes 1 and 2) summarized the general nature of the methodological and statistical problems encountered in testing this theory as follows: If the specific etiology of schizoidia is a single dominant gene, how can one estimate the probability that a person carries this gene when only highly fallible phenotypic indicators are available? There exists no acceptable criterion variable in the usual sense and no diagnostically definitive symptom or trait which we can measure reliably. A sample of individuals clinically *diagnosed* as schizoid or schizophrenic is typically somewhat heterogeneous. Thought disorder or "cognitive slippage," which is viewed by Meehl (following Bleuler) as the primary indicator of schizoidia, still is not sufficient by itself, perhaps because we cannot presently measure it well. Diagnosis of schizoidia presents a good example of the kind of problem described by Cronbach and Meehl (1955), in which we must start with a fallible set of indicators of unknown validities, and on the basis of some internal statistical relationships, hope to end up with accurate assignment of individuals, in this case as either schizoid or nonschizoid. Since the intended result is a selection of a subset from the candidate indicators, combined so as to be more construct-valid than the crude initial "criterion" (e.g., psychiatric diagnosis plus

indicator face-valid content), Cronbach and Meehl called it the "bootstraps effect" (pp. 286-287) to highlight the methodological paradox that we seem to lift ourselves up by our own bootstraps—a process that happens repeatedly in all fields of science.

In the present study we evaluated six of the more popular cluster methods as to their accuracy in testing a similar taxonomic "theory" (one which we knew in advance to be correct). In doing this, we attempted to determine which, if any, of these cluster methods are satisfactory for solving the pseudo-problem of detecting the taxonomic variable of biological sex when sex-discriminant Minnesota Multiphasic Personality Inventory (MMPI) items are used as indicators. It should be evident that taxonomic methods which can not pass this empirical trial are unlikely to be useful in detecting, as but one example, the schizoid taxon with (other) MMPI items.

In other words, our approach to evaluating cluster methods is that of determining how well each method detects a *known* taxonomy. General use of this approach is difficult in fields such as personality and psychopathology since there are few known taxonomies. Fortunately at least one physical taxonomy, biological sex, is virtually perfect on the criterion side. Also, it is reassuring to note that a scale of MMPI masculine-feminine interest items produces near bimodality for a mixed sample with equal numbers of males and females. We think it likely that many taxonomies in the social sciences have too much overlap to generate bimodality when psychometric indicators such as MMPI are used. If bimodality obtains, it is generally because the two means of the taxonomic class distributions on the indicator-scale are quite far apart. For example, if the two taxonomic class base-rates are equal, and the two taxonomic class distributions are normal in shape, then bimodality is only discernible when the two means are more than two within-taxonomic class sigma-units apart. We assume that unless a cluster method is able to detect with reasonable accuracy a known taxonomy (identified on the basis of an accepted criterion variable) with the use of relatively discriminant indicators, the method is not likely to be useful in detecting less obvious taxonomies such as the schizoid taxon.

Although cluster methods were first proposed by Zubin (1938), and Tryon (1939), general interest in their use paralleled the development of large computers. Now, according to Blashfield (1976), there are over 100 different cluster methods (*cf.* Ander-

berg 1973, Bailey 1974, and Everitt 1974). Even though they are most frequently used for generating clusters of related variables, the cluster methods can also be used for generating clusters of similar individuals. In this paper we consider only clusters of individuals since we are interested in detecting the biological sexes.

According to Blashfield (1976), the most popular cluster methods are the agglomerative ones. These methods are used in conjunction with a matrix of similarity-values for each pair of individuals' sets of indicator scores. The measure of similarity between the two individuals' indicator scores is usually some kind of correlation or distance in the indicator-hyperspace. From the similarity-values, the cluster methods produce clusters by assigning individuals with similar scores to the same cluster. A cluster method is iterative and generates a hierarchical tree, with each level of the tree representing a different clustering called a partition. If there are N individuals, then the first partition consists of $N - 1$ clusters, the next of $N - 2$ clusters, and so on until the last partition, which consists of two clusters. It is this last partition of two clusters which we compared with biological sex.

Four of the most popular agglomerative methods are called by Blashfield the 'single linkage,' 'complete linkage', 'average linkage' and 'minimum variance' methods. Two other cluster methods, mathematically related to the average linkage method are the 'centroid' and 'median' methods. It has been shown by Lance and Williams (1967) and Wishart (1969) that the six methods can be described in terms of the same algorithm or iterative procedure. To describe this algorithm, let the degree of similarity between two clusters i and j of individuals be denoted by d_{ij}. Suppose these two clusters of individuals have been combined to form a new cluster k. Then we denote a remaining individual by h and the distance between cluster k and individual h by d_{hk}. The general algorithm is

$$d_{hk} = Ad_{hi} + Bd_{hj} + Cd_{ij} + D|d_{hi} - d_{hj}|$$

where A, B, C and D are parameters whose values depend on the particular cluster method and are given in Table 1.

Table 1

Parameter Values for Each of the Six Methods for Use in the General Algorithm *

Method	A	B	C	D
Single Linkage	1/2	1/2	0	-1/2
Complete Linkage	1/2	1/2	0	1/2
Average Linkage	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	0	0
Minimum Variance	$\frac{n_h+n_i}{n_h+n_k}$	$\frac{n_h+n_j}{n_h+n_k}$	$\frac{-n_h}{n_h+n_k}$	0
Centroid	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	$-\frac{n_i n_j}{n_k^2}$	0
Median	1/2	1/2	-1/4	0

* Wishart (1969)

n_i: number of individuals in cluster i of preceding partition
n_j: number of individuals in cluster j of preceding partition
n_k: number of individuals in cluster i or cluster j of preceding partition (i.e. number in cluster of the present partition formed from the union of cluster i and cluster j)
n_h: number of individuals with a particular vector of indicator scores for which we wish to consider the distance to cluster k (d_{hk}).

This algorithm can be translated into words to give a sense of how a particular cluster method works and how it differs from the others. For example, in the single linkage method, each member of a cluster is more similar to at least one member of that cluster

than it is to any member of any other cluster. In the complete linkage method, each member of a cluster is more similar to the most dissimilar member of the same cluster than it is the most dissimilar member of any other cluster. In the average linkage method, each member of a cluster has a greater average similarity with the other members of the same cluster than it does with the members of any other cluster. In the centroid method the members of a cluster have, on the average, a greater similarity to the centroid of the cluster than they do to the centroid of any other cluster. The centroid of a cluster is the vector of indicator means calculated across the members of the cluster. The median method is similar to the centroid method except that the median of the cluster members is used in place of the centroid. In the minimum variance method, the clusters are formed so that the sum of the squared differences in the similarity-measures across pairs of individuals of the cluster is minimal.

Each of these six methods can be used with any of several measures of similarity between individuals. The measure of similarity used in this study was the average (across indicators) of the squared differences of the two individuals' indicator scores. When measuring the similarity between two individuals across several indicators, it is usually desirable to transform scores so that there are no differences in the indicator means or variances. In the present study we standardized each item-indicator so that each had a mean of zero and a variance of one for the total mixed sample (males and females combined).

Method

The accuracy with which each of six cluster methods detected the biological sex taxonomy, when used with sex discriminant MMPI item-indicators, was determined by observing how accurately individuals in a mixed-sex sample were correctly classified according to biological sex. The last partition of two clusters was used and the cluster with the most females was identified as the female cluster and the other cluster as the male cluster.

The MMPI item-indicators were chosen by comparing two samples of males and females on each of the 550 MMPI items. The two samples consisted of 430 male and 675 female adult psychiatric patients in the University of Minnesota Hospitals. The items were scored 1 for a "female" response and 0 for a "male"

response. The "female" and "male" responses were determined by comparing the response proportions of these same male and female samples and by considering the item content (i.e. face validity); the two methods agreed perfectly. It was found that 49 items discriminated between the two samples to the extent that the difference in the proportions that scored a 1 was .1 or more. This difference in proportions will be referred to as the 'validity' of the item. Of these 49 items, 18 were found to have validities of .3 or more and will be referred to as the 'highly discriminant' items. Examples of the highly discriminant items are: "I am not afraid of mice" (F) (The letter in parentheses indicates if a response of true (T) or a response of false (F) is scored as 1; otherwise the response is scored as 0). "I used to like hopscotch" (T), "I used to keep a diary" (T), "I very much like hunting" (F), "I like collecting flowers or growing houseplants" (T), "I would like to be a nurse" (T). The 12 items found to have validities between .2 and .3 are referred to as the 'moderately discriminant' items. Examples of moderately discriminant items are: "I like poetry" (T), "I like to cook" (T), "I would like to be a soldier" (F), "If I were an artist, I would like to draw flowers" (T), "I have no fear of spiders" (F). The 20 items found to have validities between .1 and .2 are referred to as the 'weakly discriminant' items. Examples of weakly discriminant items are: "I gossip a little at times" (T), "Sometimes when I am not feeling well I am cross" (T), "I would like to be a florist" (T), "At times I feel like swearing" (F), "I am certainly lacking in self-confidence" (T), "I am easily downed in an argument" (T), "I like science" (F). Finally, 26 items were selected at random from the remaining 501 items in the MMPI inventory with validity coefficients between $-.1$ and .1; and are referred to as 'non-discriminant' items. Examples of non-discriminant items are: "I have several times given up doing a thing because I thought too little of my ability" (T), "I have often met people who were supposed to be experts who were no better than I" (F), "At times I have worn myself out by undertaking too much" (F), "My plans have frequently seemed so full of difficulties that I have had to give them up" (T).

For each trial, a mixed sample of size 200 consisting of 100 males and 100 females was analyzed. A size of 200 is a common one for cluster analysis studies and was the maximum that the available computer program could accomodate. A second mixed

sample of the same size and mixture was used for replication trials.

The six cluster methods were studied using the following different sets of indicator-items: a) the 18 highly discriminant items and 2 moderately discriminant items (20 items), b) the 18 highly discriminant and 12 moderately discriminant items (30 items), c) the 18 highly discriminant, 12 moderately discriminant, 20 weakly discriminant items (50 items), d) all of the highly discriminant, moderately discriminant, weakly discriminant, and non-discriminant items (75 items). (See Table 2.)

Table 2

Description of the Four Item-Indicator Sets

Set	Total Number of Items	Number of Items in Each Range of Validity *			
		High (> .3)	Moderate (.2 to .3)	Low (.1 to .2)	None (-.1 to .1)
a)	20	18	2	0	0
b)	30	18	12	0	0
c)	50	18	12	20	0
d)	75	18	12	20	25

* The validity of an item is defined here to be the difference between the proportions of the females and males that respond to the item in the female direction.

RESULTS

Three measures can be used to describe the validity of the clusters. First, the female *cluster* baserate, that is the proportion of the total sample who were in the cluster which consisted mostly of females, was calculated (denoted by P in Table 3). Since each of the two mixed samples consisted of 100 males and 100 females, the actual sample baserate for each of the biological sexes was .5. For the clusters to be considered accurate the estimated female baserate P should be close to .5. Second, the correct classification rate or hitrate in predicting biological sex from cluster membership was calculated (denoted by H in Table 3). If there were perfect classification, the value of H would have been 1.00. If the cluster membership happened to be perfectly independent of biological sex for the sample used, then the value of H would have been .50. Third, in addition to H, we calculated coefficient kappa (see Cohen, 1960), another index of the agreement of the cluster-membership and biological sex. This statistic can be roughly described as the relative improvement in the correct classification rate of individuals through the use of the clusters over that obtained simply by the use of random assignment according to the observed baserates of the clusters and the biological sexes. For the parameters of the present study the *kappa* coefficient is equivalent to $2H-1$, and the range of possible values for this statistic is 0 to 1.00.

For results to be regarded as acceptable it was required that the baserate estimate P be within .10 of the actual sample value of .5 and that the hitrate H be at least .75 (or *kappa* to be at least .50). Results meeting these two criteria are indicated by an asterisk in Table 3.

Three of the methods (single linkage, centroid and median) did not provide acceptable values for the baserate P and the hitrate H for any of the four items sets. In each case, the absolute error in the estimate of P exceeded .15 or the hitrate H was less than .65 where .50 is the chance rate.

In contrast, the results of the other three methods (complete linkage, average linkage, and minimum variance) were often, although not always, quite good (see Table 3). The average linkage method worked slightly better than the other two methods and generally provided clearly acceptable estimates of P and H.

Table 3

The Estimate of the Base-Rate P, Correct Classification Rate H, and Kappa for the Three Most Accurate Cluster Methods

I. Complete Linkage Method

Set	Number of Items	Sample 1			Sample 2		
		P	H	Kappa	P	H	Kappa
a)	20	.35	.77	.54	.48	.92	.84*
b)	30	.52	.87	.74*	.59	.80	.60*
c)	50	.53	.75	.50*	.53	.83	.66*
d)	75	.52	.74	.48	.53	.73	.46

II. Average Linkage Method

Set	Number of Items	Sample 1			Sample 2		
		P	H	Kappa	P	H	Kappa
a)	20	.56	.85	.70*	.43	.86	.72*
b)	30	.50	.90	.80*	.42	.88	.76*
c)	50	.47	.76	.52*	.63	.85	.70
d)	75	.46	.75	.50*	.63	.65	.30

III. Minimum Variance Method

Set	Number of Items	Sample 1			Sample 2		
		P	H	Kappa	P	H	Kappa
a)	20	.65	.79	.58	.56	.60	.20
b)	30	.47	.90	.80*	.44	.79	.58*
c)	50	.55	.92	.84*	.55	.74	.48
d)	75	.58	.76	.52*	.59	.73	.56

Note: None of the other three methods (single linkage, centroid, median) ever met the joint criteria: kappa $\geq$ .30 and .35 $\leq$ P $\leq$ .65 which were met in all but one of the trials reported in the above table.

* clearly acceptable results: kappa $\geq$ .50 and .4 $\leq$ P $\leq$ .6

Discussion

We must emphasize that the major concern of this study was to estimate the tendency of several cluster methods to produce an inaccurate or spurious result when attempting to detect real empirical classes such as is done when testing a taxonomic theory. Cluster analysis can be used in many ways other than testing taxonomic theories. Some examples are hypothesis generation, exploratory study, clustering according to an algorithm which is accepted by definition and so on.

There is no generally accepted method by which any quantitative method can be evaluated. There are many methods of evaluation other than by empirical trial. For example, some psychometricians, including Lord and Novick (1968, p. 383), have suggested that psychometric methods should be evaluated in terms of their overall utility, i.e., in terms of the usefulness of the predictions made via the method. A limitation of this approach is that if utility fails to obtain, we do not know whether to blame the theory being tested, the data measurements used and/or the quantitative methods used. In areas such as psychopathology and personality measurement this dilemma is encountered frequently as utility is often lacking.

A case could be made that a partition other than the last one consisting of two clusters may have come closer to "carving nature at its joints," and that restricting our attention to partitions consisting of two clusters resulted in unfair tests of the methods. This argument is presumably based on a reluctance to accept biological sex as the main or only taxonomy underlying the responses to the MMPI items used. That is to say that there may have been several related (but etiologically different) taxonomies underlying the MMPI responses such as 'psychological sex,' 'sexual identification,' and 'masculinity-femininity interest,' and, as a consequence, the number of taxonomic classes need not be two. If biological sex is only one of the major underlying taxonomies, then the clusters may agree with other (equally real) taxonomies rather than with biological sex. However, we find it difficult to believe that there exist prominent taxonomic variables underlying the present indicator responses that are not highly correlated with biological sex.

We also believe that our choice of indicators and sample is almost certainly superior to those used in much taxonomic research in the social sciences, and that our dummy theory, that the under-

lying taxonomy is indeed biological sex, has much more evidence supporting it than most hypothesized taxonomies which are tested in actual research! We claim that if we are unable to confirm correctly our dummy theory that a biological sex taxonomy exists by using reasonably discriminant items, then, whatever the reason that we failed to do this, we cannot expect to do very well in testing a taxonomic theory such as one regarding schizoidia when using the same method. If the dominant factor underlying the items is not biological sex but something having more to do with, say, masculinity-femininity interests, then the last partition of the cluster results would not necessarily agree with the biological sex taxonomy. Since some of the clusters show little agreement with biological sex it would be necessary for the dominant factor to be weakly correlated with biological sex and it is hard to imagine what such a factor could be. Further, we have used candidate indicator items that are probably more valid than those tried in actual research, and having available sufficiently valid indicator items is a necessary component of a taxometric analysis.

Finally, some researchers who use cluster methods do not recommend that these methods be used for the detection of hypothesized empirical classes. They claim, for example, that the cluster methods are only useful for creating clusters which have certain mathematical properties (e.g., see Forgey, Note 3). This view seems to be based on the belief that the rationale behind the clustering algorithm must be simply accepted as it cannot be directly tested. However, we urge that if we have little evidence that the resulting clusters correspond to a real empirical taxonomy and know only that they have some sort of mathematical properties, then they would be of little use to us for research attempting to test substantive theories. We wish to use a taxometric method for testing a conjectured taxon such as the schizoid taxon described above. In the present paper we have attempted to evaluate some of the cluster methods for *this* purpose.

The most striking result of this study was that our dummy theory was correctly confirmed by use of some of the methods, item sets, and samples, but not for the others. Three of the cluster methods usually produced results which would lead us to confirm our dummy theory and three of the methods did not. There was similar variability in the accuracy of the results over item sets and subject samples. Since nothing within the cluster methods tells us *which* results are likely to be accurate and which are not, the

present trials suggest that we do not know how to select a cluster method and a set of indicators with sufficient confidence so that the existence of hypothesized empirical classes can be properly tested.

It should be noted that in exploratory or 'context of discovery' (Richenbach, 1938) the clusters found may have been valid enough to begin a bootstrapping process that would eventually lead to discovering the biological sexes. While this is true, in the present paper we are concerned with using the cluster methods in the 'context or justification' or theory testing research.

It has been pointed out to us by one reviewer that the cluster methods tested in this study are not the most appropriate for a dichotomous taxonomy. It was suggested that the non-hierarchic methods would be more suitable for this purpose. We have only tested the methods reported in the literature to be the most popular ones. Other cluster methods may work better than those tested here. It should be noted that the average linkage method worked quite well for the present data and the main result of this study is that we do not know which cluster method to pick for a particular theory testing application.

For each of the methods we are without any means of distinguishing spurious or inaccurate results from those that are substantially correct. As a result we conclude that one cannot be highly confident of these cluster methods when used in a theory testing context. However, there appears to be at least one approach that could be used to improve the situation. It is noted that the postulates underlying a cluster method (e.g., 'similarity' between individuals can be measured in a certain way) are very unlikely to be completely in accord with the unknown state of nature. Nevertheless, it is sufficient that these postulates be robust enough so that the resulting clusters are sufficiently accurate. This is not to say that we could somehow *know* that any given clusters are accurate, for we could not. Nevertheless, a consideration for a cluster method's useful application is the possibility of detecting departures of the state of nature from these idealizations when these departures are so *gross* as to vitiate the method. This general problem of checking postulates and assumptions of statistical models and methods has been little studied, but Rozeboom has described the problem well (1966, pp. 519 ff.). We have proposed elsewhere that the degree of postulate violation be estimated by

the use of statistical tests, christened "consistency tests" by Meehl (1973, 1978; also see Notes 1 and 2).

If these tests are passed in a particular application then one has less doubts about the detected taxonomy being spurious. In brief, when the resulting clusters are inaccurate or spurious there will be manifest inconsistencies regarding the clusters which allow us to detect the fact of serious error. The consistency tests will never be perfect and must be regarded as fallible, as all scientific inference is fallible but they can easily improve on the current situation which employs no such tests at all.

EPILOGUE[1]

Some criticize the "pseudo-problem" approach as being too strong or—oddly, in one instance from the same critic—too weak. They argue it is too strong because it demands of the cluster search methods that they should succeed in detecting the biological taxon, defined causally by its specific dichotomous etiology (Meehl, 1972, 1977), in this case the XX genotype; whereas, psychological masculinity-femininity is not conceptually identical with hormonal or genital sex, and the latter is not even identical, strictly speaking, with the XX genotype. They have further argued that there are other good reasons for using a cluster search method than the method's ability to detect a taxon defined causally or by reference to some hypothesized "real" latent state of affairs (state, event, structure, or disposition).

The "too strong" objection involves deep questions about both philosophy of science and less philosophical (but still "methodological") formulations of the *purpose* of taxonomic procedures. This is not an appropriate place for an intellectually responsible consideration of those deep questions in all their ramifications, which the authors are undertaking in another place (Golden and Meehl, in press). We will content ourselves with a brief statement of our methodological position, giving only a sketch of how we would try to defend it. We are realists rather than instrumentalists (or "fictionists," the extreme of instrumentalism) in our philosophy of science. Further, apart from this general orientation, our current interest in taxometrics derives from our desire to test the dominant gene model for schizotypy (Golden and Meehl, 1979).

1. These comments were written in response to a criticism by some who have read or heard earlier presentations of this paper.

We take it that an investigator who uses cluster methods in the course of a research program on the heredity of schizoidia presupposes the existence of genes and takes his empirical task to be that of deciding about a factual question concerning the objective biological world, to wit, is there one "big gene" involved in schizoidia or is it a quasi-fungible polygenic situation? From the instrumentalist's standpoint, if explanatory theories do not concern the way the world is (the actual state of nature, which according to our philosophy of science we take to be the business of science to find out, rather than merely to concoct "convenient fictions"), it is hard to see why typological (categorical, taxonomic) models and methods should be used at all rather than dimensional ones. We do not here dogmatize the negative, but, to our knowledge, there is no convincing affirmative evidence showing that statistical methods appropriate to forming *types, species, taxa, disease entities, syndromes* or whatever are more powerful than straightforward regression or function free actuarial prediction methods (Lykken, 1956; Lykken and Rose, 1963) when the task is the purely instrumental one of predicting an item of behavior from another item or from a set of other items. So that, aside from our own epistemology which is realist in its aims, we would challenge the fictionalists or instrumentalists who claim to reject the goal of getting at the "true underlying state of nature" responsible for a cluster or syndrome. We invite them to show why, on purely instrumentalist atheoretical grounds, an investigator should employ taxonometric methods rather than the less controversial, more straightforward, and usually more powerful ordinary discriminant and regression methods. Of course, our interest in the pseudoproblem of detecting the biological sex dichotomy does not depend upon others' acceptance or on our arguments against the instrumentalist position. Suffice it to say that *we* are interested, as scientific realists, in the dichotomous causal entity alleged to underlie a family of fallible indicators. Given that interest (which surely cannot be proved illegitimate), it is appropriate for us to require of a method that it should be successful in such detection of the real (underlying causal) state of affairs.

One exception to the lesser utility of classes and dimensions that is prevalent in the areas of psychopathology and medicine lies in the possibility of discovering new indicators in a different research or clinical setting from that in which the original set of indicators was bootstrapped by some appropriate cluster method.

If we take predictor variables and outcome variables singly and study their correlations pairwise, as the number of such potential indicators increases the number of pairwise correlations goes up very rapidly (e.g., $n(n-1)/2$ gives us 190 correlations among only 20 indicators). There is, therefore, a certain descriptive economy involved in identifying a type or taxon which makes it possible for other investigators to study treatment effects or hitherto unsuspected symptoms or prognostic signs without having to include in their clinical recording or their research design the entire set of indicators that can be found in the literature as having some construct validity for the taxon. This makes possible incremental research via a diversity of investigations at different places and using different procedures, a strategy of numerous investigators contributing "piecemeal" correlates without each one having the money, facilities, staff, and number of patients that would be necessary if the entire grand matrix of "all correlations of everything with everything" were to be insisted upon. Furthermore, in the context of discovery, it may be psychologically easier for observant clinicians to "notice" a hitherto neglected sign or treatment effect if they carry in their minds the notion of a certain type or taxon, than if they were required subjectively to compute impressionistic clinical correlations of each such finding with each previously validated indicator, especially when, taken singly, those indicators are highly fallible and hard to remember. Whether the efficacy of that kind of discovery itself depends upon some objective reality to the taxon we do not wish to argue here, except to say that it is an unsettled question of very great interest both from a theoretical and practical standpoint (Meehl, 1959, 1973, 1979).

The objection "too weak a criterion" takes the form of arguing that it is possible for a method that "worked" in the detection of biological sex nevertheless to be fallible in other research contexts. Merely because a search method does well in detecting this latent causal dichotomy would not *guarantee* its infallibility for a different latent causal dichotomy, say Huntington's disease, or the presence of a specific filterable virus. The short answer to this complaint is, "Of course, it won't." The general efficacy of a proposed cluster method can be discorroborated by showing it does badly against a clean objective criterion with strong indicators, as three of the six did in the present instance. We know that biological sex is a clear dichotomy, and we know that its correlation with

psychological femininity, while imperfect, is extremely high. We also know that the reflection of the XX genome, via *psychological femininity*, in patients' responses to the MMPI verbal item pool is sufficiently strong statistically so that when one constructs femininity keys not by bootstrapping or clustering methods but by "external criterion keying" (as the original M_f scale of the MMPI was derived), one can achieve discrimination of the sexes that is 85-95 percent accurate.

It has been objected that the task set for the six competing cluster-analytic methods was difficult or impossible, because MMPI items are just too many steps removed from the sex determining genome. One simply cannot expect, so it is said, that verbal self-reports of interests, attitudes, feelings, and the like will possess sufficiently high construct validity *vis-a-vis* the XX genotype to permit a bootstrapped taxometric identification of the biological taxon, let alone a high accuracy sorting of individual subjects into those taxa by the use of such highly fallible bootstrapped indicators. Associated with this objection (and plausibly potentiating its seriousness) is the claim that the verbal behaviors available, whatever their intrinsic qualitative validity for the genetic taxon might be, are too unreliable to achieve a satisfactory net attenuated construct validity, without using a very large number of such items to counteract the single item unreliabilities. Finally, a variant of this broad kind of criticism is the qualitative content of the MMPI pool itself. There is, fortunately, a short and compelling rebuttal to these objections that the test is not "fair" to the cluster method. If the content of the MMPI item pool (as responded to with a dichotomous "true" or "false" by the subjects) is qualitatively so impoverished, or the number of items too few, or the reliabilities of the items too unstable, or the causal chain from the genome to the "psychological taxon" of sex too weak (as a consequence of its involving a long chain of connections, each of which is of a stochastic rather than nomological character), then it would not be *possible* to identify biological sex by using these items. But we know as an empirical fact that it is possible to do so, as the authors of the MMPI did when they had available an external criterion available for empirical keying. Even the old MMPI Mf scale itself achieves between 85 and 90 percent accuracy in identifying biological sex. This sufficies to show that the net attenuated construct validity of these kinds of verbal items is *not* too poor for the sex identification task we set to it.

Secondly, there are at least three taxometric methods known to us that can be applied in a bootstrap fashion without knowledge of the criterion membership of the individuals, and that will identify the taxon, estimate its base rate accurately, and classify individuals with an accuracy of 85 to 90 percent. Thus we can bootstrap it taxometrically and get a true validity that compares favorably with that achieved by Hathaway and McKinley employing the objective sex membership dichotomy for criterion keying. These are the MAXCOV-HITMAX method (Meehl, 1973, Golden and Meehl, Note 5), the consistency hurdles method (Golden and Meehl, 1979), and the normal mixture method (Golden, Tyan and Meehl, Note 6) developed by the engineer, Hasselblad (1968). Even a cruder form of the latter, developed independently by Meehl (Note 2), was applied successfully to the sex detection problem and yielded a remarkably encouraging accuracy for estimation of base rates, means, and standard deviations, although it was not there applied to the classification of individuals (Meehl *et al.*, Note 7). Put succinctly, our reply to the criticism of too difficult a task because of the nature of the indicator items and their causal distance from the XX genome, is that even if the MMPI items pool is relatively impoverished and insufficiently diversified in content (an impression one does not, we think, form when fair mindedly scanning it), and the individual items unreliable and many steps removed from the taxonic factor of interest—these factors are attenuators of validity but they are not jointly sufficient to prevent those very same verbal items from achieving an identification of the biological sex taxon and a gratifyingly accurate classification of individual subjects into the two taxa. The validity of the old Mf scale itself suffices to show that. But it is further known that at least three taxometric methods can be employed in a bootstrapping way without prior knowledge of the taxon base rate, let alone the item statistics, that result in identification of the taxon, and permit development of a mechanical decision procedure which classifies the individuals with an accuracy as good as that achieved by use of the MMPI test, which was developed using the external criterion keying procedure against the dichotomy of biological sex.

A conjecture that a certain cluster method will usually enable one to detect a taxon, to find the strong items, and to assign weights to them for classifying individuals into the detected taxon or out of it, is strongly discorroborated when the method fails at

its task in a context where the dichotomy is known to exist and the fallible indicators available to this cluster search method are known to be sufficiently valid so that when put together (even by a crude item analytic and unweighted procedure) they are highly accurate. As always, it is easier to refute the claims of a *method* (just as it is possible to refute the truth claim of a substantive *theory)* than it is to "confirm it," to prove that it's correct (accurate, valid).

This brings us to the heart of our response to the objection. The objection is fundamentally mistaken from the standpoint of inductive logic (if we allow that phrase, which Sir Karl Popper would not [Popper 1959, 1962; Schilpp 1974]), the process of arriving at conjectures and testing them against empirical facts. The objection correctly says that one cannot "prove" (a word that should be used cautiously or not at all in this context) that a cluster method is valid with respect to the conjectured underlying truth of the matter. Of course you can't prove it, because you can't "prove" anything about the empirical order in the strict, deductive sense of "prove." Some critics have said that even if the cluster method yields results that are coherent as indicated by consistency tests (which none of the six methods here studied presently provide), and if it succeeds as applied in a variety of pseudo-problems (such as the detection of biological sex) where we know the real answer, it doesn't *prove* that the method will always work. We repeat, "Of course it doesn't." There is no such thing as proving that a method always works. It is a mistake to make such a demand. Similarly, one may say that the identification of a taxon might be a misidentification; that is, the cluster method might contain satisfactory internal consistency tests and make reasonable sense in terms of the item content and yet we could conceivably err in delineating the nature of the inferred causal entity. The answer is again, "Of course, we might." It is simply a mistake to require of any procedure in empirical science that it should be incapable of misleading us. One major difference between empirical science and the formal sciences of pure logic and mathematics is that one can, without committing a methodological error, nevertheless come to the wrong substantive conclusion, an outcome that cannot occur in the formal sciences. If the premises (grounds, bases, arguments) for a syllogism in deductive logic or algebra are true and the form of the syllogism is a valid form, then the conclusion is true. But surely everybody knows that this is not true in

inductive logic, just as it is not true (more specifically) in statistical inference. There is no possibility, and fortunately there is no need, of showing that what appears to be a powerful cluster method when it is checked against independently known criteria via a bootstrapping research study (with the criteria initially kept from the investigator, as here) guarantees that it will always work in all situations. Something approximating such a guarantee can be given to the extent that the search procedures flow deductively from the theorems, that in turn flow from the postulates of a specified latent taxonomic causal model. That is a sufficiently strong kind of inference that most of us would call deductive. But it must be remembered that, even in this case, the abstract possibility of cooking up some alternative conjecture is always present. All that the scientist can say to someone who adduces that possibility is to invite him to present his substantive alternative conjecture and subject it to tests as the first investigator has done. For these reasons, which we think rely upon the best current expertise in philosophy of science, we view the "too weak" objection as fundamentally misconceived, based upon a confusion between the nature of inference in the formal sciences and in the empirical disciplines.

Reference Notes

1. Meehl, P. E. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Rep. PR-65-2). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1965.

2. Meehl, P. E. *Detecting latent clinical taxa, II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Rep. PR-68-4). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1968.

3. Forgey, E. W. *Detecting 'natural' clusters of individuals.* Western Psychological Association. Santa Monica, 1963.

4. Golden, R. R. & Meehl, P. E. *Detecting latent clinical taxa, IV: Empirical study of the maximum covariance method and the normal minimum chi-square method, using three MMPI keys to identify the sexes* (Rep. PR-73-3). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1973.

5. Golden, R. R. & Meehl, P. E. *Detecting latent clinical taxa, V: A Monte Carlo study of the maximum covariance method and associated consistency tests* (Rep. PR-73-4). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1973.

6. Golden, R. R., Tyan, S. & Meehl, P. *Detecting latent clinical taxa, VII: Maximum likelihood solution and empirical and artificial data trials of the multi-indicator multi-taxonomic class normal theory.* (Rep. PR-74-5). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1974.

7. Meehl, P. E., Lykken, D. T., Burdick, M. R. & Schoener, G. R. *Identifying latent clinical taxa, III: An empirical trial of the normal single-indicator method, using MMPI Scale 5 to identify the sexes.* (Rep. PR-69-1). Minneapolis: University of Minnesota, Reports from the Research Laboratories of the Department of Psychiatry, 1969.

REFERENCES

Anderberg, M. R. *Cluster analysis for applications.* New York: Academic Press, 1973.

Bailey, K. D. Cluster analysis. In D. Heise (ed.), *Sociological methodology.* San Francisco: Jossey-Bass, 1974.

Blashfield, R. K. Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. *Psychological Bulletin,* 1976, *83,* 377-388.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20,* 37-46.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, *52,* 281-302.

Everitt, B. S. *Cluster analysis.* London: Halsted Press, 1974.

Golden, R. R. & Meehl, P. E. Testing a dominant gene theory without an accepted criterion variable. *Annals of Human Genetics* (London), 1978, *41,* 507-514.

Golden, R. R. & Meehl, P. E. Detection of the schizoid taxon with MMPI indicators. *The Journal of Abnormal Psychology,* 1979, *88,* 217-233.

Golden, R. R. & Meehl, P. E. *Taxometric analysis of causal entities: Detection of the schizoid taxon.* New York: Academic Press, in press.

Hasselblad, V. Estimation of parameters for a mixture of normal distributions. *Technometrics,* 1968, *8,* 431-444.

Lance, G. N., & Williams, W. T. A general theory of classificatory sorting strategies. I. Hierarchical system. *The Computer Journal,* 1967, *9,* 373-380.

Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Lykken, D. T. A method of actuarial pattern analysis. *Psychological Bulletin,* 1956, *53,* 102-107.

Lykken, D. T. & Rose, R. Psychological prediction from actuarial tables. *Journal of Clinical Psychology,* 1963, *19,* 139-151.

Meehl, P. E. Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology,* 1959, *13,* 102-128.

Meehl, P. E. Schizotaxia, schizotypy, schizophrenia. *American Psychologist,* 1962, *17,* 827-838.

Meehl, P. E. Specific genetic etiology, psychodynamics and therapeutic nihilism. *International Journal of Mental Health,* 1972, *1,* 20-27.

Meehl, P. E. MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl, *Psychodiagnosis: selected papers.* Minneapolis: University of Minnesota Press, 1973.

Meehl, P. E. Specific etiology and other forms of strong influence: Some quantitative meanings. *Journal of Medicine and Philosophy,* 1977, *2,* No. 1, 33-53.

Meehl, P. E. Theoretical risks and tubular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 806-834.

Meehl, P. E. A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment,* 1979, *43,* 563-581.

Popper, K. R. *The logic of scientific discovery.* New York: Basic Books, 1959.

Popper, K. R. *Conjectures and refutations.* New York: Basic Books, 1962.

Reichenbach, H. *Experience and Prediction.* Chicago: University of Chicago Press, 1938.

Rozeboom, W. W. *Foundations of the theory of prediction.* Homewood, Illinois: The Dorsey Press, 1966.

Schilpp, P. A. (ed.) *The philosophy of Karl Popper.* LaSalle, Ill., Open Court, 1974.

Tryon, R. C. *Cluster analysis.* Ann Arbor, Michigan: Edwards Brothers, 1939.

Wishart, D. An algorithm for hierarchical classifications, *Biometrics,* 1969, *25,* March, 165-170.

Zubin, J. A technique for measuring like-mindedness. *Journal of Abnormal and Social Psychology,* 1938, *33,* 508-516.