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

A Funny Thing Happened to Us on the Way to the Latent Entities

I daresay some of you—especially those of my age group—were surprised to find the Bruno Klopfer award bestowed upon a notorious dustbowl empiricist whose modest claims to fame include a wicked book on actuarial prediction (Meehl, 1954) which I am currently struggling to up-date, and the advocacy of cookbook interpretation for the most widely used of structured personality inventories, the MMPI (Meehl, 1973d). I must confess that I was pretty surprised myself. But on reflection, I concluded that your outfit (which traces its ancestry to the old mimeoed *Rorschach Research Exchange*) showed scholarly integrity and good taste. For the record, you can't pin the Minnesota Multiphasic Personality Inventory on me, as I was an undergraduate during its development. The Minnesota lore is that McKinley wanted it, Hathaway built it, and Meehl sold it—not a bad summary of the situation! The eminent contributor for whom this award is named was known to me, although not well, because back in 1947, I went off first to Michael Reese to learn some Rorschach with Beck and followed it a couple of months later by going to Bard College where Klopfer was doing a workshop.

My first publication, "The dynamics of structured personality tests" (Meehl, 1945), replied to a paper by a distinguished contributor to projective techniques, Max Hutt. While that polemic was overly optimistic, it was, I think, an important contribution for its time. It urged a more sophisticated way of looking at verbal items, more like the way we listen to psychoanalytic material or responses to ink-blots, but linked to a more atheoretical blind external criterion keying than I now defend. Those old polarizations between structured and projective methods are weaker today, and I think this is partly because all of us have undergone some disillusionment with the power of our favorite assessment methods. The younger clinicians are less test oriented, whether structured or projective, than was my generation. Part of this comes from concerns about reliability and validity and the often depressing results of validation studies. But I think some of it comes from a greater skepticism about the usefulness of inferences concerning latent entities in favor of a greater shift toward purely behavioral, dispositional analysis. I suppose I play a role here, both in the emphasis upon empirical keying and in my views on actuarial prediction, which downgrades the usefulness of mediating forecasts via structural or dynamic inferences. There is a philosophical point here which I would still press, namely, that in the other sciences, powerful predictions can be mediated by theoretical constructs only when two conditions are met (Meehl, 1973a). First, the theory is well worked-out and well corroborated, having high verisimilitude, as Popper calls it (1962, 1976); secondly, there exists a powerful technology of measurement. Since we meet neither of those conditions in clinical psychology, we should not be surprised to find out that our predictive powers are limited.

Students have commented on an ambivalence in my writings on assessment, but I like to think that this is reality based. I do not dispute the powerful influence on me of cultural factors, such as the accident of Minnesota geography. But I was first attracted to the field by reading Karl Menninger's *The Human Mind,* and my cognitive passions were mobilized by psychoanalysis. I find it hard to imagine what kind of psychologist I might have become had I been born and raised in San Francisco or New York City, but it was

Part of this article was presented at a meeting of the Classification Society, Gainesville, Florida, April 10, 1979.

cheap and convenient to go to the University of Minnesota where I was exposed to behaviorist and measurement-oriented super-objectivists like Donald G. Paterson, Starke R. Hathaway, William T. Heron, and B. F. Skinner. Despite this education, in my 1954 book on prediction, a page count shows that even including the pro-actuarial results in the empirical chapter (and after all, I am not responsible for how those studies came out), by far the larger part of my text defends the clinician's special powers of psychodynamic inference against the claims of Lundberg, Sarbin, and others that he's nothing but a second-rate computer. I had analysis partly with a Vienna trained analyst and mostly with one trained at Columbia under Rado (I may say that Nunberg, Ackerman, and Helene Deutsch are among my psychoanalytic grandparents), and after doing a couple of controls, I then practiced a mix of psychoanalytic and rational emotive therapy which I continue today. It would be surprising if a person with that life history didn't show some ambivalence about clinical inference!

While this talk is about cluster problems, psychoanalytic inference is not without relevance, as I need hardly remind a group interested in projectives. Most objections I hear from my behaviorist anti-Freudian colleagues seem to me beside the point, focussing on the legitimacy of unobservable entities, operational definability of terms, reduction of psychodynamic to learning concepts, or the problem of mapping into physiology. All of these I view as red herrings (Meehl, 1970). The big problem about psychoanalytic inference is contained in Fliess'attack on Freud at their last meeting at Achensee in 1900, when he told Freud that "the thought-reader merely reads his own thoughts into other people." Freud, who, right or wrong, usually knew pretty much what he was up to and what the scientific stakes were, immediately perceived that this attack went to the jugular. If that was what Fliess really believed, he should throw the just-published parapraxis book into the waste-basket; and that Achensee "congress" was the last time the two men ever met. The critical problem about psychoanalytic concepts is precisely here. It is the epistemological or methodological problem of the inference, rather than any of those other side issues commonly mentioned. The shift in psychoanalytic therapy away from reconstructing the past, and even from the interpretation of dreams, to moment-to-moment handling of the transference, and the ascendency of two powerful therapeutic competitors, namely, rational emotive therapy and behavior therapy, have lessened the clinical importance and, I fear, the intellectual interest of psychoanalytic therapists in this epistemological problem. But students and colleagues still ask me, "Meehl, you have always been interested in psychoanalysis. How come you haven't ever published any quantitative research on the psychoanalytic hour itself?" They don't seem to believe me when I say straight forwardly, "Because I don't know how." And, alas, I don't think anybody else does either.

But it is not inference from dreams, associations, expressive movements and parapraxes of the kind we make in the psychoanalytic session that I want to discuss with you here. Rather, I want to consider another kind of inferred, latent entity allegedly underlying human behavior and experience, namely, the inner structures or states—if you like, the genotypic traits and types—that psychologists have searched for using correlational methods. The paradigm case of factor analysis I am going to bypass, partly because I am not sufficiently expert and have only conducted a few factor analyses in my career, although I may say that the few I have done were more illuminating than I had anticipated, having been taught at Minnesota that factor analysis was pretty much a waste of time. If the rotation problem can be solved nonarbitrarily, which is doubtful, I think one of the main problems that will persist in factor analysis is the conceptualization of the obtained factors. There is here a highly subjective interpretative feature present, comparable in quality, and I fear sometimes even in amount, to that which obtains in psychoanalytic interpretation of the patient's verbal material. I co-authored a paper some years ago on an approach to reducing that subjectivity of factor interpretation, which we

thought was rather clever at the time, but which seems to have dropped into the bottomless pit and this talk gives me a chance to recommend you have a look at it (Meehl, Lykken, Schofield, & Tellegen, 1971). Briefly, what we showed is that using not test data, but therapists' ratings as our raw material, to be correlated and factored, then if one presents skilled clinicians with a randomly chosen half of the phenotypic traits showing high loadings on a factor, and these clinicians, after independently characterizing the factor's psychological nature, have a meeting in which they thrash it out and settle on a christening of the factor; then if you present a set of such factor names to a new batch of clinicians who do not see the original half of the items and whose task is to match the remaining half of the items with the factor names, you achieve almost a perfect correct identification. We didn't expect it to come out this well, but since it did, we have been surprised that nobody has picked up the method subsequently.

It has taken a too long and rambling introduction to get around to my main focus here which is not the inference to psychodynamic entities from psychoanalytic session material nor from projective tests, nor the identification and interpretation of dimensions underlying a battery of ratings or psychological test scores, as in multiple factor analysis; but the inference to *types* or *taxa,* entities of a categorical rather than dimensional kind. Detecting these taxonic entities has been the traditional aim of the methods called cluster analysis, and the field has undergone a tremendous publication explosion within the last decade. (I belong, for instance, to an organization called the Classification Society that was originally dominated by plant people, bug people, and statisticians, and had only a half a dozen psychologists at its first meeting in 1970. Our profession now numbers about a third of the membership.) I want, in the remainder of my remarks, to make, forcefully, a point about cluster analytic methods that we have been reluctant to face. It goes without saying that in science as in psychoanalytic therapy, if a problem is not faced, it can hardly be solved. In the latter portion of my remarks I will say something about some current efforts of my own and my former student, Robert Golden, now at Columbia, to solve one important subclass of the cluster problem. What follows is an expansion of some remarks I made as a member of a panel on the proposed revision of DSM-III at the Classification Society's April (1979) meeting in Gainesville. Readers unfamiliar with the current situation in cluster analysis and numerical taxonomy may sample the intellectual ambiance of the Classification Society in Sokal (1974), Hartigan (1975), or Sneath and Sokal (1973). An excellent brief exposition for psychologists is found in Blashfield (1976), and see also Blashfield (Note 1), Blashfield and Draguns (1976), and Blashfield and Aldenderfer (1978).

I shall not offer here a rigorous general definition of 'true type' or 'taxon' or 'entity,' for which taxometric rather than continuous factor analytic methods are appropriate. I have concluded that such a general definition cannot be given except implicitly, via the mathematical formalism itself, together with references to dichotomous etiology that are problematic and not defensible within my time constraints. Roughly, then, I mean by a 'type' or 'taxon' a class entity having a nonarbitrary basis of categorization, that is, not simply a conjunction of attributes that one might impose conventionally for some useful communicative or administrative purpose, but a class of persons that really belong together, a classification punchily expressed by the metaphor that we wish, as Plato said, "to carve nature at its joints." Seeking a more rigorous (mathematical or causal) definition of a "true taxon" convinces me that taxonicity is itself not taxonic, but a matter of degree. Some of you might want to read my paper in the *Journal of Medicine and Philosophy* (1977) on some quantitative meanings of specific etiology and other forms of strong influence, and a forthcoming book on taxometrics by Dr. Golden and myself (Golden & Meehl, Note 2). While a genetically determined disease entity, such as

Huntington's or (as I believe) schizophrenia and manic-depression provides the clearest examples of causal taxonicity, we should not equate that notion with the medical model, or even with genetic etiology. Any kind of type or syndrome, however produced, that has a sufficiently strong knitting together, statistically and causally, is a taxon. For instance, consider Freud's early hypothesis that the specific etiology of hysteria is a passively experienced pre-pubescent sexual trauma involving genital stimulation with affects of fear and disgust predominating, whereas the obsessional neurosis springs from a similar sexual experience, but with enjoyment, aggression, and activity initiated by the subsequent patient. That would have been a perfectly good taxon, even though it had nothing to do with germs or genes. 'Indigenous fascist' is a perfectly good taxon, as is 'Stalinist,' 'Freudian,' or 'behaviorist.'

It seems odd that what purport to be objective, quantitative methods of classifying mental patients have not been conspicuously successful, although the effort goes back at least a half century. I don't wish to commit overkill, but I do aim at a confrontation. We have here a mind-boggling fact: No accepted entity in psychopathology has owed its initial *discovery* to formal clustering methods, whether those invented by psychologists, biologists, or statisticians. In fact, I know of no agreed-upon instance where taxometrics has been mainly responsible for definitively *settling* a controversy about an already "noticed" syndrome, e.g., as to its existence if disputed, its specific etiology, whether it should be sub-divided, or subsumed under another accepted category as a *forme fruste*, and so on. If I am wrong in this strong negative claim, it should be easy to refute me. The thesis is historical, and one clear counter-example will suffice. In organic medicine, physicians succeeded in identifying hundreds of disease entities long before Pearson invented chi square or the correlation coefficient. Even in psychopathology, recent advances in the genetics of major psychoses such as schizophrenia and the distinction between the bipolar and unipolar affective disorders have relied hardly at all upon formal clustering methods. Instead, researchers have moved forward by improving the verbal specification and the objective recording of the behaviors that were traditionally looked at by psychiatrists and clinical psychologists. For example, the recent work on bipolar and unipolar affective disorders showed that the single objective symptom 'agitated pacing' yields a nearly perfect discrimination between unipolar and bipolar depressions when the diagnosis is based upon the previous history of the patient or the patient's family for a manic attack (Depue & Monroe, 1978, pp. 1004-1005). Or with respect to schizophrenia, researchers have not said, "If we start with a big mess of correlations of everything with everything, will such a thing as schizophrenia emerge from the mess?" but instead, "Conjecturing that schizophrenia is a disease entity that has objective reality, how can we refine the way in which the interview is conducted and observations recorded so as to study its heritability?"

It seems to me that this historical observation is important. We can't explain the ability of medicine to identify disease entities by saying that all medical syndromes are tightly knit, which they are not; or by referring to pathognomic signs, that are rare in organic medicine, probably as rare as in psychopathology. The existence of external validating criteria in the form of specific pathology and etiology help (Meehl, 1973e, pp. 284-289), but hundreds of entities were identified before the specific etiology and pathology were known. So this is a puzzle about the history of medicine, and I think somebody should look into just how they managed to do it without any fancy statistics. I have some hunches, such as the fact that in medicine we focus on a small number of powerful indicators rather than a large number of weak ones such as regularly practiced in the classical psychometric model, and that medicine does not start, so to speak, blindly entering a huge matrix of intercorrelations among patients or variables. The physician hasn't said, "Look, we have hundreds of people coming into this outpatient clinic sick in

a lot of different ways with a variety of complaints. Let's concoct some ingenious formula or index for putting together their symptoms and then try to extract truth out of the mess." Instead, the physician has seen a small number of patients with certain striking configurations and writes a description of this which immortalizes him with an entry "Fisbee's Syndrome" in Dorland's *Medical Dictionary*. Conjecturing that there is such an entity as Fisbee's Syndrome and postulating that underlying any statistical order of symptoms, complaints, course and response to treatment, there is some strong causal factor within the individual, we try to improve the identification of Fisbee's Syndrome versus everything else. This leads to a rather different kind of research strategy from the statisticizing of a large can of worms.

Some believe that the main reason our methods haven't been earthshakingly successful in psychopathology is that the model is wrong, that there aren't any entities analogous to measles, mumps, and cholera. I don't believe that for the major psychoses, for the psychopath, and for one or two of the neuroses. The question is fundamental: Why do we want to classify anyway, instead of merely predicting useful dispositions, such as response to a drug? I am unaware of any proof that it is statistically profitable to sort people into taxa or groups or types or species if there aren't any real types to be discovered. Is the procedure merely imposing our arbitrary order for an alleged economy in description? We often talk about this kind of economy, but I must confess I've never understood what we mean. From the standpoint of clinical handling, it seems unlikely that a taxometric procedure would be better than a purely dimensional one. Given a finite set of indicators that are correlates of some disposition such as response to group therapy or differential reaction to two antidepressants, why mediate transition from predictors to predicand via a taxon? I am not enough of a mathematician to give a general formal proof, but I submit the following rough and ready argument against such an arbitrary imposing of taxonicity on dimensional facts: Given a half dozen facts about a patient such as certain Multiphasic scores, items from Mental Status and life history, we want to predict whether he will respond to Elavil. We can combine the predictors in some simpler linear form—Goldberg, Dawes, and Co. would say even unweighted standard scores! (Dawes, 1979; Dawes & Corrigan, 1974; Goldberg, 1965; Wainer, 1976)—or, if you think it's configural, using a function-free actuarial table as advocated by my colleague Lykken (1956; Lykken & Rose, 1963). In the classification approach, I first move from the predictors to the taxon, with statistical slippage, and then from the taxon to the drug response, with some more statistical slippage. Why would that be a profitable endeavor, especially if the taxon has no reality, but is only in the statistician's head?

Doubtless I disagree with many of you on a philosophy of science issue, in that I am a scientific realist rather than a fictionist or instrumentalist. For me, the purpose of taxometric procedures is to carve nature at its joints, to identify the real underlying taxonic entities whose conjectured existence motivates a classification rather than a dimensional approach. As Ernest Nagel says, the difference between realist and instrumentalist is usually only philosophical and without impact on their work as scientists. But sometimes it makes a difference in one's research strategy. In the *Medicine and Philosophy* paper, I set out a dozen meanings of 'strong influence,' formulable mathematically and metatheoretically. The three or four "strongest" are easily identifiable as influences that would in organic medicine be called 'specific etiology' (cf. Meehl 1973b). Examples: A necessary and sufficient condition, such as the Huntington gene; or a necessary but not sufficient condition, as the gout genotype; or a threshold value, as in a deficiency disease below a certain minimum dietary input. Somewhat weaker is the case where a causal variable is the most powerful one everywhere, such as caloric intake in nonglandular obesity. I see the basis for a nonarbitrary taxometric approach as causal, which I realize is controversial. Unless I hypothesize a quasi-dichotomous causal factor,

historically or latently generating the pattern of inter-patient resemblances, I doubt I should be doing a classification job rather than a dimensional analysis in the first place.

When "true taxon" denotes a conjectured specific etiology, we replace the usual statistician's reference to "assumptions" by neo-Popperian "auxiliary conjectures." These hypotheses, while auxiliary to the main substantive theory of interest, are still factual claims rather than mere conventions for data manipulation (Meehl, 1978, pp. 818-821). They constitute a part of the whole network of scientific hypotheses and are falsifiable. But they may not be subject to *direct* falsification, as in an auxiliary assumption of homogeneous covariances or linear regression. They may be falsifiable only by falsifying indirect, remote consequences. This is true of several consistency tests that Robert Golden and I have devised. I believe that consistency tests of a postulated latent causal model are imperative in taxometric investigations. In my paper on Popper and Fisher just cited, I develop this thesis and point out that the reason consistency tests are not labelled as such in sciences like chemistry, astrophysics, and molecular biology is that they are used ubiquitously, taken simply as a basic process of respectable science (do two or more numerical procedures lead to the same inferred answer?), so they don't have a special name (Meehl, 1978, p. 829). A taxometric procedure in psychopathology which has to begin and end by saying, "If the reader is willing to assume with us that..." is weak. Even if it happens to carve nature at its joints, the investigator and his readers have no way of knowing whether it has in fact succeeded in doing that. Fisher said, in criticizing systematic plots like the Knut Vik square, that we want not merely to have a small error, but to have an accurate estimate of error. The analog to this methodological demand in classification is that we want not only to have a search method which will detect the true entities underlying the phenomena, but we want to have auxiliary methods which tell us whether we have succeeded in that detection. If twenty people are guessing how far it is to the moon, it doesn't help if one of them is right on the nose unless we have some way of telling which one he is! Again, I am not enough of a mathematician to give a rigorous formulation of the difference between a consistency test and a main estimator of a conjectured latent quantity, but it could be done first round in terms of the relation between number of equations and number of unknowns. An expression of equality that is not an algebraic identity, or one stemming simply from the formalism itself such as integrating a differential equation, but an equality in which the expressions on the two sides contain variables that are assigned numerical values from our taxometric search procedure, should be satisfied within allowed tolerances. If the equality between these two expressions is not a mere identity, but flows as a theorem from suitable postulates formulating the conjectured latent causal model, then it can serve as a consistency test when we plug in the numerical values and see whether they fit. If they don't, either something is wrong with the model or we have had an unfortunate sampling error, despite the model itself being substantially correct. I do not see much point in doing significance testing here because (other than my general distaste for the significance test tradition, Meehl, 1970, 1978) we already know that the substantive theory as stated in the formalism is literally false. All theories are false, and for a neo-Popperian, it's a question of how much verisimilitude the formulation has. Suitable conventions concerning tolerances are preferable to showing a statistically significant difference from the idealized model, which will always happen if the sample is large enough and the measures sensitive. I take this opportunity to urge (as I did at the Classification Society) something that, while involving some risk of the taxpayer's money, is pretty sure to be informative. I propose a large scale study of the presently available taxometric search methods. Methods would be chosen on the basis of some combination of their current popularity and the plausibility of their mathematics, given the metatheoretical position that we want to carve nature at its causal joints rather than just lump people together administratively or for an alleged economy of description. I take that "realist" view partly

because I don't know how the rules of the game for such a study could be written on a nonrealist, fictionist view that all cluster methods ever do anyway is impose our desire for order upon the world, that is, they don't discover entities, they invent them. I am not interested in inventing entities. As I once said in an argument with my friend Gardner Lindzey about fictionism in psychoanalytic theory, if there isn't any Santa Claus, then it isn't he that brings the presents! One would include methods that rely on very different latent structural models in the postulates employed in their formalism and in the algorithm of their search and weighting method. One would also want to try methods that are similar but differ in some interesting respect. I would, given my prejudices, prefer methods that rely upon a moderately strong causal model rather than methods of which nothing could be said except they represent one more ingenious way of combining a bunch of differences or distances in a phenotypic space. (I think that a physical scientist would find the exercise of ingenuity in concocting distance measures and cluster algorithms a bit odd, but I won't press the point. I hasten to add that I am not pointing the finger pharisaically at others, having committed similarity index concoction myself!) I would give high weight to the existence of consistency tests, and if there are otherwise promising or popular methods which do not presently have consistency tests available, some preliminary efforts should be made to derive them. I suspect that there are a lot of potential consistency tests around which, because of the lack of emphasis upon this aspect of the problem in the classification tradition, nobody has bothered to derive. After such a preliminary screening of taxometric methods by a combination of intuitive plausibility, similarities and differences, current popularity, and the derivability of consistency tests, we could compare the methods on three broad fronts. First, we select taxon or species instances from several domains of the biological and social sciences, dispersing the qualitative characters of domains and emphasizing loose clusters that are known to reflect a truly dichotomous causal origin. For example, nobody disagrees that multiple sclerosis is different from tabes dorsalis or that mumps is different from scarlet fever. We try disease entities where we have an external criterion provided by pathology and etiology, ranging over neurology, pediatrics, internal medicine, and so forth. We do the same for suitable examples in entomology, botany, geology, and the like. For each instance, the requirement is that we have a syndrome of indicators of only moderate tightness (there is no point, as my philosopher friend Feigl says, in cutting butter with a razor), but we have an independent criterion (such as tissue pathology or causal origin or what kind of rock something is found in) that specifies what it means to carve nature at its joints.

We employ two or three dozen such pseudo-real problems, "real" in the sense that there are biological entities being measured and clustered, but "pseudo" in the sense that the taxon and its membership are known to us independently of the cluster search method. We are asking whether we *would* get the right answer if we lacked access to the independent objective criterion (Golden & Meehl, [1980]).

Use of pseudo problems with a known real answer over various areas of the life sciences is desirable because neither analytical derivations nor Monte Carlo methods can quite simulate the features of irregularity and discontinuity and so on which we find in the world of real taxa. The most plausible behavioral examples would be those in behavior genetics. For instance, there are now something like 120 clearly identified Mendelizing mental deficiencies which would present a nice problem because, on the behavior side, it would be surprising if studying patterns on subtests of intelligence tests results would succeed, whereas combining these with some of the nonbehavioral pleiotropic effects found in these conditions, e.g., skin markings, developmental malformations of the ears, or whatever, might be quite powerful. From the purely behavioral side, things get a little tougher, but even there, one can think of examples. For instance, a cluster method that fails to discriminate active dues-paying members of the John Birch

Society from similarly zealous members of the Socialist Workers Party would not appear to be very promising! We might confine this first domain to situations commanding quasi-universal agreement among informed persons that a real taxon exists.

Secondly, extensive Monte Carlo runs combining various latent parametric situations and a wide range of sample sizes should be conducted, an important part of that procedure being to study how well consistency tests detect sample results as untrustworthy, as giving the "wrong answer." Using four consistency tests as successive hurdles that a sample must "pass" to be acceptable, Golden and I were able to detect every single "bad" sample (one giving poor taxonic parameter estimates) in 600 Monte Carlo runs, at the expense of only 6% false alarms (Meehl, 1978, Table 1, p. 827). The Monte Carlo runs should include situations where there is no taxon but in which various dimensional relationships could produce a pseudo taxonic situation. The question here is, can the method be fooled into finding types when there aren't any?

The third sector is problematic but, I think, still worth doing. Here we apply the methods in contexts where there are real problems, where some expert agreement exists that there is a taxonic entity, but there is persisting disagreement as to what are strong indicators. For instance, another clinician and I might be in complete agreement as practitioners and as scientific researchers that schizophrenia is real, an entity having something to do with genes and not merely society's wickedly labelling people of strange life styles. But we might disagree as to the importance of a given symptom, such as anhedonia. There are many such questions around, both in psychopathology and in nonpsychiatric medicine. For example, I gather that there is still not universal agreement in medical genetics as to whether the juvenile and the adult forms of diabetes involve a different locus or simply a question of the juvenile patient being less adequately protected by polygenic modifiers. There is dispute as to whether malignant hypertension represents simply the upper end of the distribution of blood pressures in otherwise normal young persons, or does that suspicious "bump" or "tail" reflect a taxonic entity whose frequency in the population is not sufficient to generate a visible bimodality but only a skewness. (See, e.g., Murphy, 1964; Wender, 1967; Meehl, Lykken, Burdick, & Schoener, Note 4.) Do certain taxometric methods help clarify such situations? Are the more clarifying methods the same ones that show up well in the Monte Carlo runs, and in the pseudo-problem runs with a known true dichotomous etiology?

The hope would be to identify taxometric procedures that consistently "work" in that they give the true answer or help us to reach consensus about a plausible answer on currently active scientific problems. If no methods consistently work, we could at least scrutinize the mathematics and conceptual underpinnings that correlate with a tendency to work better. It could be that nothing works, but I can't believe that. More probably different methods will work for different kinds of situations. But I think we should be able to generate latent structures which tell us something about certain minimal conditions that a method should possess if it is likely to give the right answer. If nothing consistent showed up, that would be discouraging; but it would be enlightening.

If I permit myself to play diagnostician and prophet, attempting to say what's the matter with the social scientist's search for latent entities via the received cluster methods, I would identify several features of our approach that are individually damaging and jointly have made us so largely unsuccessful. Some have already been alluded to in the preceding text, but I want to pull them together here. Note that each of them contrasts with organic medicine, where without high powered statistics, entities have been identified with conspicuous success. First, we lack an independent criterion that corresponds to the internist's pathology (and, when known, etiology). About that lack, there is nothing we can do. If a syndrome in psychopathology is not caused by a germ,

and does not involve tissue pathology (in the usual sense of Virchow), so that, at most, anything "anatomically haywire" is a matter of CNS fine structure, the psychoclinician is permanently disadvantaged. But I repeat that organic medicine succeeded in identifying numerous disease entities prior to Koch's postulates, which date in the 1880s, and even prior to the development of scientific "pathological anatomy" as it used to be called. Secondly, instead of starting out with a huge battery of miscellaneous indicator variables, not chosen with a taxon of interest in mind, we should rely on clinical experience and sketches of causal theory to identify a relatively small number of potentially powerful taxon indicators. Thirdly, it may be better strategy to think in terms of Taxon X versus everyone else, and then subsequently to research Taxon Y versus everyone else, studying conjectural entities one at a time, tailor-making indicator domains to the focus of interest, instead of the social science tradition of an almost blind statistical scanning of heterogeneous dispositions hoping that cluster statistics will be enlightening. (Our focussing on this strategy is one reason, aside from proper modesty and scientific caution, that Golden and I lay no claim to having "solved the cluster problem," despite our growing confidence that we have "solved" what might be called the *conjectured taxon problem*.) Fourthly, we want qualitatively diverse indicators, not devices all of the same sort which are heavily saturated with methods or instrument variance (a policy I violate myself later in this paper). Fifthly, a few strong indicators minimally correlated are, as the history of medicine shows, better than many weak ones with variable correlations. Sixthly, the cluster algorithm ought not be arbitrarily concocted but should reflect our conjectures, however primitive, about the underlying causal structure. Seventh, it is undesirable to employ procedures that will always give a clustered result, as true of many popular cluster methods, in that they always end up slicing the phenotypic pie regardless of whether a true taxonicity exists; and, furthermore, two different methods that each guarantee to slice the pie may (sometimes surely will) slice it in different ways. An investigator who adopts his favorite cluster algorithm because his teacher believed in it or because he happens to have a canned program around is proceeding in a way that it is perhaps not uncharitable to describe as whimsical. Finally, I cannot emphasize too strongly the desirability—I, myself, would be more inclined to say the necessity—to develop consistency tests sufficiently strong that a taxometric search procedure can "flunk" them. I mean that one may appear to succeed in identifying a taxon, but the internal quantitative relationships of the data force him to admit that something is very wrong with the inferred situation and so he cannot rely on the deliverances of the method.

Simple conventional statistics (like indicator phi-coefficients) can yield odd results that ought to make us worry about statistical identification of taxa. For example, I have concocted a simple paradox by considering four diseases, the first three of which have clinic population base rates equal at .30, and the fourth one a base rate of .10. We consider symptoms *a, b, c, d,* through *i*, the first disease D_1 producing symptoms *a, d, e,* the second one D_2 symptoms *b, f, g,* the third one D_3 symptoms, *c, h, i.* The fourth entity D_4—which we haven't discovered yet—invariably produces the triad of symptoms *a, b, c.* There is nothing bizarre about such a configuration, and in fact, if anything, it makes life easy for the "search statistician" by involving only four disease entities, with indicators infallible as exclusion tests (one-way pathognomicity, cf. Meehl, 1973c, pp. 208-211; 1973c, pp. 230-231). Would we discover the new disease entity, D_4, using one traditional procedure, R-correlating everything with everything in a big symptom matrix? The answer is that we would not. Constructing simple four-fold tables from these proportions will quickly convince the reader that the three phi-coefficients for symptoms (*a*, *b*, *c*) taken pairwise are all *low-negative* (–.25). Only if the clinician pays attention to the triplet (or, if you like, the conditional probability of *c* given conjoint presence *ab*, and

similarly for the other two divisions) will he "notice" that the syndrome of D_4 exists. I do not suggest that this is a typical state of affairs, although I see no reason to think it extremely rare, the numerical assumptions not being outlandish. My point is a general methodological one, namely, the fact that such a thing can happen ought to make us nervous about identifying taxonic entities by calculating phi-coefficients on a big symptom matrix.

In this kind of example, a relatively simple detection procedure seems to work, comparing the incidence of symptom triads, tetrads, and higher conjoint patterns with expected values based on simple probability multiplication. But the number of patterns gets big quickly with increase in indicators, and I have not yet worked out the general mathematics rigorously.

I think we have to face the fact that the deliverances of traditional cluster methods have, by and large, been disappointing in psychopathology. Not to sound overly pessimistic, let me conclude by summarizing briefly some recent research of Dr. Golden and myself which attempts, we do not yet know how successfully, to draw on the lessons of experience and philosophy of science. I have been interested in the diagnosis of pseudoneurotic schizophrenia, and even more, Rado's "compensated schizotype." Being one of the minority still betting on a dominant gene theory, I realize that unless very low penetrance is abused as a fudge factor to defend that theory from refutation by the concordance data, it is necessary to identify the non-psychotic schizotype with respectable accuracy for genetic statistics. That includes identifying those patients in a mixed psychiatric population who receive some other nosological label (such as anxiety neurosis or reactive depression) but who are, in the eyes of Omniscient Jones, really schizotypal because their psychopathology is mainly attributable to the causative influence of the specific dominant schizogene. You will have to put up with my using the MMPI as behavior data, not because I see it as God's gift to the clinician, but because we have a large data bank at Minnesota that permits the kind of statistical manipulation required. I would prefer to have behavior ratings, soft neurology, soft cognitive slippage, and other qualitatively diverse data, as urged in my earlier remarks. The first methodological point is to avoid the flabby procedure of psychologists and sociologists, showing that there is some feeble difference between two groups. Nor do we want to rely mainly upon external criterion keying (despite Meehl, 1945!), since if there were a satisfactory criterion permitting a numerical statement of the valid and false positive rates for MMPI items or scale cuts, that criterion would be the one to use in testing a genetic model. We want numerical estimates of item parameters, and of the hit rates attained by items and item patterns, or scales and scale combinations. We don't just want to say the schizophrenics differ from the manic-depressives, or the males differ from the females, or inpatients show more bizarre items than outpatients, or any of that kind of feeble "statistically significant" business. We want to estimate the base rate of schizotypy—a *number*—in the clinical population, and to determine the valid and false positive rates of the items (singly and collectively) so that for any given patient showing a pattern, we can assign a Bayes Theorem probability that he does or does not belong in the schizoid taxon. Verbally explained, what we want to do is have the best of both worlds by combining some features of the dustbowl empirical keying approach with the bootstrapsing approach via item patterns, and relying upon the consistency tests to tell us if we have failed in our endeavors despite generating a phony success. We begin by a crude item analysis of the MMPI 550 item pool, using formal diagnosis of schizophrenia against Minnesota standardization normals. This is only to get our foot in the door by finding a set of potentially powerful indicators. That is, we begin with the formal psychiatric label attached, but we do not continue reliance on it in our subsequent manipulations. (Of course, later, in testing a genetic model, we must return to this formal nosology, because the

identification of carefully diagnosed schizophrenic probands is necessary in studying the incidence of subclinical schizoidia in family members, to test a dominant gene hypothesis of the schizotaxic defect.) Secondly, we make use of crude content validity, based upon extended experience (some forty years of psychotherapy practice, including thousands of hours in the treatment of schizoid and schizophrenic patients). I have some faith in my clinical judgment and attempted to identify items in the MMPI pool that were relatively schizo-specific, and yet not mainly psychotic schizophrenia, reflecting the accessory symptoms. We are reassured to find a high agreement between Meehl's content validity intuitions and the empirically discriminating items. We then consider the mixed clinical population of psychiatric patients who were not formally diagnosed schizophrenia nor organic brain disease or affective psychosis. We have left a heterogeneous collection of psychoneuroses, character disorders, psycho-physiological reactions, marital problems, and the like. One knows from clinical experience, as well as a few statistical studies, that in such a miscellaneous group of patients there is a subset who, if followed over time, will develop a florid schizophrenia. Looking at it from the other ("input," causal) side, foster persons at high genetic risk for schizophrenia although they have about the same 12% diagnosis of clinical schizophrenia as they would if reared by their own mothers, more often show other kinds of neurotic, pseudoneurotic, or character anomalies (Heston, 1966; 1970). On a dominant gene hypothesis, in order to reach the theoretical 50%, one has to consider those persons as unrecognized schizotypes in varying degrees of pseudoneurotic or pseudopsychopathic decompensation. I am not reasoning circularly, but in the context of discovery, the theory tells you what to look for, and how to look. Having identified the provisional pool of indicator items by reliance on formal diagnosis, we now apply our taxometric procedure to bootstraps ourselves into an identification of good indicator items, estimates of the base rate, valid and false positive rates, and so forth. Let me sketch out three taxometric bootstraps procedures briefly, and refer you to our recent paper (Golden & Meehl, 1979) for details. I hope I can convey the essential ideas without tables or equations, because they are basically quite simple. Replacing the significance test orientation by the mental set "wherever possible, estimate point values," we proceed as follows: Some years ago, I proved some theorems based on the fact that the covariance of two fallible indicators of a taxon has a maximum value for a 50/50 taxonic mix (Meehl, 1973c; Note 5; Note 6). This is a useful theorem in bootstrapping from fallible indicators of a conjectured latent taxon, because we can simply plot the covariance of two of the indicators as a function of the third. If it behaves as predicted, starting out near zero and going through a maximum somewhere in the middle range of the third variable and then declining to zero again, we conclude that the latent model is as conjectured. Further, we infer that the class interval on the third indicator variable for which the observed empirical covariance of the first two has its maximum, is the one which is below the intersection of the two latent frequency functions (what I call the *hitmax interval,* as cutting at that interval maximizes the hit percentage). From the observed (*yz*)-covariance in the hitmax interval of *x,* we can compute the product of the latent taxon mean differences on *y* and *z.* Using this product $(\bar{y}_s - \bar{y}_n)(\bar{z}_s - \bar{z}_n)$, we obtain a quadratic in the latent interval schizotype-rate p_i, and solving that quadratic in each *x*-interval we infer the latent schizotype frequencies, adding them to get the total schizotype-frequency and the taxon base rate. (Statistical bootstrapping, properly done, *can* cook up knowledge out of ignorance!) Meanwhile, the computer has drawn the latent frequency functions, so we can solve for latent hit rates achieved by the hitmax cut. Finally, we use Bayes' Theorem to classify the patients individually as schizotypal or not. Notice that we estimate a hit rate despite having no criterion! (See Dawes & Meehl, 1966; Meehl, 1973c, pp. 216-217; Note 5, pp. 37-44.)

The "Super-Bootstraps Theorem" is counter-intuitive and totally ignored in the psychometric literature, but it could theoretically be a source of greater power in research on fallible indicators of latent taxa. This is especially likely in behavior genetics, where a non-behavioral pleiotropic effect could be discerned as quasi-infallible by using the Super-Bootstraps Theorem on a behavior syndrome whose component indicators possess only moderate validity.

Secondly, it can be shown that the hitmax cut is approximately located, within an interval error at most, by a sliding cut which maximizes the difference in the proportions of patients above and below the sliding cut who ring the bell on a candidate item. Beginning with a set of potentially good items one identifies strong items, tests them for consistency, rejects those that are most inconsistent with the model, thus creating a new sliding cut scale, and iterating until the system settles down, that is, until the valid and false positive rates, hitmax cuts, and inferred latent base rate of schizotypy are all numerically coherent. Thirdly, conjecturing an approximately normal distribution for the latent taxon and the latent extra-taxon class, one can use a kind of cut and try procedure, that minimizes discrepancy between the observed and theoretically predicted values when we *arbitrarily* assign base rates, means, and standard deviations. Here again, we have no criterion. We consider a model supported if there exists an assignment of the latent values which gives an insignificant (or very small) chi square to which the discrepancies move in an orderly fashion, as we approach the optimal choice of latent parameters. Using these three non-redundant methods, we arrive at an estimate of the validity of each selected indicator and of the estimated latent base rate. We get some faith in our results partly from the fact that the base rate estimated by these three independent methods is about the same, and that it is about the same as an antecedently recorded base rate. Some of you will be shocked by this value, but I guessed around 40% are unrecognized schizotypes in such a mixed psychiatric population. If the taxometric corroboration of my 40% clinician's guesstimate is sound, that is an important and disturbing finding. It suggests that a rather large batch of schizotypal patients are being non-optimally treated, and research results in psychotherapy, pharmacology, diagnostic reliability, and genetics are likely to be fuzzed up or even uninterpretable due to the undetected presence of that many schizotypes. We are reassured to find that the estimated validities of items found by the three procedures generates a predicted degree of agreement between them by pairs that is close numerically to the observed values. Further corroboration is found by showing that the MMPI profile of the group thus identified as schizotypal is very similar in pattern to a VA group diagnosed "anxiety neurosis" who subsequently developed florid schizophrenia on follow up (Peterson, 1954a, 1954b).

Summary

Inferred latent entities, whether those of psychoanalysis, factor analysis, or cluster analysis, have declined in value for many clinical psychologists, both as tools of practice and as objects of theoretical interest. Behavior modification, rational-emotive therapy, crisis intervention, psycho-pharmacology, and actuarial prediction all tend to minimize reliance on latent entities in favor of purely dispositional concepts. Behavior genetics is, however, a powerful movement to the contrary. As regards categorical entities (types, taxa, syndromes, diseases), history reveals no impressive examples of their discovery by cluster algorithms; whereas organic medicine and psychopathology have both discovered many taxonic entities without reliance on formal (statistical) cluster methods. I offer eight reasons for this strange condition, with associated suggestions for ameliorating it. Adopting a realist instead of a fictionist approach to taxonomy, I give high priority to theory-based mathematical derivation of quantitative consistency tests for all taxometric results. I urge a large-scale cooperative survey of taxometric methods based on

Monte Carlo runs, biological pseudo-problems where the true taxon is independently known, and live problems in genetics, organic medicine, and psychopathology. An empirical example of taxometric bootstrapping and consistency testing was presented from my own current research on schizotypy.

Reference Notes

1. Blashfield, R. K. *Failure of cluster analysis in psychiatric research.* Paper presented at the meeting of the American Psychological Association, Toronto, Canada, August 30, 1978.
2. Golden, R. R.. & Meehl, P. E. *Taxometric analysis of causal entities: Detection of the schizoid taxon*. New York: Academic Press, to appear. [This was never completed. –LJY]
3. ~~Golden, R. R., & Meehl, P. E. *Detection of biological sex: An empirical test of six cluster methods.* Manuscript submitted for publication, 1979.~~
4. 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* (Tech. Rep. PR-69-1). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1969.
5. Meehl, P. E. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Tech. Rep. PR-65-2). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1965.
6. Meehl, P. E. *Detecting latent clinical taxa, II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Tech. Rep. PR-68-4). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1968.

References

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

Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research*, *13,* 271-295.

Blashfield. R. K., & Draguns, J. G. (1976). Toward a taxonomy of psychopathology: The purpose of psychiatric classification. *British Journal of Psychiatry*, *129,* 574-583.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81,* 95-106.

Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin*, *66,* 63,67.

Depue, R. A., & Monroe, S. M. (1978). The unipolar-bipolar distinction in the depressive disorders. *Psychological Bulletin*, *85,* 1001-1029.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monograph*, *79*(9, Whole No. 602).

Golden, R. R., & Meehl, P. E. (1969). Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology*, *88,* 217-233.

Golden, R., & Meehl, P. E. (1980). Detection of biological sex: An empirical test of cluster methods. *Multivariate Behavioral Research, 15*, 475-496. [Reference updated by LJY]

Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

Heston, L. L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers. *British Journal of Psychiatry*, *112*, 819-825.

Heston, L. L. (1970). The genetics of schizophrenia and schizoid disease. *Science*, *167,* 249-256.

Lykken, D. T. (1956). A method of actuarial pattern analysis. *Psychological Bulletin*, 55, 102-107.

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

Meehl, P. E. (1945). The dynamics of structured personality tests. *Journal of Clinical Psychology*, *1*, 296-303. Reprinted with Prefatory Comment in L. D. Goodstein & R. I. Lanyon (Eds.), *Readings in personality assessment* (pp. 245-253). New York: Wiley, 1971.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press. Reprinted with new Preface, 1996, by Jason Aronson, Northvale, NJ.

Meehl, P. E. (1970). Theory-testing in psychology and physics: A methodological paradox. In D. E. Morrison & R. E. Henkel (Eds.). *The significance test controversy.* Chicago: Aldine. (Originally published in *Philosophy of Science,* 1967, *34*, 103-115.)

Meehl, P. E. (1973a). Some ruminations on the validation of clinical procedures. In P. E. Meehl, *Psycho-diagnosis: Selected papers.* Minneapolis: University of Minnesota Press. (Originally published in *Canadian Journal of Psychology,* 1959, *13*, 102-128.)

Meehl. P. E. (1973b). Specific genetic etiology, psychodynamics and therapeutic nihilism. In P. E. Meehl, *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press. (Originally published in *International Journal of Mental Health.* 1972, *1*, 10-27.)

Meehl, P. E. (1973c). MAXCOV-HITMAX: A taxometric search method for loose genetic syndromes. In P. E. Meehl, *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1973d). Wanted—a good cookbook. In P. E. Meehl, *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press. (Originally published in *American Psychologist,* 1956, *11*, 263-272.)

Meehl, P. E. (1973e). Why I do not attend case conferences. In P. E. Meehl, *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1977). Specific etiology and other forms of strong influence: Some quantitative meanings. *Journal of Medicine and Philosophy*, *2,* 33-53.

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

Meehl, P. E. (1970). Some methodological reflections on the difficulties of psychoanalytic research. In M, Radner and S. Winokur (Eds.), *Minnesota studies in the philosophy of science,* Vol. IV (pp. 403-416). Minneapolis: University of Minnesota Press. Reprinted *Psychological Issues,* 1973, *8*, 104-115.

Meehl, P. E., Lykken, D. T., Schofield, W., & Tellegen, A. (1971). Recaptured-item technique (RIT): A method for reducing somewhat the subjective element in factor-naming. *Journal of Experimental Research in Personality*, *5,* 171-190.

Murphy, E. A. (1964). One cause? Many causes? The argument from a bimodal distribution. *Journal of Chronic Diseases*, *17,* 301-324.

Peterson, D. R. (1954a). The diagnosis of subclinical schizophrenia. *Journal of Consulting Psychology*, *18,* 198-200.

Peterson, D. R. (1954b). Predicting hospitalization of psychiatric outpatients. *Journal of Abnormal and Social Psychology*, *49.* 260-265.

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

Popper, K. R. (1976). A note on verisimilitude. *British Journal for the Philosophy of Science*, *27,* 147-195.

Sneath, P. H. A., & Sokal, R. R. *Numerical taxonomy.* San Francisco: W. H. Freeman, 1973.

Sokal, R. R. (1974). Classification: Purposes, principles, progress, prospects. *Science*, *185,* 1115-1123.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83,* 312-317.

Wender, P. H. (1967). On necessary and sufficient conditions in psychiatric explanation. *Archives of General Psychiatry*, *16,* 41-47.

pdf by LJY, September 2004