In *Contemporary Directions in Psychopathology*, edited by
Theodore Millon and Gerald L. Klerman. NY: Guilford, 1986.
Pp. 215-231.

#130

CHAPTER 10

Diagnostic Taxa as Open Concepts: Metatheoretical and Statistical Questions about Reliability and Construct Validity in the Grand Strategy of Nosological Revision

*Paul E. Meehl**

Since I find it hard to conceive that a rational mind could think otherwise, I presuppose that, *ceteris paribus,* careful delineation of the signs, symptoms, and course of a disorder (I cannot interest myself much in the semantic hassle over whether to call it "disease") so as to increase the reliability of classifying clients or patients is desirable. While reliability and validity are not the same thing, it is a psychometric truism that the former bounds the latter, although it is worth mentioning that the bound is the square root of the reliability, so validity can theoretically be larger. Usually the operative validity (net attenuated construct validity) runs far below that upper bound set by the square root of the reliability coefficient. Hence, alterations in the format of assessment or in the content sampled, which might under some circumstances reduce reliability, could nevertheless increase the net attenuated construct validity. Similarly, changes in content or format that increase reliability may theoretically decrease validity. For instance, an alteration in the open-ended, unstructured format of Rorschach administration (as was attempted during World War II to make it possible to test large numbers of individuals and score reliably without inquiry) seemed to eliminate whatever slight validity the instrument had as usually administered.

There is no mystery about this, although it is paradoxical at first look. We may be concerned about the reliability of behavior sampling by two different samplers ("interjudge agreement") or with the trustworthiness of a sample as drawn by an individual judge (how many marbles do we draw from the urn, and how do we draw them?). In either case, the point is this: Whether an interview behavior or a psychological test item is viewed primar-

* Deceased 2007; at the time of this publication Meehl was at the University of Minnesota Medical School Minneapolis.

ily as “sample” or as “sign” (Cronbach & Meehl, 1955), there are kinds of alterations in the examining situation and in the procedure of response classification that can alter qualitatively the intrinsic construct validity of the sample in such a way as to reduce its net validity, despite reliability, in either of the two senses mentioned, having been enhanced.

I am not arguing that such has occurred in the process of improving our old Mental Status Examination or in the construction of DSM-III, but merely that this methodological point should be kept in mind when discussing reliability/validity questions.

It requires neither psychometric nor philosophical expertise to see that the reliability/validity helps and trade-offs can be somewhat complicated, and especially so when the aimed at diagnostic construct itself (category or dimension) is an open concept, lacking a definitive “operational” criterion, specified implicitly (“contextual definition”) by presumably fallible indicators. In that kind of knowledge situation, we subtly alter meanings as we discover facts, we amend theoretical definitions as we revise indicator weights. The basic point can be better brought out by considering the decision to include an unreliable indicator in a standard examination for “any disease.” A general medical examination always includes blood pressure and not anthropometric determination of wrist width, despite the mediocre reliability of the former and $r = .98$ for the latter. We do not find this evidentiary preference puzzling, we simply say, “Blood pressure unreliably measured is a stronger indicator of more different and important conditions than wrist width reliably measured.” Similarly, a psychotherapist who employs dream interpretation (with the manifest-latent content model) would not seriously consider substituting reliably scorable multiple-choice inquiry for free association under the Fundamental Rule, despite the grave reliability problems posed by the classical procedure. One might prefer to avoid Freud’s technique altogether, and partly because of unreliability considerations (cf. Meehl, 1983); what one would almost surely *not* do is retain Freud’s core idea and its entailed technique, while substituting a multiple-choice inquiry in the service of reliability.

The reasons for desiring diagnostic reliability are well known. The most important reason is generalizability of research findings by other investigators thinking about their research and by practitioners in applying research findings to clinical decision making. The easiest way to understand the former is in terms of the number of pairwise relationships of input and output variables involved in a decision-making process whether of a theoretical or practical clinical nature (Meehl, 1959). If a set of behavior data (history and current status, interview, ward behavior, neurological, psychometric) permit us to classify patients in some “rational” (ultimately “causal”?) way, it is not necessary that each of the possible *n* output variables (e.g., treatment of choice, second choice, prognosis, employability, response to group therapy, suicide risk, genetic risk to offspring) has to be correlated singly pairwise with all the input variables, a process which would require studying *mn* relationships, where *mn* is in the thousands (Meehl &. Golden, 1982, pp. 130-131). Instead, we can first relate the *m* input variables to the diagnostic dimension or rubric and then relate the diagnostic dimension or rubric to the several output variables of interest. Hence only (*m* + *n*) correlations need to be studied. But that process cannot be carried out with any confidence if the relation of some of the input variables to dimension X or categorical rubric C as found at the University of Texas has only a little better than chance relationship between (only partially overlapping) relationships as reported by investigators in Milwaukee. The pulling together of research data to give a coherent interpretation of an alleged psychiatric entity, whether taxonomic or dimensional in nature, presupposes the possibility of scanning the research literature with at least some reasonable confidence that patients called schizophrenic by one investigator are like those called schizophrenic by another. Similarly, suppose a clinician reads a research report claiming that a certain drug is efficacious for paranoid schizophrenics, except when they have a history of an episode, when much younger, of acute catatonic excitement. That report is not helpful to the clinician who lacks rational belief that the investigator was looking at the same indicators of paranoid and catatonic schizophrenia that he or she can now look at in his or her clinical decision making.

It is an interesting question whether one can ever lose by improving reliability, except in the (rare? I don't know) sense discussed above. The main respect in which some workers, and I gingerly include myself here, seem to worry about it is that research aimed at improving, correcting, or—in the extreme case—refuting views implicit in the DSM-III conceptual system will somehow be cramped by an overly enthusiastic view of it, which sometimes takes the form of a dogmatic insistence upon its merits throughout. I have heard research-oriented clinicians express concern about this, but it is difficult to track down persuasive examples where, for instance, an otherwise admirable research proposal was rejected by the peer reviewers on the grounds that it did not employ "official categories" approved by DSM-III. While people talk about this, and one sometimes hears it alleged that it has occurred, I do not myself know of any clear cases. Admittedly, it would be hard to ascertain whether a subtle kind of social process, of the kind that the Supreme Court likes to call a "chilling effect," is taking place. Some researchers might be otherwise disposed to advocate a mild Feyerabendian "proliferation of theories" (Feyerabend, 1970) which he advocates even for cases when the going theories are extremely powerful and well corroborated and, a fortiori, for theories in such primitive fields as psychopathology. Some of them might not be getting research grant money because they have timidly avoided challenging the establishment category system.

Here again, I have no affirmative evidence that such things happen. If they do, it would appear quite easy to find a way around it, and whether it failed would hinge upon whether some peer reviewers have become over-identified with the present product. For example: Suppose I am interested in studying people with a cyclothymic personality makeup who have very mild ups and downs on an endogenous (genetic/biochemical) basis, but who at no time become diagnosably psychotic or even semipsychotic. The psychiatric tradition has connected endogenousness with severity, which there is no strong theoretical reason for insisting upon, although there is a correlation empirically. I don't see why a clinical investigator, behavior geneticist, or neurochemist should in any way be hampered by the received rubrics. He or she can be careful in adhering to criteria for diagnosing manic-depressive disorder as given by DSM-III. It may well be that the only available rubric for some of the other people he or she wants to study is "normal," or even perhaps some *other* piece of non-manic-depressive terminology as specified in DSM-III. Nothing prevents the investigator from saying, in writing up a grant proposal, "It is my empirical conjecture that there are persons who don't manage quite to squeak through the conditions for diagnosing a manic-depressive attack (because of extreme damping in their cyclothymic cycle). But all of my classifications are indicated, and all of the correlations of them with all of the other things I studied, whether psychometric or genetic or familial or whatever, are clearly indicated, so that other investigators may rely on the fact that I stuck literally to the received criteria for making that diagnosis. I have also listed, however, the set of special criteria, together with their time sampling and interjudge reliabilities, that *I* used to demarcate *my* special subgroup of individuals that do not fit the official rubrics."

At no point does this investigator have to depart from the semantics of DSM-III, nor does he or she have to do any inordinate amount of work in order to include the DSM-III criteria as available for investigators who want to examine his or her data critically. It is, I suppose, imaginable that somebody might want to do something where the task of "double diagnosis" (i.e., according to his or her conjectured criteria for entities or dimensions not in the official list along with the received one) will be a considerable amount of excess work, but I am not aware of any clear showing that that has happened. The diagnostic criteria for DSM-III simply do not involve that much additional work, and most of the overload will arise from his or her idiosyncratic system. Despite the fact that my own views on many categories are quite heterodox, when there is an adoption by an empowered body of clinicians and scientists as to a certain terminology, I think one is not unduly burdened or imposed upon by some extra scientific or clinical toil when the investigator chooses to deviate from it in his or her own research.

An interesting statistical question arises in the "context of discovery" (Reichenbach, 1938, pp. 6-7) where a plausible case—I do not urge that it is more than plausible—can be made for concern about increased difficulty of detecting subtle relationships. I mean by "detection" the development of a clinical hunch and, in a more formalized research context, the problem of the statistical power function failing to detect something that is there. Consider the following: By tightening up the diagnostic criteria, we have increased reliability and, hence (almost certainly), the net attenuated construct validity in identifying the whole class of patients called "schizophrenic." In the course of so doing, we have been forced to eliminate some signs and symptoms that some clinicians have been relying on. Perhaps we ourselves had been doing so, but we are willing to pay this price. We are even willing to pay the price of dropping something that was considered fundamental by the master himself, as, for instance, DSM-III does not include Bleuler's *ambivalence* or his *autism;* or, to take an instance closer to my heart, Rado's *anhedonia* (Meehl, 1962, 1964, 1974-1975, 1975). Less counterconventional, one thinks of the pan-anxiety considered extremely important—perhaps the most important single symptom—in the "pseudoneurotic schizophrenia" syndrome described by Hoch and Polatin (1949). The latter two examples are of course controversial; but as to the former, it is hard to believe that we should omit two of Bleuler's cardinal signs unless this choice is dictated by difficulty objectifying them in the interest of reliability.

I repeat that I am not here disputing the claim that the net attenuated construct validity for identifying the whole class of schizophrenics has been increased by the tightening process, and I am not at the moment concerned with the efficacy of clinical handling, but I am attending to the research context. It is surely possible, and to a statistically and philosophically sophisticated person not even paradoxical, that some subset of patients sharing underlying etiology and psychopathology (genetics, biochemistry, CNS fine structure, and psychodynamics or "personality structure") with the core group of schizophrenias but who, because of modifying genes and normal-range individual differences factors (Meehl, 1975) as well as life history experiences, do not develop the signs and symptoms that have remained in the selected list of DSM-III, or—equally possible despite average heightened reliability—do not have them in sufficient quantity to be clear instances. Such a state of affairs is not only consistent with, but is probabilistically inferable from either the medical model, classical psychometrics, genetics, learning theory, or ordinary trait theory. The point is that clinicians trained to classify patients with the reduced high-reliable list of criteria will not be psychologically disposed to consider the subset of peripheral or borderline cases as belonging to the schizophrenic group (as they should not in applying the objectified criteria). In the context of discovery, this could sometimes operate adversely, since the way you categorize your world, as we all know, will in considerable part determine what you are capable of noticing.

But suppose a perceptive clinician does notice something about these borderline cases and undertakes a systematic research study of something middling complicated and not easy to discern, say, for example, a second-order interaction between phenothiazines and a certain mode of psychotherapeutic intervention (e.g., RET). Now if the polygenic modifiers or environmental factors that make the atypical schizophrenias show a different kind of interaction effect from the core group, that will not be detected statistically, even having been noticed clinically by a gifted clinician, because such cases will only rarely (and mostly due to carelessness in applying the new criteria!) be included in the study. If one believes (as I do) that the psychiatric treatment of the future will involve complicated kinds of actuarial grounds for selecting and sequencing the treatment of choice (Meehl, 1972a, pp. 135-137), early research progress along such lines could be hampered in this way.

I want to emphasize that I'm not here invoking some kind of vague clinical intuitions about "patterns." I am making a simple point about research statistics, that is, that you can't detect a trend that makes a subset of subjects different from the other subjects in a certain group if there aren't any of the subset present in the study. Furthermore, as we move into higher order interaction effects, such as Drug X Psychotherapy X Subdiagnosis patterns, the degrees of freedom shrink so that errors of Type II begin to preponderate due to marked reduction in statistical power.

It might be argued that while this may impose an irksome hurdle in the context of discovery at the intuitive stage for the clinician trained in the use of DSM-III, and thinking more or less automatically that way, it will not have any long-run bad effect because the cases not included in such studies will be detectable in studies focusing on some other diagnostic rubric. I think that is an optimistic view because it implies that some sort of massive research network of all possible combinations of everything with everything is going to take place in psychiatry and clinical psychology, which it is not, both for economic and professional interest reasons. Furthermore, what kind of thing is detected will depend upon what initial overall rubric is being studied. If these borderline cases were subsumed under "anxiety state" rather than borderline schizophrenia, the interaction effect between an antipsychotic drug and cognitive therapy will not be a likely subject matter of investigation. Finally, what is perhaps the more serious statistical point, such people will not be found in any one rubric if misdiagnosed because of the tight criteria [by misdiagnosed, I of course mean subsumed to the wrong specific etiological group (Meehl, 1972b, 1977) in the eyes of Omniscient Jones], but are likely to be dispersed. When one disperses a group of people who are heterogeneous in some respects, but homogeneous in some core feature of high causal relevance, into a number of heterogeneous diagnostic categories, the best bet is that they will simply get lost in the shuffle. While I do not claim to know that this is a serious problem, it is not a silly consideration that can be dismissed out of hand without thorough mathematical analysis.

Moving away from what one might call the "political-social-economic" impact of DSM-III, it is worthwhile to examine at a more philosophical level the ways in which a practitioner or researcher may view its categories and dimensions. I can see three (although not sharply demarcated), one of which is admirable, one of which is criticizable but fairly harmless, and only the third of which is scientifically malignant. The first is to view the delineation of a syndrome as an empirically observed (clinically or statistically!) cluster, a syndrome plus course, that suggests to us some kind of underlying causal homogeneity in the subjects who show it; although we may, depending on our theoretical predilections, sit quite loosely to this etiological promissory note. Its justification is mainly communicative and pragmatic, together with whatever degree of faith we have from the history of medicine (and genetics, and psychometrics) that future research will give us a more detailed understanding of whatever historical and "latent" (inner) current processes and structures are at work to produce the covariation of the signs, symptoms, aspects of course, prognosis, and response to treatment. Covariation is the essence of descriptive science and the touchstone of scientific thinking, whether we read such diverse writers as Freud, Skinner, Allport, Murray, Eysenck, Thurstone, or Cattell—strange bedfellows indeed, whose unanimity on *this* point should surely tell us something about how to study the mind! *Ceteris paribus* again, the more standardized the examination can be made, the more objectively described the classification of the responses, and as a result, the greater interjudge agreement by different examiners, and the more striking the empirical "tightness" of the cluster, the better we like the syndrome as an entity. As already stated, it is hard to understand why a rational mind would object to any approach that enhances these desirable properties.

Second, one may believe that DSM-III is the best that can be achieved, at least in the foreseeable future, and may be suspicious or even antagonistic to deviations from it, for either clinical or research purposes. This attitude troubles me, but I should think it can be adequately buffered by the practice I suggested above, that is, that investigators have a responsibility to employ it until it is officially revised by some "culturally empowered" group such as those who constructed it in the first place. But we do not pressure researchers or punish them financially or otherwise, once they have met these conditions in their semantics, for delineating some further conjectural entities or dimensions of their own, hoping to persuade the profession on the basis of clinical experience of better evidence, that they are right.

It is the third attitude which I think is malignant, partly because of its potential chilling effect, but mainly because it is philosophically so terribly mistaken. It says not merely that "this is a good thing so far as it goes, and should not be lightly discarded or whimsically amended." This third view claims it is *the* truth, as a matter of some kind of rigorous definition process. The extreme (simplistic, "vulgar operationist") form of this view is that the very *meaning* of the concepts is contained, exhaustively and explicitly, in the "operational definitions" provided by DSM-III. It would be hard to find one single logician or historian of science today (or for that matter, since around 1935!) who would countenance the conception of scientific method enshrined in this view. I find it puzzling that physicians, or for that matter, psychologists, unless they are of the most dogmatic behaviorist kind, should adopt this position when neither the history of organic medicine, nor of genetics (I don't mean here merely behavior genetics), nor of traditional trait theory in academic psychology, nor of classical psychometrics, gives any support to it. It is simply not true that diseases in organic medicine are "defined by" the syndrome or by the syndrome and course together. Organic diseases are defined by a conjunction of their etiology and pathology when these are known; and otherwise—with much less scientific assurance—as syndromes remaining to be researched so as to be medically understood. A disease entity, as delineated in the early stages of clinical experience and scientific study, at the level of mere syndrome description when there is as yet no (or minimal and conjectural) knowledge of the etiology or pathology underlying it, is an open concept (Meehl, 1972b, 1977; Cronbach & Meehl, 1955; Meehl & Golden, 1982; Pap, 1953, 1958, Chap. 11). It is neither philosophically rigorous nor scientifically sophisticated to make a literal identification of a disease entity with its currently accepted signs and symptoms. Corresponding to organic medicine's pathology (in a more extended sense than that envisaged by Virchow) is personality structure (genotypic traits, psychodynamics). Corresponding to etiology are, except for an environmentalist fanatic, the genetic predispositions not only to specific mental disorders, but to "temperamental genotypic traits" generally, such as anxiety conditionability, rage readiness, hedonic capacity, general intelligence and the like, and the learning history imposed on an organism whose varied behavior acquisition functions are characterized by such-and-such inherited parameters. Our problem in psychopathology of the so-called functional behavior disorders is obvious, to wit, that we do not possess an equivalent to the pathologist's and microbiologist's report telling us the "right answer" at the conclusion of a clinicopathological case conference (Meehl, 1973, pp. 284-289). If I make a psychodynamic inference, it is not possible for me to ask the psychopathologist whether his stained slides showed the patient's psyche had holes in the superego. To a thoughtful clinician with philosophical sophistication, it is perfectly obvious that disease syndromes are inherently open concepts, as mentioned above. Nothing but dogmatism on the one hand, or confusion on the other, is produced by pretending to give operational definitions in which the disease entity is literally identified with the list of signs and symptoms. Such an operational definition is a fake.

If somebody does not like the medical model (and if that's the case, one wouldn't be taking DSM-III—concocted by a group of psychiatrists for medical purposes—seriously to begin with), he should be reminded that in classical psychometrics (such as factor analysis) or in more recent developments (such as multidimensional scaling), we cannot even write the basic equations, let alone the embedding interpretative text required to give empirical meaning to the variables in those equations, unless a clear distinction is already made between the manifest behavior indicators and the inferred (latent, causal) factors. The same is true of biophysical trait theory as classically elaborated by Allport (1937), Murray (1938), Cattell (1946), and others. Obviously, the great breakthrough in genetics with Mendel, and the rediscovery of Mendel's concepts at the turn of the century, hinged upon the distinction between the genotype and the phenotype. This distinction forced theoretical recognition that under many circumstances or available pedigrees, the weakly stochastic relationship between the two made an inference to genotype impossible.

One simple-minded mistake that I am surprised to find physicians making is to think that if, in a given concrete instance (single case, not class), we do not have a touchstone for testing whether a certain inferred construct property such as a latent disease is present or absent, that lack means that it is scientifically meaningless to ask the question, a view that the logician Carnap, a strongly positivist and tough-minded philosopher of science, refuted definitively almost a half century ago!

The same is true of most variants of learning theory, the old-fashioned kind (Tolman, Hull or Guthrie) as well as the souped up developments in mathematical learning theory, information processing, and cognitive processes generally that took place subsequently. The only plausible exception to the genotypic/phenotypic, inner/outer, inferred/observed distinction in learning theory is strict Skinnerian learning theory which is almost entirely dispositional, although not as "pure" in this respect as some of its adherents like to think when they talk metatheory about it.

I am fond of referring clinical psychology students to a little known 2-page article published many years ago by the late T. A. Peppard (1949), a reputedly brilliant diagnostician who practiced internal medicine in Minneapolis for many years. He made a statistical study of the source of his diagnostic mistakes, using very strict criteria against postmortem findings. Errors of omission (well known to be commoner than errors of commission in medical diagnosis) sometimes occurred because he failed to look for something, other times because he looked for it but didn't give it the proper weight, other times because he made an "error" on a judgment call, and so on. But the interesting thing is that 29% of the errors of omission were attributable, even by very tight standards imposed on himself, to the factor he called "symptoms and signs not found." Of course, all physicians know the concept of "silent disease" such as an undiagnosed staghorn kidney or an early Pick's frontal lobe atrophy, not to mention subjects with an epileptic brain wave who never have a fit and would not be discovered except for being the monozygotic twin of somebody who has a clinically recognizable convulsive disorder. I repeat that I find it strange that one must remind physicians about the distinction between the construct "disease" and its presently accessible symptom picture, although it is not so surprising that some psychologists confuse them.

Finally, of course, the most obvious example, which would still be persuasive to some of my generation, is psychodynamics, whose essence consists in the distinction between the easily observed manifest behavior or self-awareness and the "hidden, latent, underlying source" of some aspect of observable covariation.

Since neither psychodynamics, classical psychometrics, taxometrics, organic medicine, genetics, learning theory, or trait theory has proceeded by explicit identification between theoretical entities and their indicators, it would be strange to hold that rational use of DSM-III requires us to consider its syndromes as literally definitive and totally noninferential.

It might be argued that if the builders of DSM-III had achieved consensus on constructing a purely descriptive (atheoretical, noninferential) "phenomenological" taxonomy, they should have proceeded by applying an appropriate formal cluster algorithm to a huge batch of carefully gathered clinical data, "letting the statistics do the whole job for them," which would have saved a lot of conference time as well as generating a more objective scientific product. This sounds plausible to a psychologist, and maybe to some statisticians, but the main trouble with it is that there is no "accepted" cluster algorithm which is known to be sufficiently powerful to be used in this way (cf. Meehl, 1979). Even if there were such an agreed upon cluster analysis algorithm, one doubts that the committee could have proceeded in that way. The fact is that different clinicians do not share an equally "operational" view, partly for the reasons I have given and partly because of certain clinical (perhaps one could even say ideological) identifications, for example, between organicists and psychoanalysts, biotropes and sociotropes, scientists interested in genes and psychotherapists interested in battle-ax mothers.

I am inclined to think that the next round ought to at least settle on some way of deciding when the orientation should be taxonomic versus dimensional. But that would hinge upon having a sufficiently well-trusted algorithm for determining whether the latent order of a syndrome or dimension should be thought of as taxonic or nontaxonically factorial. Another possibility, which again seems simplistic and arbitrary until you ask what are the reasons for doing it another way, would be to collect all of the information or input kinds of variables, including life history data and the like, that go into diagnosis, and all of the output dispositions that are clinical reasons for making a diagnosis, such as differential response to psychotropic drugs, response to individual and group therapy, danger of acting out, suicide risk, and long-term employability. Absent cogent reasons for giving higher weight to some of these output ones than others, it is arguable that the proper statistical model should be canonical correlation in which we simultaneously optimize the predictability of the most predictable composite on the output side by optimal weights on the input variables. If the various output consequences of clinical importance are not prima facie very different in "importance," if they are, so to speak, qualitatively of equal significance to us in decision making, then the difference in the weights they get might best be to weight them so as to make them collectively most predictable. The justification for defining a syndrome (or a nontaxonic factor) by some subset of input and output considered jointly would be that the canonical correlation between the two sets reach a certain minimum size. It would be interesting, by the way, to ascertain whether such a distribution of candidate canonical correlations would show, if not an actual break, at least some tendency to bimodality, suggesting that some syndromes are "real" and others are more or less arbitrary carvings out by the clinician of regions of slightly greater densification in the multivariate descriptor space (but see Murphy, 1964). My own research interests are such that I consider that the initial distinction between whether one should proceed taxometrically or factorially should be given very great priority in the next revision.

The question as to the desirability of adopting a fixed rule approach to diagnostic criteria involves a complicated mix of statistical, philosophical, and clinical issues that are beyond the space limitations of this chapter and about which I myself have formed no definite opinion. This question has been aired recently in papers by Finn (1982, 1983) and Widiger (1983) [see also Meehl and Rosen (1955), comment by Cureton (1957), and Rorer, Hoffman, La-Forge, and Hsieh (1966)]. In thinking about this difficult question, it is necessary first to distinguish between issues regarding base rate fluctuations in different clinical or research populations and the separate but intimately related issues of clinical utility in treatment and prognosis. In saying these are distinct but intimately related, I mean to emphasize that from the standpoint of scientific realism (surely the implicit assumption of organic medicine, whether medical researchers or practitioners use the philosopher's terminology for it or not!), one does not wish to conflate the probability or corroboration of a diagnostic statement as a factual claim with the seriousness of a mistake. As Widiger worried about in his exchange with Finn, we do not want to adopt a decision rule based on a policy of systematically misdiagnosing patients on the grounds that correctly diagnosing a subset of them would, in certain pragmatic contexts, be too costly or risky or have too many side effects or make them more uncomfortable than the disease makes them, or whatever. Crudely put, the first business of a diagnostic assertion is to be right! We cannot make use of differential utilities and disutilities of clinical errors without at least some crude assessments of diagnostic confidence, whereas we can investigate the optimally of a diagnostic procedure with regard to truth value without referring to any utilities other than the "cognitive utility" of being correct in our assertions. It would seem best, if it can be done and is psychologically acceptable to practitioners, to optimize the diagnoses by some suitable adjustment for known or guesstimated base rates in a given clinical population, and subsequently to raise the question of the various utilities involved in adopting a certain treatment plan or making predictions to the patient, court, employer, insurer, family, or whatever. In that mode of reasoning, the best inferable diagnostic statement is made first and the utilities are plugged in afterward.

But this of course doesn't take care of the base rate problem. Theoretically we know that both the cutting score on a variate which is an indicator of the disease entity and any formal or informal weighting of the scores or way of combining them into a pattern, as in Bayes's formula, should not be done independently of the base rates. In ordinary clinical medicine, practitioners who never heard of the Reverend Thomas Bayes or the subsequent controversy about his ideas (this use of the formula itself is, of course, hardly controversial) make implicit use of it. They know that if you diagnose syphilis in Puerto Rico on the grounds of a positive Wassermann you are likely to fall into errors that you would not make in Minnesota because of the geographic epidemiology of lues versus yaws. Every general practitioner at times says to the patient, "Well, I think you've got the winter crud; there's a lot of that going around these days," an informal Bayesian inference. It is an unsettled question how much the explicit and formalized inverse probability machinery of the statistician should become part of the decision making by a busy doctor. Of course, even given a certain diagnosis, perhaps tentatively arrived at with the intention to be flexible about revising it should the predicted results of a therapeutic intervention fail to materialize in the usual fashion, it is common practice, within the category of patients who meet the diagnostic criteria, to pay attention to the pattern of symptoms that is relevant to treatment choice and to include in this those "extraneous" characteristics (e.g., age, family, income, unrelated concurrent illness) that themselves did not enter into the diagnostic decision proper.

There is nothing either wrong or particularly complicated about any of this. The only question is the extent to which formalization improves or impairs certain of these generally accepted clinical practices. Unfortunately, the behavior of a Bayes theorem computed inverse probability depends in somewhat complicated ways upon the distribution of sign validities, the relationship between valid and false-positive rates, the extent to which the independence assumption of the signs pairwise is not satisfied, how robust the inferred diagnostic *p* value is with respect to departures from those assumptions, differential responsiveness of error rate at different regions of the base rate continuum, and the like. It would seem that some rather large-scale but also intensive research by statisticians and clinicians would be in order.

I do not think it is safe to assume that because such actuarial refinements are not part of the everyday mental habits of practitioners in organic medicine, then, we don't have to worry about it in psychopathology. There are probably important differences in the latter area. Further, we still do not know the extent to which ordinary clinical practice of organic medicine commits more diagnostic errors than need be because of the extent to which the mathematics of clinical reference is not explicitly employed by the practitioner (Blois, 1980; Dawes, 1979; Dawes & Corrigan, 1974; Engelhardt, Spicker, & Towers, 1979; Goldberg, 1970, 1976; Gough, 1962, Holt, 1970, 1978; Kahneman, Slovic, & Tversky, 1982; Kleinmuntz, 1982; Meehl, 1954, 1956a, 1956b, 1956c, 1957, 1960, 1967; Sawyer, 1966; Sines, 1970). Finally, how one thinks about this and what kinds of research are conducted depend on how confident we are that the underlying psychopathology is intrinsically taxonic (categorical, "typal") versus nontaxonically multidemensional where class concepts and qualitative predicates are only handy rubrics for roughly designating regions in an ontologically continuous descriptor hyperspace.

While the very title of this volume orients us toward revision, one hopes that the intellectual fretfulness of primates and the availability of taxpayer dollars will not induce us to attempt substantial revisions until a large mass of evidence, including experimental research, clinical trials, quantitative analysis of clinical file data, and exchange of experience by seasoned practitioners of various persuasions, puts us in a position to do something more than speculate or nitpick. A tremendous amount of work by able people and a lot of taxpayer money went into generating the present product, and it is foolish to tinker with it very much, let alone undertake a complete overhaul, because it isn't perfect, or because the results of an unavoidable compromise are not located precisely where one might prefer it, given his own theory and practice. Sometimes the best advice is that of the Baptist preacher, "Leave it lay where Jesus flang it." I bethink myself of how difficult it is for me, after 40 years on the Minnesota faculty, to interest myself in interminable discussions about how we should revise the written preliminary examination for the Ph.D. so as to get a better assessment, reduce student anxiety, or whatever. A half century of observation (if I include my student days) reveals mostly primate meddlery, irrational optimism, a disinclination to consult the past, and the Hegelian swing of even short-term history!

It goes without saying that the most important developments one can anticipate that would make it rational to revise are substantive advances in our understanding of mental disorder. But there are also, I think, several metaquestions that it would be desirable to have "settled" (if not exactly *solved*) before the next round of major revision. The reader will discern that the "answers" to these metaquestions involve a mix of mathematical development of statistical methods especially suitable for taxonomic problems, the usual impact of substantive developments upon methodology (no contemporary philosopher of science conceives of methodology as entirely prior to theory), and considerations of clinical utility. I repeat that it is a grave mistake to conflate this last class of questions with questions regarding the intrinsic science, that is, factual validity, of any proposed concept. There are four metaquestions that should meanwhile be addressed by high-competence investigators so that we will be in good methodological shape when the time for major revision arrives. Without dogmatism, I might go so far as to say that in my judgment until these four are answered, at least in the sense of a fairly high consensus among qualified individuals (there is no point in absolute democracy in a field like this!), we are probably not in a cognitive position that warrants a major revision being attempted.

First, what role should a conjectural etiology, when moderately to strongly corroborated, play in the taxonomic strategy? Here one must avoid a simplistic division into "known" and "unverified" etiology, assuming a sharp dividing line where none exists even in organic medicine, genetics, or other fields of knowledge. It is obvious on mere inspection of the present list of rubrics that etiological factors partially understood, and in which varying degrees of "strong influence" as causal factors (Meehl, 1972b, 1977) must have been taken into account at least behind the scenes, have been unavoidable. It will be necessary to have a uniform standard of proof rather than a double standard of methodological morals such as prevails in some quarters today. For example, there are clinicians in the medical and psychological professions who resist recognizing the genetic influences in major mental disorders or, while reluctantly recognizing them, would not want to split the nosology of affective disorders into unipolar and bipolar despite the strong evidence available presently as to the reality of that distinction genetically and its correlates with certain aspects of the syndrome, course, and so on. Yet, some of these same clinicians, while justly pointing out that an absolute hammer blow unavoidable demonstration (there is no such thing as this in empirical science, of course) has not been given for the unipolar/bipolar distinction, will in their own diagnostic thinking rely upon highly speculative psychodynamics, or family factors, or other alleged causal influences, whose degree of evidentiary support at the present time is nowhere in the running with that for the biological distinctions made. This parallels some clinical psychologists who, because of hostility to medicine (or simply poor training at a second-rate school?), continue to decry all psychiatric diagnosis as "mere labeling" or "completely unreliable," refusing to read the quantitative evidence of diagnostic reliability developed in recent years, and then by some obscure mental process (which I confess myself quite unable to understand) proceed to substitute for such "unreliable" psychiatric nosology a batch of unproved, politicized social determiners, or flimsy psychodynamics inferred from an instrument with as low reliability and validity as the Rorschach! That is the sort of thing I mean by double standard of epistemological morals.

Second, the strategic distinction between thinking in terms of dimensions and categories (types, species, taxa, disease entities) remains with us. While one can get by with a kind of compromise between these, the basic theoretical claim of a classification system should be methodologically clear, even if a sizable proportion of patients are not clearly sortable into one or the other (a different question). Sooner or later we should get clear about which of our nosological rubrics are intended to be rough designations of persons location in a multidimensional descriptor space (whether phenotypic or genotypic, psychodynamic or genetic, that's not the point) and which rubrics have a genuine typological (taxonomic) theoretical intent. Thus, for instance, the very meaning of some standard terms in epidemiology and psychometrics, such as "false positive" and "base rate," which can be made tolerably clear on a taxonomic model, becomes fuzzy and—if the point is pressed—hardly interpretable on a nontaxonomic model. An adequate understanding of the philosophical and statistical aspects of this in relation to substantive theories of causation might properly lead us to abandon the idea of rubrics entirely for some subsets of conditions. For example: When I used to teach clinical psychology, in order to make this point I sometimes pushed the following (doubtless exaggerated) doctrine: There are several major mental disorders (e.g., schizophrenia, manic-depression, unipolar depression, delirium tremens, Alzheimer's disease) that are truly taxonomic in nature, and for which category rubrics are semantically strictly appropriate, not merely as rough ways of delineating regions in a continuous descriptor space. There is also, in my opinion, a true entity of the solid gold essential psychopath (sociopathic personality, asocial, amoral type). But when we get to the so-called neuroses and psychophysiological disorders of the neurotic kind, there is only one rubric (with the possible exception of the textbook obsessional neurosis), namely, “psychoneurosis, mixed,” a term no longer found in the official nomenclature. The distinctions within that mixed category are quantitative only; they are merely differing degrees of anxiety, depression, somatization, and defense mechanisms in the neurotic mixture. In the long run, it may be worth the trouble to teach clinicians to think more dimensionally than categorically and mold their verbal and inferential habits in those directions.

Third, we should get clearer than we presently are about the matter of sliding cuts on various indicators of an entity in relationship to base rates and various clinical populations in geographic, social classes, and the like, and the relevance of Bayes’ theorem. In matters where extremely asymmetrical likelihoods exist for the combination of a small number of high-valid signs, the importance of the base rate, except for the most extreme values, is considerably reduced, and it is probably statistical pedantry to push some kind of Bayes’ theorem algorithm onto working clinicians under such circumstances. I think that more mathematical analysis in relationship to the diagnostic habits of practitioners is in order here before altering the character of a psychiatrist’s or clinical psychologist’s education in this regard. Nobody acquainted with my writings would suspect me of being even faintly “antistatistical” in my biases; but I believe we should think like behavioral engineers in considering ourselves and others as clinical practitioners, taking into account what kinds of psychological disruptions in diagnostic cognitive activity could take place that might reduce net efficiency, even though the underlying mathematical model makes it look like an improvement.

Finally, at the risk of projecting my own current research interests, I would say that a desideratum for the next major revision is agreement upon the general taxometric problem as such, which I see as having two elements: (*a*) Is a taxometric procedure in psychopathology aimed at anything more than identifying phenotypic clusters, and, if it is, (*b*) which of the available formal taxometric methods (if any!) have shown themselves capable of detecting an underlying causal structure (whatever its biological or social nature), being meanwhile free of any appreciable tendency to detect taxonic structures that aren’t there (Meehl, 1979)? I think it not unduly optimistic to opine that we will have a pretty clear answer to the second question before the end of this decade (Grove & Andreasen, Chap. 17 of this volume; Meehl & Golden, 1982; Sneath & Sokal, 1973).

REFERENCES

Allport, G.W. (1937). *Personality: A psychological interpretation.* New York: Henry Holt.

Blois, M.S. (1980). Clinical judgment and computers. *New England journal of Medicine, 303,* 192-197.

Cattell, R.B. (1946). *Description and measurement of personality.* New York: World Book Company.

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

Cureton, E.E. (1957). Recipe for a cookbook. *Psychological Bulletin,* 54, 494-497.

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.

Engelhardt, T.H., Spicker, S.F., & Towers, B. (Eds.). (1979). *Clinical judgment: A critical appraisal.* Boston, MA: Reidel.

Feyerabend, P.K. (1970). Against method: Outline of an anarchistic theory of knowledge. In M. Radner & S. Winokur (Eds.), *Analysis of theories and methods of physics and psychology. Minnesota studies in the Philosophy of Science* (Vol. 4). Minneapolis: University of Minnesota Press.

Finn, S.E. (1982). Base rates, utilities, and DSM-III: Shortcomings of fixed-rule systems of psycho-diagnosis. *Journal of Abnormal Psychology,* 97, 294-302.

Finn, S.E. (1983). Utility-balanced and utility-imbalanced rules: Reply to Widiger. *Journal of Abnormal Psychology, 92,* 499-501.

Goldberg, L.R. (1970). Man versus model of man: A rationale plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422-432.

Goldberg, L.R. (1976). Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance, 1.6,* 13-22.

Gough, H.G. (1962). Clinical versus statistical prediction in psychology. In L. Postman (Ed.), *Psychology in The making.* New York: Knopf.

Grove, W.M., & Andreasen, N.C. (1986). Multivariate statistical analysis in psychopathology. In T.M. Millon & G.L. Klerman (Eds.), *Contemporary Directions in Psychopathology* (pp. 347-362). NY: Guilford.

Hoch, P., & Polatin, P. (1949). Pseudoneurotic forms of schizophrenia. *Psychiatric Quarterly, 3,* 248-276.

Holt, R.R. (1970). Yet another look at clinical and statistical prediction. *American Psychologist, 25,* 337-339.

Holt, R.R. (1978). *Methods in clinical psychology: Vol. 2. Prediction and research.* New York: Plenum.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* London: Cambridge University Press.

Kleinmuntz, B. (1982). Computational and noncomputational clinical information processing by computer. *Behavioral Science, 27,* 164-175.

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, Northvale, NJ: Jason Aronson.]

Meehl, P.E. (1956a). Clinical versus actuarial prediction. In *Proceedings of the 1955 Invitational Conference on Testing Problems* (pp. 136-141). Princeton, NJ: Educational Testing Service.

Meehl, P.E. (1956b). Wanted—a good cookbook. *American Psychologist, 11,* 263-272.

Meehl, P.E. (1956c). The tie that binds. *Journal of Counseling Psychology, 3,* 163-164.

Meehl, P.E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4,* 268-273.

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

Meehl, P.E. (1960). The cognitive activity of the clinician. *American Psychologist,* 75, 19-27.

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

Meehl, P.E. (1964). *Manual for use with checklist of schizotypic signs.* Psychiatry Research Unit, University of Minnesota Medical School, Minneapolis, Copyright 1964.

Meehl, P.E. (1967). What can the clinician do well? In D.N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 594-599). New York: McGraw-Hill.

Meehl, P.E. (1972a). Reactions, reflections, projections. In J.N. Butcher (Ed.), *Objective personality assessment: Changing perspectives* (pp. 131-189). New York: Academic Press.

Meehl, P.E. (1972b). Specific genetic etiology, psychodynamics and therapeutic nihilism. *International Journal of Mental Health, 1,* 10-27.

Meehl, P.E. (1973). Why I do not attend case conferences. In *Psychodiagnosis: Selected papers* (pp. 225-302). Minneapolis: University of Minnesota Press.

Meehl, P.E. (1974-1975). Genes and the unchangeable core. *VOICES: The Art and Science of Psychotherapy, 38,* 25-35.

Meehl, P.E. (1975). Hedonic capacity: Some conjectures. *Bulletin of the Menninger Clinic, 39,* 295- 307.

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

Meehl P.E. (1983). Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed.), *Testing scientific theories. Minnesota studies in the philosophy of science* (Vol. 10). Minneapolis: University of Minnesota Press.

Meehl, P.E., & Golden, R. (1982). Taxometric methods. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: Wiley.

Meehl, P.E., McArthur, C.C., & Tiedernan, D.V. (1956). Symposium on clinical and statistical prediction, *journal of Counseling Psychology, 3,* 163-173.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

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

Murray, H. A. (1938). *Exploratory in personality.* London: Oxford University Press.

Pap, A. (1953). Reduction sentences and open concepts. *Methodos, 5,* 3-30.

Pap, A. (1958). *Semantics and necessary truth.* New Haven, CT: Yale University Press.

Peppard, T. A. (1949). Mistakes in diagnosis. *Minnesota Medicine, 32,* 510-511.

Reichenbach, H. (1938). *Experience and prediction.* Chicago, IL: University of Chicago Press.

Rorer, L.G., Hoffman, F.J., LaForge, G.E., & Hsieh, K.E. (1966). Optimum cutting scores to discriminate groups of unequal size and variance, *Journal of Applied Psychology, 50,* 153-164.

Sawyer, J. (1966). Measurement and prediction, clinical *and* statistical. *Psychological Bulletin, 66,* 178-200.

Sines, J.O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry,* 776, 129-144.

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

Widiger, T.A. (1983). Utilities and fixed diagnostic rules: Comments on Finn (1982). *Journal of Abnormal Psychology, 92,* 495-498